9/3 FoML 1

Tuesday, September 3, 2024 11:02 AM

- Underlying model of the data

X: input / covariate Ex: X: 2, 5, 8

Y: output / response Y: 11

(X, Y) is a joint random vector, drawn from a Probability distribution P (which might be horribly complicated)

- Prediction is not always possible

↳ sometimes there is inherent noise in the data

Ex. X and Y are independent

- No learning algorithm is universal

No-free-lunch theorem: for any learning algorithm A, there exists a data distribution P s.t. A fails. Ex. Number Seq.

Consequence: Any ML algorithm needs to make assumptions on how data is generated.

This course:

→ Describe standard assumptions

→ Introduce a statistical framework to apply / study performance of the ML algorithms.

→ Cover main learning algorithms.

9/5 FoML 2

Thursday, September 5, 2024 2:03 PM

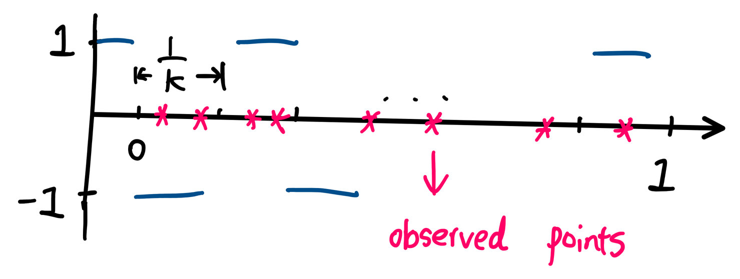
Machine Learning Paradigms

- A no-free-lunch theorem:

(We will give an instance of an "impossible" learning problem)

Let n: number of training examples

Consider $\boxed{k >> n}$



for $r_j = \begin{cases} +1 & \text{with prob. } \frac{1}{2} \\ -1 & \text{with prob. } \frac{1}{2} \end{cases}$ & $\boxed{\text{independent}}$

$r_j = \pm 1$, $j = 1, \dots, k$

$\boxed{\text{Target Function}}$ (Random)

Now let's consider a training set $\{(x_i, y_i)\}_{i=1,\dots,n}$

$x_i \sim \text{Unif}([0, 1])$

$y_i = r_{\lfloor x_i \cdot k \rfloor}$ → lower integral

→ If x is drawn s.t. $\lfloor x \cdot k \rfloor \neq \lfloor x_i \cdot k \rfloor$ for $\forall i$

Then can any learning algorithm predicts $r_{\lfloor x, k \rfloor}$?

→ Any learning algorithm $A(x)$ only depends on R.V. $\{r_j ; j \text{ in the training set}\}$

⇒ Prediction for such x is limited to random guess

$\mathbb{P}(A(x) \neq y \mid X \text{ is } \underline{\text{not}} \text{ in a deserved bracket}) = \frac{1}{2}$

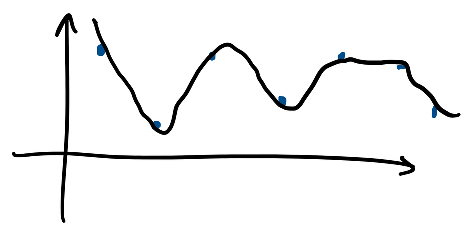
Hence,

$$\mathbb{P}(A(x) \neq y) = \frac{1}{2} \mathbb{P}(X \text{ is not observed})$$
$$\geq \frac{1}{2}\left(1 - \frac{n}{k}\right)$$

for any algorithms $A(\cdot)$

→ Intepretation: there is no iteraction between training and testing

→ Contrast with



because smoothness of this curve creates $\boxed{\text{dependence}}$ between train and test

Machine Learning Paradigms

① Simplest setting (focus of this course): $\boxed{\text{Supervised Learning (SL)}}$

dataset of labeled examples $\{(x_i, y_i)\}$

$x_i \in \mathcal{X}$ → input feature space $\mathcal{X} \stackrel{\text{e.g.}}{=} \{\text{natural images}\}$

$\stackrel{\text{e.g.}}{=} \{\text{text sequences}\}$

$y_i \in \mathcal{Y}$ → label space $\mathcal{Y} = \mathbb{R}$ for regression $\stackrel{\text{e.g.}}{\Rightarrow}$ (predicting temp.)

$\mathcal{Y} = \{1, k\}$ for classification (category)

$\mathcal{Y} = \mathbb{R}^d$ (protein folding) ["structured prediction"]

② Important special case of SL: $\boxed{\text{Self-supervised Learning (SSL)}}$

→ We define the label Y ourselves from unlabeled data.

(video)

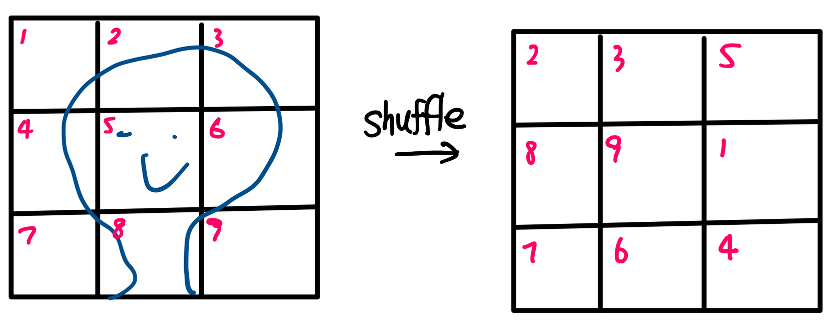
ex1: $Z_1, Z_2, \boxed{Z_3, \dots} \xrightarrow{\text{predict}} \boxed{Z_t}, \dots$ $\boxed{\text{future}}$

Sequential Data — past X_t "input" — present Y_t "label"

"Representation Learning"

Condition: Near future being highly dependent on the presence

ex 2:



Y: Correct ordering

X: set of 9 patches shuffled

③ Unsupervised Learning

→ Consider an unlabeled dataset $D = \{X_i\}$. Extracting "original" information out D

→ Geometric Encoding $X \to \boxed{\Phi} \to Z \to \boxed{\Psi} \to X$ [autoencoder]

Encoder — Latent Space — Decoder

Design encoder / decoder to minimize $\min_{\phi, \psi} \mathbb{E}\|X - \psi(\phi(X))\|^2$

[PCA]

Canonical Ex. $\phi: \mathbb{R}^d \to \mathbb{R}^k$ $k \ll d$ linear maps

$\psi: \mathbb{R}^k \to \mathbb{R}^d$

FML Lecture 3 : Linear Regression I

Recap from Lec. 2

* Impossibility of Learning (NFLT)
* Several paradigms of ML
  - → Supervised Learning
  - → Self-supervised learning
  - → Unsupervised learning

· Unsupervised Learning

"discovering" hidden structure in data

$x \in \mathcal{X} \to \boxed{\Phi} \longrightarrow z \longrightarrow \boxed{\psi} \to \hat{x}$

encode (Φ), layout (z), decode (ψ)

ex. Principal Component Analysis (PCA)

where Φ and ψ are linear maps

• Probabilistic view on unsupervised learning

Input $x_i \in \mathcal{X}$, $i \in \{1, \cdots, n\}$ is viewed as n i.i.d. samples of an unknown prob. distribution P.

↳ Unsupervised learning to estimate $\hat{P}$ from samples $\{x_i\}$

· Main Application : generative modeling use $\hat{P}$ to draw new sample (Dall-E, ChatGPT, …)

Semi-supervised Learning

Large unlabeled dataset $D = \{X_i\}_i$

Small labeled dataset $D' = \{X_i', Y_i'\}_i$

· Assumptions : X_i and X_i' are drawn from some distribution

· Goal : Combine D and D' to "propagate" labels

(If X_i is 'similar' to X_i', then Y_i should be similar to Y_i')

Online and Reinforcement Learning

→ So far learning has been passive

→ Learning is also the ability to act and adapt to changing adversified environment

eg 1 : Bandit Problem

slot 1   slot 2   · · ·   slot K

· Each slot is modeled as a distribution $\mathcal{U}_k$ with $r_j \sim \mathcal{U}_j \in [0, +\infty)$, $j = 1, \cdots, K$

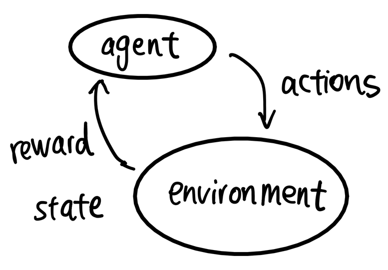
· Each round t the player picks a slot $S_t \in \{1, \cdots, K\}$ (we can pick the same slot repeatedly)

Goal : After T rounds, maximize reward $\sum_{t=1}^{T} r_{S_t}$

Online Learning key aspect : tradeoff between exploration and exploitation

eg. → Allocation of Research Budget

→ Clinical Trials



→ Extension to settings when enivronment depends on an agent

Reinforcement Learning Structure

eg. Games (GO, Chess, Mazes 迷宫)

Robotics

▫ Focus on Supervised Learning

Regression Problems : We are given a dataset $\{(x_i, y_i)\}_{i=1,\cdots,n}$

$x_i \in \mathcal{X}$ inputs (features) $y_i \in \mathcal{Y}$ output (response) $(x_i, y_i) \sim P$ (i.i.d.)

still inside the sample space (i.e., we are doing training only)

· Goal : Given a new $X \in \mathcal{X}$ drawn from P_X, predict $f(X) \approx Y$

· Regression : $\mathcal{Y} = \mathbb{R}$, measure errors using square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$

Q : What is the optimum least squares predictor ?   perdictor $f : \mathcal{X} \to \mathcal{Y}$

find $\min_{f: \mathcal{X} \to \mathcal{Y}} \mathbb{E}_P (f(X) - Y)^2$

eg1 : If $\exists f^*$ s.t. $f^*(y) = x$, then the best predictor is f^* which makes $\ell = 0$

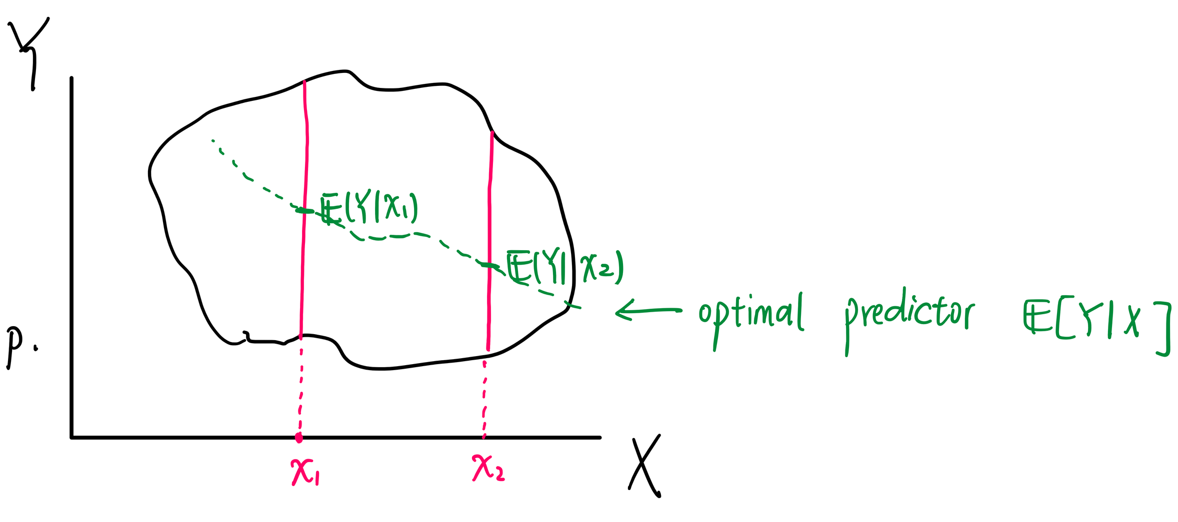
eg2 : If $y \sim \mathcal{N}(0,1)$ and indep. of x. then the best predictor is $f = 0$

$\mathbb{E}_{X,Y}[(f(X) - Y)^2] = \mathbb{E}_X \mathbb{E}_{Y|X}[(f(X) - Y)^2]$

Hence Goal $\Leftrightarrow \min_{c \in \mathbb{R}} \mathbb{E}_Z[(Z - c)^2] \Rightarrow$ take $c = \mathbb{E}Z$ because $\mathbb{E}[(Z - c)^2] = \underbrace{(c - \mathbb{E}[Z])^2}_{\text{bias}} + \underbrace{Var(Z)}_{\text{variance}}$

where $Z = Y | X$

Hence we choose our optimum predictor as $f^*(x) = \mathbb{E}[Y|X]$



Conclusion 1 : Optimal predictor is $f_*(x) = \mathbb{E}_P[Y | X = x]$

Conclusion 2 : This optimal predictor is unknown in general as we don't know p.

Instead, we will find the optimal function in the linear span of a set of predetermined predictors

FML Lecture 4 : Linear Regression II

Recap : Regression Problem $\min_{f: X \to Y} \mathbb{E}[|f(x)-y|^2]$

$\to$ Optimal solution $f^*(x) = \mathbb{E}_p[Y | X = x]$

Linear Regression

f_1
f_2
$\vdots$ candidate solutions
f_d

Regression Model :

$$f_\Theta(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_d f_d(x), \quad \Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{pmatrix} \in \mathbb{R}^d$$

linear combinations of candidate solutions

$\to f_\Theta(x)$ is linear in Θ : $f_{\alpha\Theta + \alpha'\Theta'} = \alpha f_\Theta + \alpha' f_{\Theta'}$

$\to$ But f_Θ is <u>NOT</u> linear w.r.t. x !! (as f_i could be nonlinear)

eg. How bitter is an expresso shot ?

X : barista coffee makers $\quad y$ = acidity level

$f_1(x)$ = temperature of water

$f_2(x)$ = altitude of beans

$f_3(x)$ = pressure

$f_4(x)$ = pressure^3

$\vdots$

- Given observations $x_1, \cdots, x_n$ and candidate solutions $f_1, \cdots, f_d$

Then linear regression is

$$\min_{\Theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \Big[y_i - \underbrace{\sum_{j=1}^d \theta_j f_j(x_i)}_{f_\Theta(x_i)} \Big]^2 = \min_{\Theta \in \mathbb{R}^d} \hat{R}(\Theta) \to \text{Empirical Risk}$$

($\hat{R}$ is a random function)

Collect | features in a matrix $\hat{H} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_d(x_1) \\ f_1(x_2) & \cdots & \cdots & f_d(x_2) \\ \vdots & & \ddots & \vdots \\ f_1(x_n) & \cdots & \cdots & f_d(x_n) \end{bmatrix} \in \mathbb{R}^{n \times d}$

| labels in a vector $\hat{y} = (y_1, \cdots, y_n) \in \mathbb{R}^n$

Then $\boxed{\hat{R}(\Theta) = \frac{1}{n} \| \hat{y} - \hat{H}\Theta \|^2}$

Tasks : (1) How to minimize $\hat{R}(\Theta)$

(2) Geometric Interpretation

(3) Statistical Analysis

(1)
- The Normal Equations

Assumption : The matrix $\hat{H}$ has rank d. In particular, $\boxed{n \geq d}$.

$\hookrightarrow$ direct consequence : The associated Gram matrix $\hat{K} = \frac{1}{n} \hat{H}^T \hat{H} \in \mathbb{R}^{d \times d}$

where $\hat{K}_{i,j} = \frac{1}{n} \langle \hat{H}_i, \hat{H}_j \rangle$ columns

Moreover, $K^T = K$ and K is invertible [Consequence of SVD]

for $\forall z \in \mathbb{R}^d$, $(z \neq 0)$ $z^T \hat{K} z = \frac{1}{n} z^T \hat{H}^T \hat{H} z = \frac{1}{n} \| \hat{H} z \|^2 > 0$ as $\hat{H}$ has rank d

Therefore $\boxed{\hat{K} \text{ is positive-definite.}}$

- Now $\hat{R}(\Theta) = \frac{1}{n} \| \hat{y} - \hat{H}\Theta \|^2$

$$= \frac{1}{n} \|\hat{y}\|^2 + \underbrace{\frac{1}{n} (\hat{H}\Theta)^T \hat{H}\Theta}_{\Theta^T \hat{K} \Theta} - \frac{2}{n} \hat{y}^T \hat{H} \Theta$$

Claim : $\hat{R}$ is convex.

Then $\nabla \hat{R}(\hat{\Theta}) = 2 \hat{K} \hat{\Theta} - \frac{2}{n} (\hat{y}^T \hat{H})^T = 0$ make dimensions coincide

$$\boxed{\Rightarrow \underset{\text{Optimal}}{\hat{\Theta}} = \frac{1}{n} \hat{K}^{-1} \hat{H}^T \hat{y}} = \frac{1}{n^2} (\hat{H}^T \hat{H})^{-1} \hat{H}^T \hat{y}$$

[Normal Equations] Legandre , early 19th century

Associated Risk :

$$\hat{R}(\hat{\Theta}) = \frac{1}{n} \|\hat{y}\|^2 + \frac{1}{n^2} \hat{y}^T \hat{H} \underbrace{(\hat{K}^T)^{-1} \hat{K}}_{I_n} \hat{K}^{-1} \hat{H}^T \hat{y} - \frac{2}{n^2} \hat{y}^T \hat{H} \hat{K}^{-1} \hat{H}^T \hat{y}$$

$$= \frac{1}{n} \hat{y}^T \hat{y} - \frac{1}{n^2} \hat{y}^T \hat{H} \hat{K}^{-1} \hat{H}^T \hat{y}$$

$$= \frac{1}{n} \hat{y}^T \left(I_n - \frac{1}{n} \hat{H} \hat{K}^{-1} \hat{H}^T \right) \hat{y}$$

9/17 FoML 5

Tuesday, September 17, 2024 1:56 PM

FML Lecture 5 : Linear Regression (cont'd) Fixed Design

Recap from last lecture :

Dataset $\{(X_i, Y_i)\}_{i=1,\cdots,n}$, $(X_i, Y_i) \sim P$

feature vector

Linear Regression Model : $f_\Theta(x) = \Theta^T H(x)$ where $H(x) = \big(H_1(x), \cdots, H_d(x)\big) \in \mathbb{R}^d$

$\Theta = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d$

Risk : $R(\theta) = \mathbb{E}\, \|Y - \Theta^T H(x)\|^2$

Empirical Risk : $\hat{R}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \Theta^T H(x_i)\right]^2 = \frac{1}{n}\|\hat{y} - \hat{H}\Theta\|^2$ where $\hat{y} = (y_1, \cdots, y_n) \in \mathbb{R}^n$

Feature Matrix : $\hat{H} = \left[H_j(x_i)\right]_{\substack{j=1,\cdots,d \\ i=1,\cdots,n}} \in \mathbb{R}^{n\times d}$ (We assume $H_1, \cdots, H_j$ are l.i. , i.e., $\hat{H}$ has rank d)

(ordinary) Why ordinary ? (A1)

Least - square Solution : $\hat{\Theta} = \frac{1}{n}\hat{K}^{-1}\hat{H}^T\hat{y}$, where $\hat{K} = \frac{1}{n}\hat{H}^T\hat{H}$

(A1): because the cost function does not have regularization term .

Best - Empirical Risk : $\hat{R}(\hat{\Theta}) = \frac{1}{n}\hat{y}^T\Big[I_n - \underbrace{\tfrac{1}{n}\hat{H}\hat{K}^{-1}\hat{H}^T}_{\Pi = \hat{H}(\hat{H}^T\hat{H})^{-1}\hat{H}^T \in \mathbb{R}^{n\times n}}\Big]\hat{y}$

$\frac{1}{n}\times\left(\frac{1}{n}\right)^{-1} = 1$

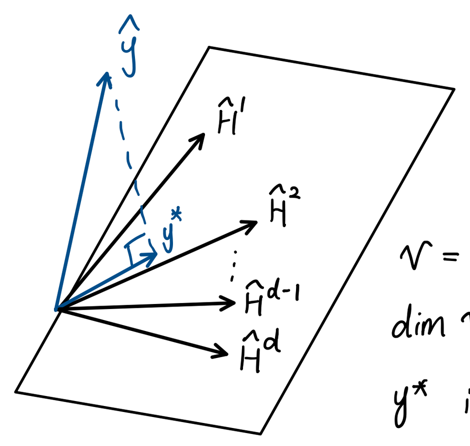
Today : (1) Geometric Interpretation

(2) Statistical Analysis

Let $\Pi = \hat{H}(\hat{H}^T\hat{H})^{-1}\hat{H}^T \in \mathbb{R}^{n\times n}$

Q : How to interpret the OLS (ordinary least square) solution ?

$\hat{H} = \begin{bmatrix} | & & | \\ \hat{H}^1 & \cdots & \hat{H}^d \\ | & & | \end{bmatrix}$ where $\hat{H}^j \in \mathbb{R}^n$, $j = 1, \cdots, d$



Rmk. $y^* = \hat{H}\hat{\Theta}$ as orthogonal projection is the minimizer of the ER. We will show $\Pi\hat{y} = \hat{H}\hat{\Theta}$.

i.e., Π is the orthogonal projector of $\hat{y}$ onto $\mathcal{V}$

$\mathcal{V} = \mathrm{span}(H^1, \cdots, H^d)$

$\dim \mathcal{V} = d$

y^* is the orthogonal projection of $\hat{y}$ onto $\mathcal{V}$

A: $\hat{H}\hat{\Theta} = \Pi\hat{y}$, as the orthogonal projection of $\hat{y}$ onto $\mathrm{Col}(\hat{H})$

pf. $\min_{\Theta\in\mathbb{R}^d} \|\hat{y} - \hat{H}\Theta\|^2 = \min_{v\in\mathcal{V}} \|\hat{y} - v\|^2 = \mathrm{Proj}_{\mathcal{V}}(\hat{y})$

We need to show that Π is the orthogonal projector onto $\mathcal{V}$

Verify Property (1) : Let $x = \hat{H}\Theta$ for some Θ

$\Pi x = \hat{H}(\hat{H}^T\hat{H})^{-1}(\hat{H}^T\hat{H})\Theta = \hat{H}\Theta = x$ ✓

Property (2) : for $x \in \mathcal{V}^\perp \Leftrightarrow x \perp H^j$ for $\forall j = 1, \cdots, d$

i.e., $\mathcal{V}^\perp = \mathrm{null}(H^T)$

$\Rightarrow x \in \mathcal{V}^\perp$, $\hat{H}^T x = 0 \Rightarrow \Pi x = 0$

Hence Π is an orthogonal projector.

#

• Equivalent properties of orthogonal projector P :

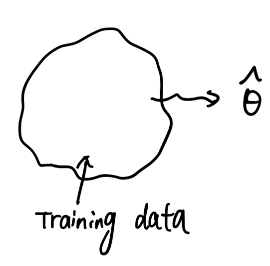
(1) $x \in \mathcal{V}$, $Px = x$

(2) $x \in \mathcal{V}^\perp$, $Px = 0$

Statistical Analysis of Least Square

We distinguish two frameworks, to study generalization in linear regression

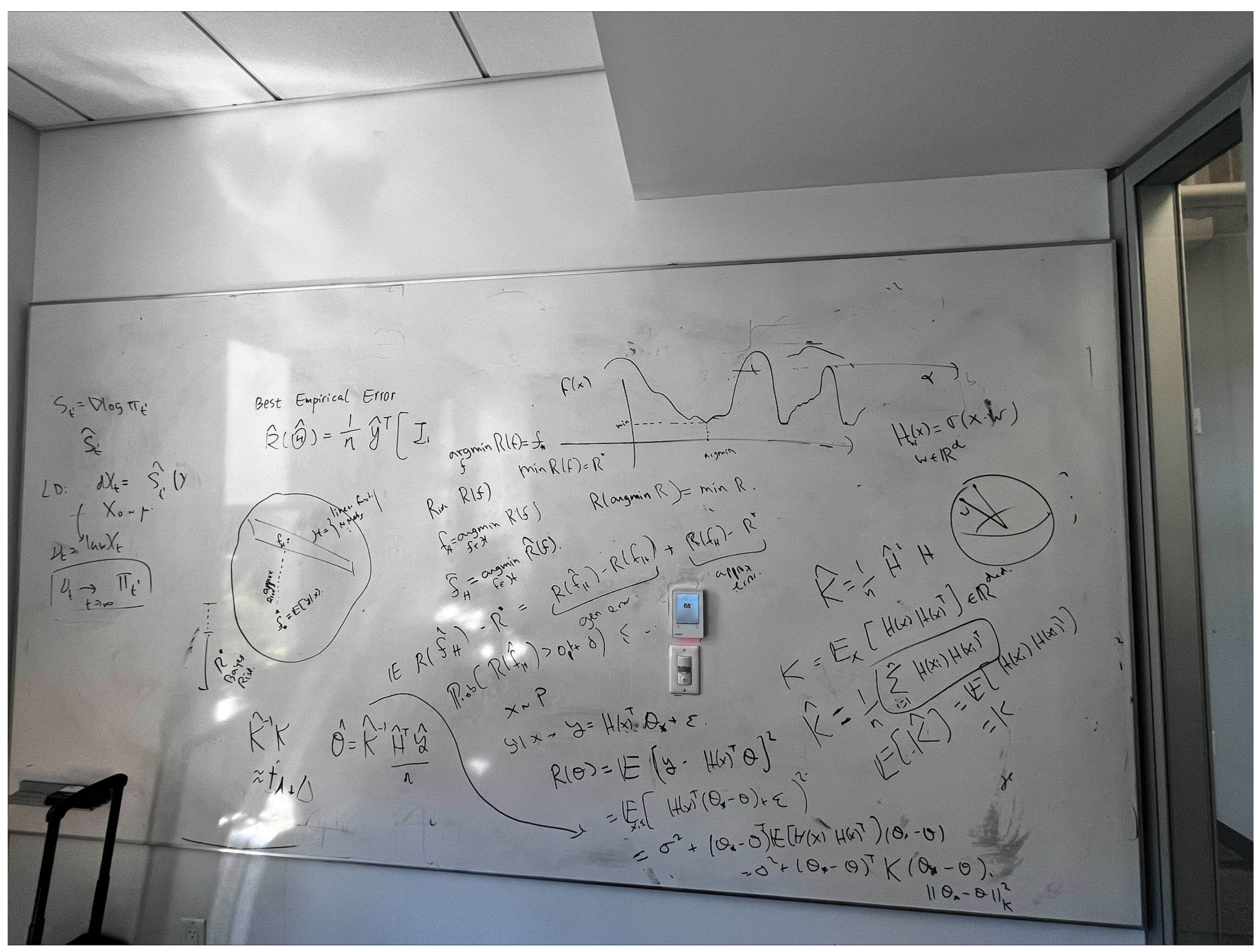
(i) " Random Design " : We view (X_i, Y_i) as a random vector drawn from unknown distribution P



Test time $R(\hat{\Theta}) = \mathbb{E}_P\left[(y - \Theta^T H(x))^2\right]$

$(X,Y)\sim P$ (Random Setting)

Talk a bit more on this ? ⇒



(ii) " Fixed Design " : We view input features x_i as fixed , but outputs y_i still random

Eg. Coffee Experiment : same barista /machines but perhaps different observed conditions

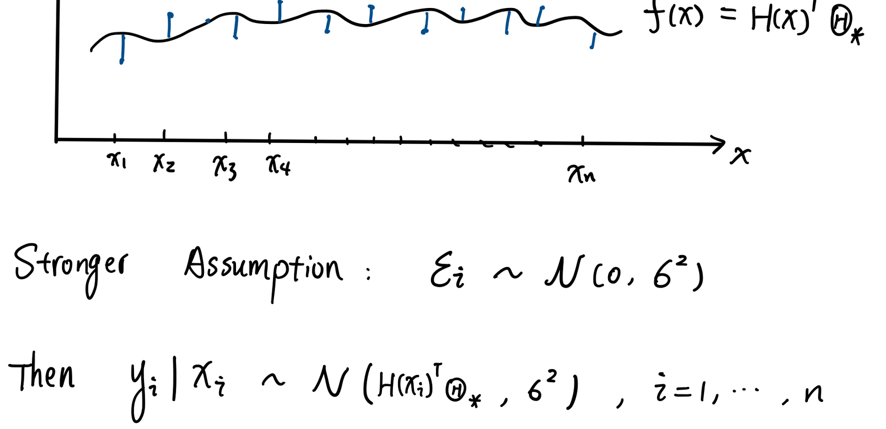
Focus on fixed design setting :

• As before, we assume that $\hat{H} \in \mathbb{R}^{n\times d}$ has rank d (hence $\hat{K}$ is invertible)

• We also suppose that outputs are generated using

$\mathbb{R}^d$, row vector of $\hat{H}$, representing a feature vector of one sample

$y_i = \underbrace{H(x_i)^T\boxed{\Theta_*}}_{\substack{\text{deterministic} \\ \text{"signal"}}} + \underbrace{\varepsilon_i}_{\substack{\text{random} \\ \text{noise}}}$ such that $\begin{cases} \mathbb{E}[\varepsilon_i] = 0 \\ \mathrm{Var}(\varepsilon_i) = \sigma^2 \text{ for } \forall i = 1, \cdots, n \\ \varepsilon_1, \cdots, \varepsilon_n \text{ are i.i.d} \end{cases}$



→ Stronger Assumption : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Then $y_i \mid x_i \sim \mathcal{N}(H(x_i)^T\Theta_*, \sigma^2)$, $i = 1, \cdots, n$

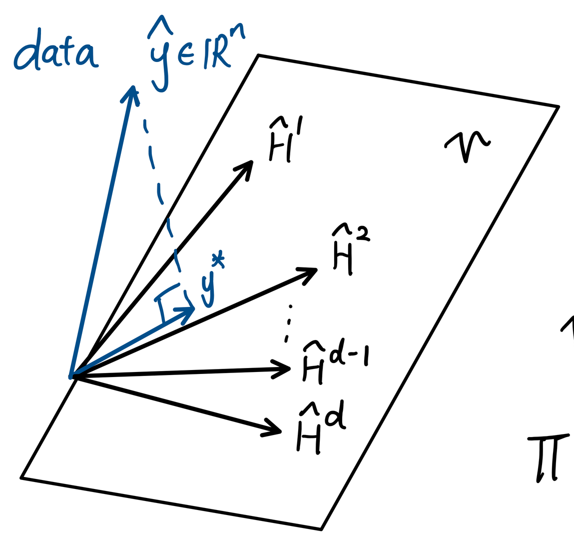
9/19 FoML 6

Thursday, September 19, 2024 1:58 PM

FML Lecture 6 : Linear Regression (cont'd) : Statistical Analysis

Recap from L5 :

- Geometric View



$\hat{H}^i$: feature vectors

$\mathcal{V} = \text{span}(H^1, \cdots, H^d)$

$\Pi y^* = \hat{H}^T \hat{\Theta}$ solution of OLS

- Genaralization

error of Least Square $\begin{cases} * \text{ "Fixed Design" : input } X_i \text{ fixed, } y_i \text{ random} \\ * \text{ Random Design : } (X, Y) \sim p \end{cases}$

Today (1) Analyze LS on fixed design setting

(2) Regularization

- Fixed Design : $x_1, \cdots, x_n$ fixed s.t. $\hat{H} \in \mathbb{R}^{n \times d}$ is rank d

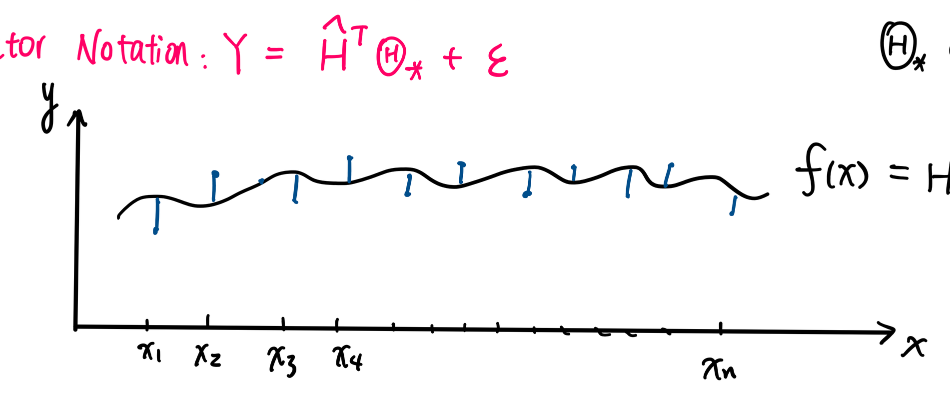
→ Assumption : labels $\{y_i\}$ are generated according to

for $i = 1, \cdots, n$, $y_i = \underbrace{\hat{H}(x_i)^T \Theta_*}_{\text{"signal"}} + \underbrace{\varepsilon_i}_{\text{"noise"}}$, $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. R.V.'s

with $\mathbb{E}[\varepsilon] = 0$, $Var(\varepsilon) = \sigma^2 > 0$

Vector Notation : $Y = \hat{H}^T \Theta_* + \varepsilon$

$\Theta_* \in \mathbb{R}^d$ This is what we need to figure out (this is the underlying structure)



One thing ambiguous : How do we determine whether the deviation is because of the lack of parameters or the random noises (Our Assumption)

↳ Stronger Assumption : assume $\varepsilon_i \sim \mathcal{N}(H(X_i)^T \Theta_*, \sigma^2)$ (Gaussian Noise)

→ Goal : Estimate Θ_* (God's Parameters)

- Risk Decomposition

Def. For any $\Theta \in \mathbb{R}^d$, the generalization error is $\boxed{R(\Theta) = \frac{1}{n} \mathbb{E}_Y \|Y - \hat{H}\Theta\|^2}$ (Note : R is NOT random)

Prop. Under the linear model assumption, $\boxed{R(\Theta) = \sigma^2 + (\Theta - \Theta_*)^T \hat{K} (\Theta - \Theta_*)}$

Rmk. $R(\Theta)$ takes minimum at $\underset{\text{(unique)}}{\Theta = \Theta^*}$ of σ^2

pf. $R(\Theta) = \frac{1}{n} \mathbb{E}_Y \|Y - \hat{H}\Theta\|^2$

$\overset{\text{model assumption}}{=} \frac{1}{n} \mathbb{E}_\varepsilon \|\hat{H}(\Theta_* - \Theta) + \varepsilon\|^2$

$= \frac{1}{n}\left[\mathbb{E}_\varepsilon \|\varepsilon\|^2 + (\Theta_* - \Theta)^T \hat{H}^T \hat{H} (\Theta_* - \Theta) + 2\,\underbrace{\mathbb{E}_\varepsilon[\langle \varepsilon, \hat{H}(\Theta_* - \Theta) \rangle]}\right]$

$= \frac{1}{n} \mathbb{E}_\varepsilon\left[\sum_{i=1}^n \varepsilon_i^2\right] + (\Theta_* - \Theta)^T \hat{K} (\Theta_* - \Theta)$ $\quad \langle \underbrace{\mathbb{E}_\varepsilon[\varepsilon]}_{0}, \hat{H}(\Theta_* - \Theta) \rangle$

#

Next : Evaluate $\mathbb{E}[R(\hat{\Theta})]$ when $\hat{\Theta}$ is the OLS estimator
(training data)

- Bias - Variance Decomposition

Sps. that $\hat{\Theta}$ is a Random Vector (any random vector), for example $\hat{\Theta} = \hat{\Theta}_{OLS}$, the OLS estimator (it is random because $\hat{\Theta}_{OLS} = \frac{1}{n} \hat{K}^{-1} \hat{H} \hat{y}$ and $\hat{y}$ is random, $\hat{H}$ is fixed)

Then $\mathbb{E}_{\hat{\Theta}}[R(\hat{\Theta})] = \underbrace{\sigma^2}_{\text{Bayes Risk}} + \underbrace{\|\mathbb{E}[\hat{\Theta}] - \Theta_*\|_{\hat{K}}^2}_{\text{Bias}} + \underbrace{\mathbb{E}[\|\hat{\Theta} - \mathbb{E}[\hat{\Theta}]\|_{\hat{K}}^2]}_{\text{Variance}}$

Bayes Risk: Smallest possible error that any method suffers

Def. $\|x\|_{\hat{K}}^2 = x^T \hat{K} x$ (In particular, $\|x\|_{I_d}^2 = \|x\|^2$)

pf. By RD, $R(\hat{\Theta}) = \sigma^2 + (\hat{\Theta} - \Theta_*)^T \hat{K} (\hat{\Theta} - \Theta_*)$ (1)

Also, $\mathbb{E}[(\hat{\Theta} - \Theta_*)^T \hat{K} (\hat{\Theta} - \Theta_*)] = \mathbb{E}\left[(\underbrace{\hat{\Theta} - \mathbb{E}[\hat{\Theta}]}_{\Theta_A \text{ (random)}} + \underbrace{\mathbb{E}[\hat{\Theta}] - \Theta_*}_{\Theta_B \text{ (constant)}})^T \hat{K} (\underbrace{\hat{\Theta} - \mathbb{E}[\hat{\Theta}]}_{\Theta_A} + \underbrace{\mathbb{E}[\hat{\Theta}] - \Theta_*}_{\Theta_B})\right]$

$= \underbrace{\mathbb{E}[\Theta_A^T \hat{K} \Theta_A]}_{\mathbb{E}\|\hat{\Theta} - \mathbb{E}[\hat{\Theta}]\|_{\hat{K}}^2} + \underbrace{\mathbb{E}[\Theta_A^T \hat{K} \Theta_B]}_{\hat{K}\Theta_B \underbrace{\mathbb{E}[\Theta_A^T]}_{0}} + \underbrace{\mathbb{E}[\Theta_B^T \hat{K} \Theta_A]}_{\Theta_B^T \hat{K} \underbrace{\mathbb{E}[\Theta_A]}_{0}} + \underbrace{\mathbb{E}[\Theta_B^T \hat{K} \Theta_B]}_{\Theta_B^T \hat{K} \Theta_B = \|\mathbb{E}[\hat{\Theta}] - \Theta_*\|_{\hat{K}}^2}$

$= \underbrace{\|\mathbb{E}[\hat{\Theta}] - \Theta_*\|_{\hat{K}}^2}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\|\hat{\Theta} - \mathbb{E}[\hat{\Theta}]\|_{\hat{K}}^2\right]}_{\text{Variance}}$ (2)

Take $\mathbb{E}(\cdot)$ on both sides of (1) and use (2), we complete the proof.

#

→ Now let's plug in our OLS estimator : $\hat{\Theta} = \frac{1}{n} \hat{K}^{-1} \hat{H}^T \hat{y}$ where $\hat{y} = \hat{H}\Theta_* + \varepsilon$

$\hat{\Theta} = \hat{\Theta}_{OLS}$

$= \hat{K}^{-1} \frac{\hat{H}^T}{n} (\hat{H}\Theta_* + \varepsilon)$

$= \Theta_* + \frac{1}{n} \hat{K}^{-1} \hat{H}^T \varepsilon$

Therefore, $\mathbb{E}[\hat{\Theta}] = \Theta_*$ (3)

⇒ Bias Term : $\|\mathbb{E}[\hat{\Theta}] - \Theta_*\|_{\hat{K}}^2 = 0$. [OLS is unbiased!]

⇒ Variance Term : $\mathbb{E}\left[\|\hat{\Theta} - \mathbb{E}[\hat{\Theta}]\|_{\hat{K}}^2\right] \overset{(3)}{=} \mathbb{E}\left[\|\hat{\Theta} - \Theta_*\|_{\hat{K}}^2\right] = \sigma^2 \cdot \frac{d}{n}$ (to be cont'd next class)

Q1 : Will regularization makes $\hat{\Theta}$ biased ?

A1 : Yes !

9/24 FoML 7

Tuesday, September 24, 2024 10:57 AM

FML Lecture 7 : Linear Regression : Regularisation

- Recap from last week :

1) Risk of any LS estimator Θ in the fixed design

$$R(\Theta) = \sigma^2 + (\Theta - \Theta_*)^T \hat{K} (\Theta - \Theta_*)$$

2) When $\hat{\Theta}$ is a random vector (eg. $\hat{\Theta} = \hat{\Theta}_{OLS}$).

$$\mathbb{E}[R(\hat{\Theta})] = \sigma^2 + \| \mathbb{E}[\hat{\Theta}] - \Theta_* \|_{\hat{K}}^2 + \mathbb{E}\left[\| \hat{\Theta} - \mathbb{E}[\hat{\Theta}] \|_{\hat{K}}^2 \right]$$

$$\hat{\Theta}_{OLS} = \hat{K}^{-1} \frac{\hat{H}^T \hat{y}}{n} = \Theta_* + \hat{K}^{-1} \frac{\hat{H}^T \varepsilon}{n} \Rightarrow \hat{\Theta}_{OLS} \text{ is unbiased}$$

- Let's compute variance

$$\mathbb{E}\left[\| \hat{\Theta}_{OLS} - \Theta_* \|_{\hat{K}}^2 \right] = \mathbb{E}\left[\left(\frac{\varepsilon^T \hat{H}}{n} \right) \hat{K}^{-1} \hat{K} \hat{K}^{-1} \left(\frac{\hat{H}^T \varepsilon}{n} \right) \right]$$

$$= \frac{1}{n^2} \mathbb{E}[\varepsilon^T \underbrace{\hat{H} \hat{K}^{-1} \hat{H}^T}_{n\Pi} \varepsilon] = \frac{1}{n} \mathbb{E}[\varepsilon^T \Pi \varepsilon] \quad \text{for } \varepsilon \in \mathbb{R}^n, \ \Pi \in \mathbb{R}^{n \times n}$$

where $\Pi = \hat{H}(\hat{H}^T \hat{H})^{-1} \hat{H}^T$ is an orthogonal projection onto $\mathcal{V} = Col(H)$

$$= \frac{1}{n} \mathbb{E}\left[\sum_{i,j} \varepsilon_i \varepsilon_j \Pi_{ij} \right] = \frac{1}{n} \sum_{i,j} \underbrace{\mathbb{E}[\varepsilon_i \varepsilon_j]}_{\begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}} \Pi_{ij} = \frac{1}{n} \sum_i \sigma^2 \Pi_{ii} = \frac{\sigma^2}{n} \cdot Tr(\Pi)$$

$$= \frac{\sigma^2}{n} \cdot Tr\left(\hat{H} (\hat{H}^T \hat{H})^{-1} \hat{H}^T \right) \quad \hat{H} \in \mathbb{R}^{n \times d}$$

A, B squared or some generalized condition
$Tr(AB) = Tr(BA)$

$$= \frac{\sigma^2}{n} \cdot Tr\left(\underbrace{(\hat{H}^T \hat{H})^{-1}}_{\mathbb{R}^{d \times d}} \hat{H}^T \hat{H} \right)$$

need stronger linear algebra tool

$$= \frac{\sigma^2}{n} \cdot Tr(I_d) = \sigma^2 \cdot \frac{d}{n}.$$

#

Conclusion : Variance $\mathbb{E}\left[\| \hat{\Theta} - \mathbb{E}[\hat{\Theta}] \|_{\hat{K}}^2 \right] = \sigma^2 \cdot \frac{d}{n}$

$$\Rightarrow \mathbb{E}[R(\hat{\Theta})] = \boxed{\sigma^2} + \boxed{\sigma^2 \cdot \frac{d}{n}}$$

$\rightarrow$ excess risk , goes to 0 as $n \nearrow +\infty$
[Variance]
[No bias]

↳ "Incompressible" error

Q1 : What happens when $n \approx d$, even $n < d$?

Q2 : What happens beyond the fixed design setting ?

- Regularisation

  - Motivation :

$\rightarrow$ When $n = d$, the normal equations $\hat{K} \hat{\Theta} = \frac{\hat{H}^T y}{n}$ — n equations, n unknowns

Assuming $\hat{K}$ invertible , unique solution $\hat{\Theta}$ with error $\hat{R}(\hat{\Theta}) = 0$

↳ We are memorizing data (which often includes noise)

$\rightarrow$ When $n < d$, $\hat{K}$ is not invertible ! (under - determined)

Very common regime (gene expression , $n \ll d$)

- Solution : Add some "friction" , "cost" to using features

  $\Leftrightarrow$ Explain the data using cheapest option

  $\Leftrightarrow$ Occam's Razor

Θ_1 Q : How to define a useful notion of cost ?

Θ_2 Def. A $\boxed{\text{ridge regularisation}}$ considers the cost as L^2-norm : $\|\Theta\|_2^2 = \sum_{i=1}^n \Theta_i^2$

Θ_3 Note : An alternative is to consider the L^1 norm $\|\Theta\|_1 = \sum_i |\Theta_i|$

↳ this leads to $\boxed{\text{sparse regression}}$

Question : What is the essential difference of changing L^2- norm to L^1- norm

- Ridge Regularisation

$$\hat{R}_\lambda(\Theta) = \frac{1}{n} \| \hat{y} - \hat{H}\Theta \|^2 + \underbrace{\boxed{\lambda}}_{\text{regularisation strength}} \|\Theta\|_2^2 , \quad \lambda \geq 0$$

$\rightarrow$ Minimize $R_\lambda(\Theta)$ $= \frac{1}{n}\|\hat{y}\|^2 - 2\frac{\hat{y}^T \hat{H} \Theta}{n} + \Theta^T \hat{K} \Theta + \lambda \Theta^T \Theta$

$$\nabla_\Theta \hat{R}_\lambda(\Theta) = 0 \Leftrightarrow -\frac{2}{n} \hat{H}^T \hat{y} + 2\hat{K}\Theta + 2\lambda\Theta = 0$$

$$\Leftrightarrow -\frac{1}{n} \hat{H}^T \hat{y} + [\hat{K} + \lambda I_n] \Theta = 0$$

$$\Rightarrow \hat{\Theta}_\lambda = [\hat{K} + \lambda I_n]^{-1} \frac{\hat{H}^T \hat{y}}{n}$$

$\rightarrow$ Now we observe that $[\hat{K} + \lambda I_n]$ is always invertible for $\lambda > 0$

This is because $\hat{K}$ is symmetric positive semi-definite , so $y^T \hat{K} y \geq 0$ for all $y \in \mathbb{R}^n$

hence $\hat{y}^T (\hat{K} + \lambda I_n) \hat{y} = \hat{y}^T \hat{K} \hat{y} + \lambda \|\hat{y}\|_2^2 > 0$

hence $\hat{K} + \lambda I_n$ is symmetric positive definite.

$\rightarrow$ Compare the generalisation error of $\hat{\Theta}_\lambda$ in the fixed design setting , linear model $y_i = H(x_i)^T \Theta_* + \varepsilon_i$

$$\mathbb{E}[R(\hat{\Theta}_\lambda)] = \sigma^2 + \underbrace{\boxed{\lambda^2 \Theta_*^T (\hat{K} + \lambda I_n)^{-1} \hat{K} \Theta_*}}_{\text{Bias Term}} + \underbrace{\boxed{\frac{\sigma^2}{n} Tr\left[\hat{K}^2 (\hat{K} + \lambda I_n)^{-2} \right]}}_{\text{Variance Term}}$$

Q : Is the variance term always reduced ? How should we choose λ ?

9/26 FoML 8

Thursday, September 26, 2024 2:03 PM

[TA] Solving Normal Equations

- Iterative Method
- SVD
- QR Decompositive
- Cholesky Decomposition

→ Linear Regression

$X\beta = y$

$X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$

n : # of samples

d : # of features

$n > d$ (otherwise no error)

Goal : $\hat{\beta} = \arg\min_{\beta} \|X\beta - y\|^2$

$= (X^TX)^{-1}X^Ty \in \mathbb{R}^d$

$\Leftrightarrow (X^TX)\beta = X^Ty \in \mathbb{R}^d$

Normal Equation

→ Iterative Method

$\hat{R} = \|X\beta - y\|^2 \in \mathbb{R}$

$= (X\beta - y)^T(X\beta - y)$

$\nabla_\beta \hat{R} = 2X^T(X\beta - y) \in \mathbb{R}^d$

- Gradient Descent.

for $\forall t \geq 0$, $\beta_{t+1} \leftarrow \beta_t - \underbrace{\eta}_{\text{learning rate}} \cdot \nabla_\beta \hat{R}(\beta_t)$

$= \beta_t - \eta \cdot 2X^T(X\beta_t - y)$

Advantage :

① easy to understand and implement

② scalable (Stochastic Gradient Method)

Disadvantage :

① might be slow to converge

Hessian : $\nabla_\beta(\nabla_\beta \hat{R}) = 2X^TX \in \mathbb{R}^{d\times d}$

and min eigenvalue of $X^TX \approx 0$ is possible

[make η as $\eta(\lambda_t)$, Adaptive Gradient Descent]

→ SVD : Singular Value Decomposition

Decompose any matrix $A = U\Sigma V^T \in \mathbb{R}^{n\times d}$ where $\overbrace{U \in \mathbb{R}^{n\times n},\ V \in \mathbb{R}^{d\times d}}^{\text{orthogonal}}$, $\overbrace{\Sigma \in \mathbb{R}^{n\times d}}^{\text{(rectangular) diagonal matrix}} = \begin{bmatrix} * & & & 0 \\ & * & & \\ & & * & \\ 0 & \cdots & & 0 \end{bmatrix}$

Change $X \leftarrow A$

$(A^TA)x = A^Tb$

$A = U\Sigma V^T$

$\Rightarrow V\Sigma^T\Sigma V^T x = V\Sigma^T U^T b$

$\Rightarrow x = V\Sigma^+ U^T b$

where Σ^+ : pseudo - inverse (Search)

Advantage :

① Handle ill-conditioned problem : $\dfrac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}$

Disadvantage :

① Expensive computation

QR - decomposition :

Decompose any $A = QR \in \mathbb{R}^{n\times d}$

Q : orthogonal matrix $\mathbb{R}^{n\times d}$

R : upper triangle $\mathbb{R}^{d\times d}$

FML Lecture 9 : Principles of Supervised Learning

Reminder : Office Hours (Joan) tmr. Wed. @ 2 p.m. (612 CDS)

Today : → From fixed to random design

→ Main elements of supervised learning

Limitations of fixed design : We need to predict outside training set

(linear)

Random Design : Training $\{x_i\}_{i\le n}$ i.i.d. from P

$y_i = H(x_i)^T \Theta_* + \varepsilon_i$, ε_i i.i.d. , $\begin{cases} \mathbb{E}[\varepsilon_i]=0 \\ \mathbb{E}[\varepsilon_i^2]=\sigma^2 \end{cases}$

But now we evaluate it on a new point $x \sim P$ (indep. of $\{x_i\}_{i\le n}$)

- Given any $\Theta \in \mathbb{R}^d$, $R(\Theta) = \mathbb{E}_P\left[(H(x)^T\Theta - y)^2 \right]$

$$= \underbrace{\mathbb{E}_{x,\varepsilon}}\left[\left(H(x)^T(\Theta-\Theta_*) - \varepsilon \right)^2 \right]$$

↳ main difference w.r.t. fixed design

$$= \sigma^2 + \mathbb{E}_x\left[(\Theta-\Theta_*)^T H(x) H(x)^T (\Theta-\Theta_*) \right]$$

$$= \sigma^2 + (\Theta-\Theta_*)^T \underbrace{\mathbb{E}_x\left[H(x) H(x)^T \right]}_{K \in \mathbb{R}^{d\times d}} (\Theta-\Theta_*) \quad , \quad \Theta \in \mathbb{R}^d$$

$$= \sigma^2 + \|\Theta - \Theta_*\|_K^2$$

→ The only difference w.r.t. fixed design is that we have K instead of $\hat{K} = \frac{1}{n}\hat{H}^T\hat{H} = \frac{1}{n}\sum_{i=1}^n H(x_i)H(x_i)^T$,

$K = \int H(x)H(x)^T \, P(x)\,dx$ (continuous version of $\hat{K}$)

→ Now we view $\hat{K}$ as the sample version of K

Let's now plug $\Theta = \hat{\Theta}_{OLS} = \hat{K}^{-1}\frac{\hat{H}^T\hat{y}}{n} = \Theta_* + \hat{K}^{-1}\frac{\hat{H}^T\varepsilon}{n}$

$$\mathbb{E}_{x,\varepsilon} R(\hat{\Theta}_{OLS}) = \sigma^2 + \mathbb{E}\left[\frac{1}{n^2} \overbrace{\varepsilon^T \hat{H} \hat{K}^{-1} K \hat{K}^{-1} \hat{H}^T \varepsilon}^{\mathbb{R}} \right]$$

$$= \sigma^2 + \frac{1}{n^2}\mathbb{E}\Big[\underbrace{\mathrm{Tr}(\varepsilon^T \hat{H}\hat{K}^{-1} K \hat{K}^{-1}\hat{H}^T\varepsilon)}_{\mathrm{Tr}(\varepsilon\varepsilon^T\hat{H}\hat{K}^{-1}K\hat{K}^{-1}\hat{H}^T)} \Big]$$

Rmk. We can interchange $\mathbb{E}[\cdot]$ & $\mathrm{Tr}(\cdot)$ as they are both linear.

$$= \sigma^2 + \frac{1}{n^2}\underbrace{\mathbb{E}[\mathrm{Tr}(\varepsilon\varepsilon^T)]}_{\sigma^2} \cdot \mathbb{E}\Big[\underbrace{\mathrm{Tr}(\hat{H}\hat{K}^{-1}K\hat{K}^{-1}\hat{H}^T)}_{\mathrm{Tr}(\underbrace{\hat{H}^T\hat{H}}_{n\hat{K}}\hat{K}^{-1}K\hat{K}^{-1})} \Big]$$

$$= \sigma^2 + \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}(K\hat{K}^{-1}) \right]$$

→ We need to understand the inverse of a random matrix $\hat{K}$

This requires tools from Random Matrix Theory

Q: Does regularisation still make sense in random design setting

A: Yes, as $\hat{K}$ might be poorly-conditioned and we should regularize on that.

Main Results we have seen :

- $\hat{\Theta}_{OLS}$ "best" model that fits data
- How to assess model outside training

Now let's describe general picture

→ Model for data : Training data n i.i.d. samples

$\{(x_i, y_i)\}_{i=1,\dots,n}$ drawn from unknown distribution P in $\mathcal{X}\times\mathcal{Y}$ ($x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$)

Test data $(x,y) \sim P$ independent of training

Strong Assumption : Same distribution as training data

(when test distribution $P_{test} \neq P_{train}$, we have different problem : transfer learning)

→ Loss function : a function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_+$

that measures agreement between true and predicted label

Ex : $\ell(y,y') = (y-y')^2$ (LS regression)

$\ell(y,y') = 1_{y\neq y'}$ (Classification)

→ Now, given any mapping $f : \mathcal{X}\to\mathcal{Y}$,

its (expected future) performance is :

$R(f) = \mathbb{E}_{(x,y)\sim P}\left[\ell(f(x), y) \right]$ (population risk / generalization error)

Ex. $R(\Theta) = \mathbb{E}\left[(H(x)^T\Theta - y)^2 \right]$ (least square in random setting)

Rmk. Now everything is on a random design

→ From population risk, we can define the optimum predictor :

$$f_* = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} R(f)$$

↳ Recall in LS setting, $f_*(x) = \mathbb{E}_P[Y|X=x]$ (Lec. 3)

↳ f_* is the $\boxed{\text{Bayes Predictor}}$, $R^* = R(f_*)$ is called $\boxed{\text{Bayes Risk / Bayes Risk}}$

↳ We can have $R^* > 0$ in general ($R^* = \sigma^2$ in the linear model)

$\boxed{\text{• Bayes Risk is unattainable in general}}$

→ At least two reasons :

(1) It requires knowledge of data distribution P !

(2) f_* might be arbitrarily crazy function → hard to even approximate !

10/3 FoML 10

Thursday, October 3, 2024 2:02 PM

FML Lecture 10 : Elements of SL

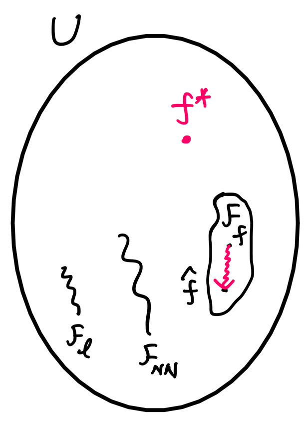
Recall : $f : \mathcal{X} \to \mathcal{Y}$ (input $\mathcal{X}$, output $\mathcal{Y}$)

$R(f) = \mathbb{E}_{(X,Y)\sim P}\left[\ell(f(x), y)\right]$ Population Risk

↳ (f^*, R^*) Bayes Predictor / Risk ($f^* = \arg\min R$, $R^* = \min R$)

→ Unpractical $\begin{cases} \cdot \text{ They depend on population} \\ \cdot \text{ They can be arbitarily complex} \end{cases}$

Instead in SL, we focus our attention on a hypothesis class $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathcal{Y},\ \theta \in \Theta\}$



Ex. $\mathcal{F}_\ell = \left\{ f_\theta(x) = H(x)^T \theta,\ \theta \in \Theta = \mathbb{R}^d,\ x \in \mathcal{X} = \mathbb{R}^d \right\} \subset U = \{ f : \mathbb{R}^d \to \mathbb{R} \}$ (linear hypothesis class), $\dim \mathcal{F}_\ell = d$

$\mathcal{F}_{NN} = \{ f_\theta(x) = \underbrace{6_L(w_L \cdots 6_1(w_1 x))}_{L \text{ layers}},\ \theta \in \Theta = \{w_1, \cdots, w_L\} \} \subset U$ $\dim \mathcal{F}_{NN} = L$

→ Now we can consider the best predictor in $\mathcal{F}$

$\bar{f} = \arg\min_{f \in \mathcal{F}} R(f)$

→ $\inf_{f \in \mathcal{F}} R(f) - R^* \geq 0$ measures how accurate the hypothesis space is for our prediction task

Approximation Error / Risk

→ It is still impossible to find $\bar{f}$ (P unknown)

→ Instead, we can consider minimizing the $\boxed{\text{Empirical Risk}}$:

$\hat{R} = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$ where $(x_i, y_i) \overset{\sim}{\text{i.i.d.}} P$

Since $\{(x_i, y_i)\}_{i=1}^n$ is a Random Sample, $\hat{R}$ is a Random Functional

Q1: What is the mean of $\hat{R}(f)$ for any f ?

A: $\mathbb{E}[\hat{R}(f)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)\right] \underset{(x_i,y_i)\ \text{i.i.d.}\ P}{=\!=\!=} \mathbb{E}_P\left[\ell(f(x_i), y_i)\right] = R(f)$

i.e., $\hat{R}$ is an unbiased estimator of R

→ $\{ Z_i = \ell(f(x_i), y_i) \}_{i=1}^n$ are i.i.d. R.V.'s

$Z = \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} Z_i$, where $\mathbb{E} Z = R(f)$, $\operatorname{Var}(Z) = 6_f^2$

under mild moment assumptions, the sample mean is asymptotically normal (n large)

[CLT]: $\sqrt{\frac{n}{6_f^2}}\left(\hat{R}(f) - R(f)\right) \xrightarrow{d} \mathcal{N}(0,1)$

→ for large n and fixed f, $|\hat{R}(f) - R(f)| \simeq \frac{6_f}{\sqrt{n}}$

prob. of being larger than $\frac{6_f}{\sqrt{n}}$ is exponentially smaller

↳ can be formulized in the non-asymptotic setting using Concentration Inequality

(finite n)

We can define

$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$ Empirical Risk Minimization (ERM)

↳ Look for hypothesis in our class that best fits the training data

↳ Now, we have reduced learning to solving an optimization problem

Q: How to control the quality of ERM ?

i.e., control generalization gap $R(\hat{f}) - R^*$

For any f, $R(f) = \hat{R}(f) + \left(R(f) - \hat{R}(f)\right)$ [tautology]

So, if we want LHS to be small, we can "hope" to have:

$\begin{cases} \hat{R}(f) \text{ small, and} \\ R(f) - \hat{R}(f) \text{ also small} \end{cases}$

ERM is designed to minimize $\hat{R}(f)$, then what about $R(f) - \hat{R}(f)$

- Key Observation: there is an inherent tension between the two terms

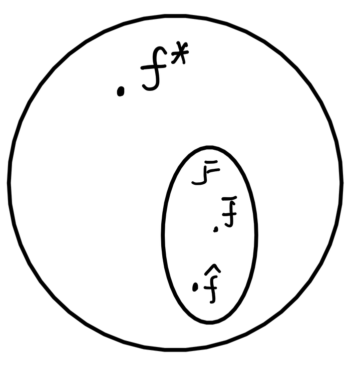
$\hat{f}$ = ERM

$\hat{R}(\hat{f})$: decreases as $\mathcal{F}$ gets bigger

but

$R(\hat{f}) - \hat{R}(\hat{f})$ might increase as $\mathcal{F}$ gets bigger

Decomposition of Risk:



Consider $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$ (ERM)

$R(\hat{f}) - R^* = \underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R^*}_{\mathcal{E}_A = \text{approximation error}}$

$= R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - R(\bar{f}) + \mathcal{E}_A$, where $\bar{f} = \arg\min_{f \in \mathcal{F}} R(f)$

$\leq R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\bar{f}) - R(\bar{f}) + \mathcal{E}_A$, as $\hat{R}(\hat{f}) \leq \hat{R}(\bar{f})$ by def. of $\hat{f}$

$\leq |R(\hat{f}) - \hat{R}(\hat{f})| + |R(\bar{f}) - \hat{R}(\bar{f})| + \mathcal{E}_A$, by triangular inequality

$\leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| + \mathcal{E}_A$

FML Lecture 11: Decomposition of Risk

Recap: $R(f)$: Expected Risk $:= \mathbb{E}_P[\ell(f(x), y)]$

$\hat{R}(f)$: Empirical Risk $:= \frac{1}{n}\sum_{i=1}^{n} \ell(f(x), y)$

- $\mathbb{E}[\hat{R}(f)] = R(f)$ ($\hat{R}$ is an unbiased estimator of R)
- $|\hat{R}(f) - R(f)| \sim \frac{\sigma_f}{\sqrt{n}}$ | $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \hat{R}(f)$ Empirical Risk Minimization

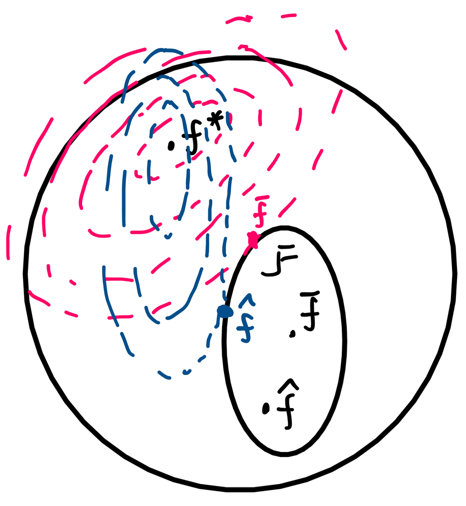
  hypothesis class

- ML "Tautology": for $\forall f$, $R(f) = \underbrace{\hat{R}(f)}_{\text{"under control"}} + \underbrace{(R(f) - \hat{R}(f))}_{\text{estimation}}$

§ Decomposition of Risk

Consider $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \hat{R}(f)$

Goal: Control excess risk: $R(\hat{f}) - \underline{R^*}$ Bayes Risk



$$R(\hat{f}) - R^* = \underbrace{R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R^*}_{\varepsilon_A = \text{approximation error}}$$

$= R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - R(\bar{f}) + \varepsilon_A$, where $\bar{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, R(f)$

$\le R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\bar{f}) - R(\bar{f}) + \varepsilon_A$, as $\hat{R}(\hat{f}) \le \hat{R}(\bar{f})$ by def. of $\hat{f}$

$\le |R(\hat{f}) - \hat{R}(\hat{f})| + |R(\bar{f}) - \hat{R}(\bar{f})| + \varepsilon_A$, by triangular inequality

$\le \underbrace{2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|}_{\varepsilon_S = \text{Statistical error}} + \varepsilon_A$ where $R(f) = \mathbb{E}[\hat{R}(f)]$

"Rule of Thumb":

→ "Small" hypothesis space $\mathcal{F}$: ε_A dominates over ε_S

→ "Large" hypothesis space $\mathcal{F}$: ε_S dominates over ε_A

→ Instance of $\underset{\varepsilon_A}{\underline{\text{bias}}}$ – $\underset{\varepsilon_S}{\underline{\text{variance}}}$ decomposition of risk

Important Remark:

ε_S is an upper bound of the estimation error

→ Upper Bound is pessimistic

✲ Q: How does ε_S behave as a function of "size" of $\mathcal{F}$ and size of training set n?

$\varepsilon_S = \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$ (Uniform)

Recall that before, we measured fluctuations at an $f \in \mathcal{F}$:

$|\hat{R}(f) - R(f)| \simeq \frac{\sigma_f}{\sqrt{n}}$ (Pointwise)

* To get the main idea, consider idealized setting

(i) $\mathcal{F} = \{f_1, \cdots, f_M\}$ is a finite set of M hypothesis

(ii) $\hat{R}(f_i)$ are indep. Gaussian R.V.'s with mean $R(f_i)$ and variance σ^2

$\max_{i=1,\cdots,n} \hat{R}(f_i) - R(f_i)$

Then $Z_i = \hat{R}(f_i) - R(f_i) \sim \mathcal{N}(0, \sigma^2)$, i.i.d.,

Now $\mathbb{E} \max_i Z_i \sim \sqrt{2\sigma^2 \log M}$

10/10 FoML 12

Thursday, October 10, 2024 1:58 PM

FML Lecture 12 : Statistical Error in SL

Recall : $\hat{f} = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \hat{R}(f)$ ERM

$R(\hat{f}) - R^* \leq 2\mathcal{E}_S + \mathcal{E}_A$ with $\mathcal{E}_A = \inf_{f\in\mathcal{F}} R(f) - R^*$, $\mathcal{E}_S = \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|$

→ Natural Tension / Trade-off between approximation & statistical error

$\mathcal{E}_A \downarrow$ as $\mathcal{F}$ grows , while $\mathcal{E}_S \uparrow$ as $\mathcal{F}$ grows

→ To understand $\mathcal{E}_S$, we need to move from pointwise bound $|\hat{R}(f) - R(f)| \sim \sqrt{\frac{\sigma_f^2}{n}}$ to uniform bound

→ Simplified Settings :

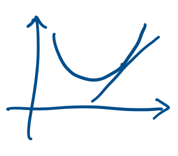
(1) $\mathcal{F} = \{f_1, \dots, f_M\}$ finite discrete hypothesis class

(2) $\hat{R}(f_i) \sim \mathcal{N}(R(f_i), \sigma^2)$, $i = 1, \cdots, M$

Q : $\mathbb{E} \max_{i=1,\cdots,M} (\hat{R}(f_i) - R(f_i))$?

Tools :

(a) Jensen's Inequality : f is convex , then $f(\mathbb{E}X) \leq \mathbb{E}f(X)$

Convexity :  for $\forall x, y$, $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle$ [linear approximation of f at x]

$f: \mathbb{R}^d \to \mathbb{R}$

Then apply x by $\mathbb{E}X$ to have : $f(\mathbb{E}X) \leq f(y) - \langle \nabla f(\mathbb{E}X), y - \mathbb{E}X \rangle$

Taking $\mathbb{E}[\cdot]$ on both sides , $f(\mathbb{E}X) \leq \mathbb{E}f(y)$ then choose y to be X

(b) Moment Generating Function

$t \mapsto \mathbb{E}[e^{tX}] = e^{\frac{t^2\sigma^2}{2}}$ for $x \sim \mathcal{N}(0, \sigma^2)$

Let $Z_i = \hat{R}(f_i) - R(f_i)$, so $Z_i \sim \mathcal{N}(0, \sigma^2)$

$\bar{Z} = \max(Z_1, \cdots, Z_n)$

We want $\mathbb{E}\bar{Z}$:

Let $t > 0$,
$$\mathbb{E}[\bar{Z}] = \mathbb{E}\left[\frac{1}{t}\log(e^{t\bar{Z}})\right] \underset{\text{(Concave)}}{\overset{\text{Jensen's}}{\leq}} \frac{1}{t}\log\left(\mathbb{E}[e^{t\bar{Z}}]\right) = \frac{1}{t}\log\left(\mathbb{E}\left[\max_{i=1,\cdots,M} e^{tZ_i}\right]\right)$$
$$\leq \frac{1}{t}\log\left(\mathbb{E}\left[\sum_{i=1}^{M} e^{tZ_i}\right]\right)$$
$$= \frac{1}{t}\log\left(\sum_{i=1}^{M}\mathbb{E}[e^{tZ_i}]\right)$$
$$\overset{\text{MGF}}{=} \frac{1}{t}\log\left(M e^{\frac{t^2\sigma^2}{2}}\right)$$
$$= \frac{\log M}{t} + \frac{t\sigma^2}{2} \triangleq \phi(t)$$

Recap : $\mathbb{E}\bar{Z} \leq \phi(t)$ for all $t > 0$

So $\mathbb{E}\bar{Z} \leq \inf_{t>0}\phi(t) = 2\sqrt{\frac{\log M\,\sigma^2}{2}} = \boxed{\sqrt{2\sigma^2\log M}}$

→ Price to pay (at most) for uniform deviations $\sqrt{2\log M}$

→ In fact, we can show (much harder) a $\boxed{\text{lower bound}}$ of the form

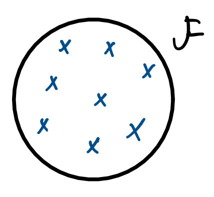
$c \cdot \sqrt{\log n \cdot \sigma^2}$ when $Z_1, \cdots, Z_M$ are i.i.d.

→ $\hat{R}(f) - R(f) \sim \mathcal{N}(0, \frac{\sigma_f^2}{n})$ therefore $\mathbb{E}[\max(\hat{R}(f) - R(f))] \leq \sqrt{\frac{2\sigma^2\log M}{n}}$

→ Key Step : Replace the max by the sum $\boxed{\text{Union bound}}$

→ Q : How about infinite hypothesis Space ?

e.g. $\mathcal{F} = \{ f(x) = H(x)^T\Theta ,\ \Theta \in \mathbb{R}^d \}$

Intuition :  Grid the hypothesis space

Recap : $R(\hat{f}) - R^* \leq \underbrace{\inf_{f\in\mathcal{F}} R(f) - R^*}_{\text{approx.}} + \underbrace{2\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|}_{\text{Statistical.}}$ $\simeq \sqrt{\frac{\log|\mathcal{F}|}{n}}$ (up to a constant)

Question :

1) When is the upper bound correctly capturing the tradeoff between approximation & estimation

2) Estimate the quantities in practice

3) How to efficiently adjust the "size" of $\mathcal{F}$ to balance errors ?

Answer :

(1) Upper bounds will be generally pessimistic

Exceptions : Sometimes we can directly analyze the generalization gap : $R(\hat{f}) - \hat{R}(\hat{f}) = \frac{\sigma^2 d}{n}$ (in the fixed design of OLS)

(2) <u>Cross-Validation</u>

In practice, we <u>split</u> the available data into two buckets

Training Set : $T = \{(x_i, y_i)\}_{i=1,\cdots,n}$

Validation Set : $V = \{(x_i', y_i')\}_{i=1,\cdots,m}$

ERM (using T) $\hat{f} = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \hat{R}(f)$

Goal : Estimate $R(\hat{f}) - R^*$

→ Recall that for each fixed f, $\hat{R}(f)$ is an unbiased estimator of $R(f)$

Why isn't $\hat{R}(\hat{f})$ a good estimator of $R(\hat{f})$?

Because $\hat{f}$ depends on randomness in T, can not treat $\hat{f}$ as fixed

- Define another estimator $\hat{R}(f) = \frac{1}{m}\sum_{j=1}^{m} \ell(f(x_i'), y_i')$

Still have that $\mathbb{E}_V \tilde{R} = R$ and $\tilde{R}(\hat{f})$ is an unbiased estimator of $R(\hat{f})$

Next Class : $|\tilde{R}(\hat{f}) - R(\hat{f})| \sim \sqrt{\frac{1}{m}}$ → the size of validation set .

FML Lecture 13 : Universal Approximation

Next Tuesday's Class : Florentin Gath Guest

Office Hours will be moved to Thursday.

In the past lectures, we saw that excess risk of ERM :

$$R(\hat{f}) - R^* \geq \underset{\text{(approximation error)}}{\mathcal{E}_A} = \min_{f \in \mathcal{F}} R(f) - R^*$$

Q : Design hypothesis class $\mathcal{F}$ s.t. $\mathcal{E}_A$ is as small as we want ?

→ Let $\mathcal{V} = \{ f : \mathcal{X} \to \mathbb{R} , f \text{ is continuous} \}$

↳ Assume that Bayes estimator $f^* \in \mathcal{V}$

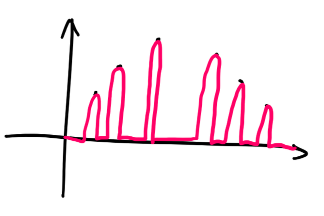
↳ We now consider a norm in $\mathcal{V}$ given the supremum of $f \in \mathcal{V}$, $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$

↳ Can we do ERM on $\mathcal{V}$ directly ?

i.e., Given $\{(x_i, y_i)\}_{i=1}^n$, $\min_{f \in \mathcal{V}} \hat{R}(f) = \frac{1}{n} \sum_{i=1}^n |y_i - f(x_i)|^2$

No ! Because there is no control of statistical error on $\mathcal{V}$!

$\sup_{f \in \mathcal{V}} | R(f) - \hat{R}(f) | = \|f^*\|$ for any n !



→ So we need to somehow "simplify" the universe.

Regularization Perspective

Consider a set $A \subseteq \mathcal{V}$, e.g. $A = \{ f : [0,1]^d \to \mathbb{R} \text{ polynomial} \}$

$A = \{ f : [0,1]^d \to \mathbb{R} , f(x) = W_L \sigma(W_{L-1} \sigma(W_{L-2} \cdots \sigma(W_1 x))) \}$ Nueral Nets of depth L

$A = \{ f(x) = H(x)^T \Theta , \Theta \in \mathbb{R}^d \}$ Linear Regression

→ Now we consider a "cost" measure over A

$\gamma : A \to \mathbb{R}$, measuring how expensive is it to use a given $f \in A$

ex. $A = \{ \text{polynomials} \}$, $\gamma(p) = \text{degree of } p$

$A = \{ f \text{ Nueral Networks} \}$, $\gamma(f) = \#$ of parameters

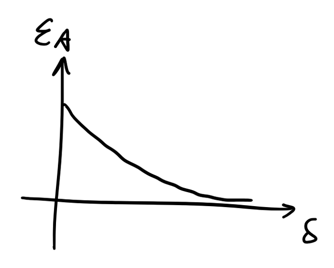
$A = \{ f(x) = H(x)^T \Theta , \Theta \in \mathbb{R}^d \}$, $\gamma(f) = \|\Theta\|$

→ We can now use γ to design a hypothesis class :

for each $\delta > 0$, $\mathcal{F}_\delta = \{ f \in A : \gamma(f) \leq \delta \}$

→ Q : How does $\mathcal{E}_A$ behave as we increase δ ?

When can we have $\mathcal{E}_A \to 0$ as $\delta \to \infty$



→ a set $A \subseteq \mathcal{V}$ is $\boxed{\text{dense}}$ if for $\forall f \in \mathcal{V}$ and any $\varepsilon > 0$, there exists $g \in A$ s.t. $\|f - g\| \leq \varepsilon$.

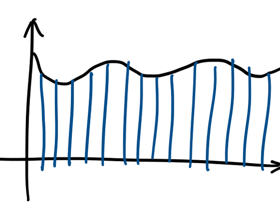
→ When a set $A \subseteq \mathcal{V} = C(\mathcal{X})$ is dense, we say that it has the $\boxed{\text{universal approximation property}}$

§ Universal Approximation of Polynomials

Consider a continuous function $f : [0, 1] \to \mathbb{R}$

For each m, we consider the polynomial (deg. m)

$$P_m(x) = \sum_{j=0}^m f\left(\frac{j}{m}\right) \binom{m}{j} x^j (1-x)^{m-j}$$



Theorem (Weierstrass, early 20th century)

$$\lim_{m \to \infty} \|f - P_m\| \triangleq \lim_{m \to \infty} \sup_{x \in [0,1]} |f(x) - P_m(x)| = 0$$

So polynomials have the the universal approximation property

Pf. (By Bernstein)

Fix $x \in [0, 1]$, Let $Z_1, \cdots, Z_m$ be i.i.d. Bernoulli R.V. of parameter x

Let $W = \frac{1}{m} \sum_{j=1}^m Z_j$ $[mW \sim \text{Binomial}(m, x)]$

We have (i) $1 = \sum_{j=0}^m \mathbb{P}(W = \frac{j}{m}) = \sum_{j=0}^m \binom{m}{j} x^j (1-x)^{m-j}$

(ii) $\mathbb{E}W = \mathbb{E}Z_i = x = \sum_{j=0}^m \binom{m}{j} \frac{j}{m} x^j (1-x)^{m-j}$

(iii) $\text{Var}(W) = \frac{\text{Var}(Z_1)}{m} = \frac{x(1-x)}{m} = \sum_{j=0}^m \binom{m}{j} \left(\frac{j}{m} - x\right)^2 x^j (1-x)^{m-j}$

$$P_m(x) - f(x) = \sum_{j=0}^m \binom{m}{j} f\left(\frac{j}{m}\right) x^j (1-x)^{m-j} - f(x) \cdot 1$$

$$= \sum_{j=0}^m \binom{m}{j} \left[f\left(\frac{j}{m}\right) - f(x) \right] x^j (1-x)^{m-j} \quad (*)$$

As $f \in C[0, 1]$, we have

① for $\forall \varepsilon > 0$, $\exists \delta > 0$ s.t. $|f(x) - f(y)| \leq \varepsilon$ whenever $|x - y| \leq \delta$ (δ is indep. of ε , uniform continuity)

② f is bounded : $\|f\| = \sup_{x \in [0,1]} |f(x)| = M < \infty$

We then break $(*)$ into two parts :

$$(*) = \sum_{j, |\frac{j}{m} - x| \leq \delta} \left| f\left(\frac{j}{m}\right) - f(x) \right| \underbrace{\binom{m}{j} x^j (1-x)^{m-j}}_{b_{m,j}(x)} + \sum_{j, |\frac{j}{m} - x| > \delta} \left| f\left(\frac{j}{m}\right) - f(x) \right| b_{m,j}(x)$$

$$\leq \varepsilon \cdot \sum_{|\frac{j}{m} - x| \leq \delta} b_{m,j}(x) + 2M \cdot \sum_{|\frac{j}{m} - x| > \delta} b_{m,j}(x)$$

$$= \varepsilon \cdot \mathbb{P}(|W - \mathbb{E}W| \leq \delta) + 2M \cdot \underbrace{\mathbb{P}(|W - \mathbb{E}W| > \delta)}_{\leq \frac{\text{Var}(W)}{\delta^2} = \frac{x(1-x)}{m\delta^2}} \quad \text{Chebyshev}$$

$$\leq \varepsilon + 2M \cdot \frac{x(1-x)}{m\delta^2} \leq \varepsilon + \frac{M}{2m\delta^2}$$

Setting $m = \frac{M}{2\varepsilon\delta^2}$ to get :

$$\sup_{x \in [0,1]} |P_m(x) - f(x)| \leq 2\varepsilon \quad \text{for } \forall \varepsilon > 0.$$

Therefore $\lim_{m \to \infty} \|P_m - f\| = 0$.

#

Remark :

→ The polynomial we have used here $b_{m,j}(x) = \binom{m}{j} x^j (1-x)^{m-j}$ are called $\boxed{\text{Bernstein Polynomial}}$

→ They are not optimal, in the sense of having smallest degree m for a target error ε. (optimal approximation in the uniform norm is obtained by $\boxed{\text{Chebyshev Polynomials}}$)

10/24 FoML 15

Thursday, October 24, 2024 2:03 PM

FML Lecture 15 : The Curse of Dimensionality

Recap : Excess Risk : $R(\hat{f}) - R^* = \underset{\underset{\inf_{f\in\mathcal{F}} R(f) - R^*}{\parallel}}{\overset{(\text{approx.})}{\mathcal{E}_A}} + \underset{\underset{R(\hat{f}) - \inf_{f\in\mathcal{F}} R(f)}{\parallel}}{\overset{(\text{estimation})}{\mathcal{E}_S}}$

↳ Two parameters guiding this error

→ n : # of training points

→ δ : "size" of hypothesis space $\mathcal{F}_\delta = \{ f \in \mathcal{A}, \ \gamma(f) \leq \delta \}$

$f : \mathcal{X} \to \mathbb{R}$, $d \triangleq \dim(\mathcal{X})$

We saw :

(i) $\mathcal{E}_A \to 0$ as $\delta \to \infty$ (Universal Approximation)

(ii) $\mathcal{E}_S \to 0$ as $n \to \infty$ (recall $\mathcal{E}_S \leq \sqrt{\frac{\log|\mathcal{F}_\delta|}{n}}$)

→ Supervised Learning works "asymptotically"

Today : Practical aspect (i.e., finite δ, n) ?

Key extra parameter : dimension d of input space

↳ Generic Phenomena : n, δ need to grow exponentially in d

• Curse of Dimensionality [Bellman 1950s]

Two vigenettes of CoD :

(1) Approximation with polynomials

Last week we saw that polynomials have UAP

In $d = 1$, $f : [0,1] \to \mathbb{R}$ conti. , then $\lim_{K\to\infty} \inf_{\mathrm{poly}(K)} \|f - p\| = 0$

$\mathcal{P}_K = \{ p : [0,1] \to \mathbb{R}, \text{ polynomials of degree } K \}$

If $f \in C^1$, then $\inf_{p \in \mathcal{P}_K} \|f - p\| \leq \frac{1}{K}$

In other words, if we want ε , we set degree of poly. to be $K = \frac{1}{\varepsilon}$

In $d = 1$, $\mathcal{P}_K$ contains $p(x) = x^K + a_{K-1}x^{K-1} + a_{K-2}x^{K-2} + \dots + a_0$, $a \in \mathbb{R}^K$

Q: What happens as d decreases ?

$f : [0,1]^d \to \mathbb{R}$, $f \in C^1$

$\mathcal{P}_K = \{ p : [0,1]^d \to \mathbb{R}, \ p \text{ is a multivariate poly. of deg. } K \}$

e.g. $d = 2, K = 3$: $x_1^3, \ x_1^2 x_2, \ x_1 x_2^2, \ x_2^3$

→ It is not hard to check that $\mathcal{P}_{K,d}$ has UAP in the class of smooth functions (using e.g. Stone - Weierstrass)

→ We also preserve the rate of approximation : $\inf_{p\in\mathcal{P}_{K,d}} \|f - p\| \leq \frac{1}{K}$ ⇒ we need at least $\frac{1}{\varepsilon}$ degree to reach approx. error ε

↳ How many parameters do we need to express $\mathcal{P}_{r,d}$?

$x_1^{S_1} x_2^{S_2} \cdots x_d^{S_d}$ where $S_i \in \mathbb{N}$, $S_i \geq 0$ & $K = \sum_{i=1}^{d} S_i$

# of possible choices : $\binom{d+K-1}{K} = \binom{d+K-1}{d-1} \overset{K \gg d}{\approx} \binom{K}{d} \simeq K^d = \varepsilon^{-d}$

→ Same is true if we replace polynomials by Neural Nets

→ $\delta = \varepsilon^{-d}$ is a "signature" of Curse of Dimensionality

• Estimation of Continuous / Lipschitz Functions

Say we want to learn a target function $f^* : [-1,1]^d \to \mathbb{R}$

from examples $\{ (X_i, y_i = f(X_i)) \}_{i=1,\dots,n}$, under the assumption

that f^* is 1 - Lipschitz : $|f^*(x) - f^*(x')| \leq \|x - x'\|$ for $\forall x, x'$

→ A natural estimator in this setting is the Nearest Neighbor estimator

$$\hat{f}(x) = f^*(X_{i(x)}) \quad \text{where} \quad i(x) = \underset{i=1,\dots,n}{\operatorname{argmin}} \|x - X_i\|$$

(Fundamental nonparametric estimator)

↳ Existence of memorization, and exploit smoothness prior

Q: How well does Nearest Neighbor do ?

$$\mathbb{E}_X |\hat{f}_{NN}(x) - f^*(x)| = \mathbb{E}_X |f^*(X_{i(x)}) - f^*(x)| \leq \mathbb{E}_X \|X_{i(x)} - x\|$$

Uniform Distribution is optimal for the lower bound.

In this case, the expected error is $\varepsilon \sim n^{-\frac{1}{d}}$

⟺ To reach error ε, we need $n \sim \varepsilon^{-d}$ points

⟺ high dimensional spaces are very lonely places !

10/29 FoML 16

Tuesday, October 29, 2024 2:06 PM

FML Lecture 16 : Optimization in ML

Recap so far : → Focus on statistical & approximation in SL

→ We have viewed ERM as a black-box

ERM : $\min_{f \in \mathcal{F}_S} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), y_i) = \hat{R}(\hat{f})$

Beyond OLS, this problem does not admit a closed-form solution

⟶ We need to resort to iterative, optimization methods

→ We will focus on the two most important methods

(i) Gradient Descent

(ii) Stochastic Gradient Descent

Optimization Basics :

Consider a generic optimization $\min_{\Theta \in \mathbb{R}^d} F(\Theta)$

① When can we solve this problem efficiently ?

② How expensive ?

Def. (Global Minimizer)

A point $\Theta^* \in \mathbb{R}^d$ is a $\boxed{\text{global minimizer}}$ of F if $F(\Theta^*) \le F(\Theta)$ for $\forall \Theta \in \mathbb{R}^d$

(Local Minimizer)

A point $\Theta^* \in \mathbb{R}^d$ is a $\boxed{\text{local minimizer}}$ if $\exists \varepsilon > 0$ s.t. $F(\Theta^*) \le F(\Theta)$ for all $\Theta \in B_\varepsilon(\Theta^*)$

Remark. Global minimizer is a (much) stronger property than local minimizer

Q: How hard is to solve a (generic) optimization problem (in high dimension) ?

→ We only access the function via local queries

In general, we need to grid / explore all the domain to find the global minimum.

→ We need an exponential number of queries (Curse of dimensionality)

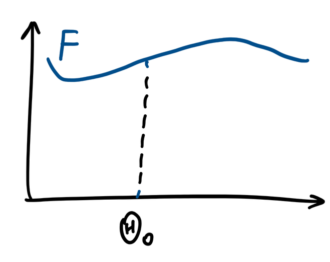
→ Contrary to the worst case, many typical global optimization problems can be solved by breaking them into a sequence of local optimisation problems

↳ Eg. Navigation

Given some point Θ_0, we aim to find a nearby point Θ_1 s.t. $F(\Theta_1) < F(\Theta_0)$

How to find such update ?

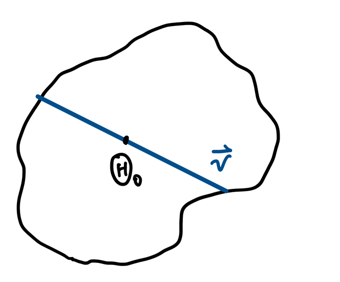
($d = 1$ setting) :



sign$(F'(\Theta_0))$ indicates whether to go left / right

$F(\Theta_0 + t) = F(\Theta_0) + t \cdot F'(\Theta_0) + o(\frac{t^2}{2})$

(Taylor Expansion)



$f(t) := F(\Theta_0 + tv)$

$f'|_{t=0} = \langle \nabla F(\Theta_0), v \rangle$

So $\Theta_1 = \Theta_0 - \varepsilon \cdot \text{sign}(f'(0)) \cdot v$

In particular, choosing $v = -\nabla F(\Theta_0)$ is the steepest descent direction whenever $\nabla F(\Theta_0) \neq 0$

In particular, at local minimizers Θ, we must have $\nabla F(\Theta) = 0$ (necessary)

The set of points $G = \{\Theta ; \nabla F(\Theta) = 0\}$ are called the first-order stationary / critical points

$GM = \{\Theta ; \Theta \text{ is a global min}\}$

$LM = \{\Theta : \Theta \text{ is a local min}\}$

$GM \subset LM \subset \underset{\underset{\text{Saddle}}{\cup}}{G} \supset \text{Max}$

- Gradient Descent (Cauchy, 1847)

(i) Picking an initial point $\Theta_0 \in \mathbb{R}^d$.

(ii) For each $t = 0, 1, 2, 3, \cdots$

Pick a step-size $\eta_t > 0$ $\boxed{\text{(learning rate)}}$ and set

$\Theta_{t+1} = \Theta_t - \eta_t \nabla F(\Theta_t)$

Remark : $\{\Theta_t\}$ is a random sequence if either F and/or Θ_0 are random

Key Questions :

→ When can we guarantee that GD finds the global optimum ?

How long do we need to run it ? How to adjust LR ?

→ How to apply it to solve ERM.

→ How to scale it to large problems ?

- GD succeeds whenever $GM = G$

In particular, F convex satisfies this property.

pf. We only need to show that $C \subseteq GM$ if F convex

Recall F convex : for $\forall \theta$, $\forall \alpha \in [0,1]$, $F((1-\alpha)\theta^* + \alpha\theta) \le (1-\alpha) F(\theta^*) + \alpha F(\theta)$

$\Leftrightarrow F(\theta) \ge F(\theta^*) + \frac{1}{\alpha}\left(F((1-\alpha)\theta^* + \alpha\theta) - F(\theta^*)\right)$

Let $g(t) = F(\theta^* + t(\theta - \theta^*))$

$F(\theta) \ge F(\theta^*) + \frac{g(\alpha) - g(0)}{\alpha}$

Mean-Value Thm. $\exists \tilde{\alpha} \in (0, \alpha)$ s.t. $\frac{g(\alpha) - g(0)}{\alpha} = g'(\tilde{\alpha}) = \langle \nabla F(\theta^* + \tilde{\alpha}(\theta - \theta^*)), \theta - \theta^* \rangle$

By sending $\alpha \to 0$,

$F(\theta) \ge F(\theta^*) + \langle \nabla F(\theta^*), \theta - \theta^* \rangle$

So, $F(\theta) \ge F(\theta^*)$ for all θ if $\theta^* \in G$ $\Rightarrow \theta^* \in GM$

#

10/31 FoML 17

Thursday, October 31, 2024 2:01 PM

FML Lecture 17 : Optimization II

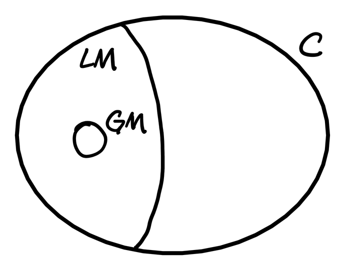
Recap : • Optimization in worst case : too hard (in high-dim)

• Approach : Use local descent method iteratively

$C = \{ \theta ; \nabla F(\theta) = 0 \}$

$LM = \{ \theta ; \theta \text{ is a local minimum of } F \}$

$GM = \{ \theta ; F(\theta) \le F(\theta') \text{ for } \forall \theta' \}$



Generically, C are the equilibrium points of gradient descent

LM are the stable equilibrium points

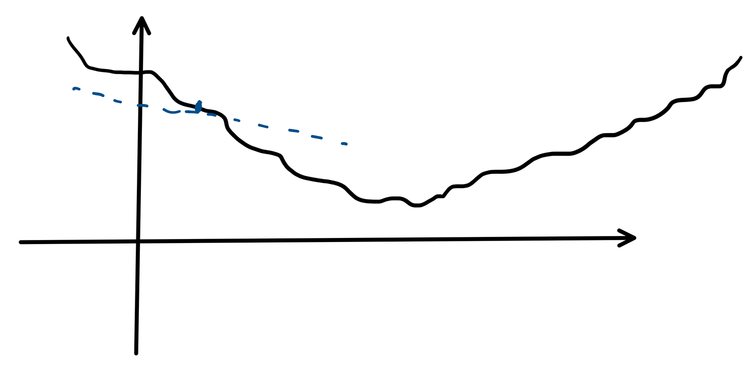
→ There is a class of functions where $C = GM$: Convex Functions

Remark : There are other functions F s.t. $C = GM$

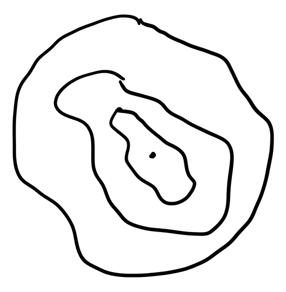
(*) quasi - convex functions

F s.t. its level sets $S_\lambda = \{ \theta ; F(\theta) \le \lambda \}$ are convex for $\forall \lambda$

A good example : (quasi - convex but not convex)



• If F is quasi - convex, then GD will find a global optima



If $\lambda \le \tilde{\lambda}$, then $S_\lambda \subseteq S_{\lambda'}$

(**) F with a Polyak - Lovajcievice (PL) inequality :

$\| \nabla F(\theta) \| \ge a \, | F(\theta) - F(\theta^*) |^b$

GD will find global optima

(***) F with discrete symmetries,

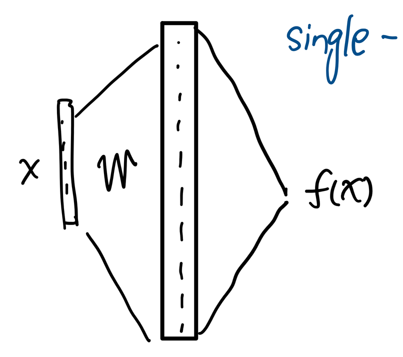
$F(T_k \theta) = F(\theta)$, $\forall \theta$, $\{ T_1, \cdots, T_k \}$ is a family of transform

Eg. θ - parameters of a Neural Network

Single - hidden NN

$\theta = \{ W, a \}$; $f(x ; \theta) = a^T \cdot \sigma(Wx)$

$W \in \mathbb{R}^{m \times d}$; $a \in \mathbb{R}^m$



• Insights from quadratic functions

$F(\theta) = \frac{1}{2n} \| H\theta - y \|^2$ (ordinary least square)

$\theta \in \mathbb{R}^d$

$H \in \mathbb{R}^{n \times d}$

$y \in \mathbb{R}^n$

• $\nabla F(\theta) = \frac{1}{n} H^T (H\theta - y) = K\theta - \frac{1}{n} H^T y$, $\nabla^2 F(\theta) = K$

• Recall that θ^* is a GM iff $\nabla F(\theta^*) = 0$ (Normal Equations)

$K\theta^* = \frac{1}{n} H^T y$

• F equals its 2nd - order Taylor Approximation (since F quadratic)

$F(\theta) = F(\theta^*) + \langle \underbrace{\nabla F(\theta^*)}_{0}, \theta - \theta^* \rangle + \frac{1}{2} (\theta - \theta^*)^T \nabla^2 F(\theta^*) (\theta - \theta^*)$

$= F(\theta^*) + \frac{1}{2} (\theta - \theta^*)^T K (\theta - \theta^*)$

→ Recall that K is symmetric psd , i.e., K has eigenvalues $\lambda_1, \cdots, \lambda_d$ where $\lambda_i \ge 0$

→ Define $\mu = \min(\lambda_i)$, $L = \max(\lambda_i)$, $\rho = \frac{L}{\mu} \ge 1$ Condition Number of K

→ Gradient Descent with fixed step - size $\eta > 0$ & initial point θ_0

$\theta_{t+1} = \theta_t - \eta \cdot \nabla F(\theta_t)$

$= \theta_t - \eta \left(K\theta_t - \frac{1}{n} H^T y \right)$

as $\frac{1}{n} H^T y = K\theta^*$

$= \theta_t - \eta K (\theta_t - \theta^*)$

$\Rightarrow \theta_{t+1} - \theta^* = [I - \eta K] (\theta_t - \theta^*)$

$= [\underbrace{I - \eta K}_{A}]^{t+1} (\theta_0 - \theta^*)$

→ Now we can track progress of GD : (not well defined for non-unique θ^*)

$\| \theta_t - \theta^* \|^2 = (\theta_0 - \theta^*)^T A^{2t} (\theta_0 - \theta^*)$ (Iterate Convergence)

→ $F(\theta_t) - F(\theta^*) = \frac{1}{2} \| \theta_t - \theta^* \|_K^2$ $\left(= (\theta_t - \theta^*)^T K (\theta_t - \theta^*) \right)$

$= \frac{1}{2} (\theta_0 - \theta^*)^T A^t K A^t (\theta_0 - \theta^*)$

$= \frac{1}{2} (\theta_0 - \theta^*)^T A^{2t} K (\theta_0 - \theta^*)$

Recall that K has eigenvalues in $[\mu, L]$

↳ Let us first assume that θ^* is unique $\Leftrightarrow \mu > 0$

Q : What are the eigenvalues of A ?

eigenvalues of K : $\{ \lambda_i \}_{i=1}^d \longleftrightarrow \{ 1 - \eta \lambda_i \}_{i=1}^d$: eigenvalues of $A = I - \eta K$

(matrix calculus) $\longleftrightarrow \{ (1 - \eta \lambda_i)^{2t} \}_{i=1}^d$: eigenvalues of A^{2t}

→ To guarantee that $\| \theta_t - \theta^* \|^2 \xrightarrow{t \to \infty} 0$, we want $| 1 - \eta \lambda_i | < 1$ for $\forall i = 1, \cdots, d$

Eg. Pick $\eta = \frac{1}{L}$ where $L = \max \lambda_i$

$\lambda_i \in [\mu, L] \Rightarrow \eta \lambda_i = \frac{\lambda_i}{L} \in [\frac{\mu}{L}, 1]$

$\Rightarrow 1 - \eta \lambda_i \in [0, 1 - \frac{\mu}{L}] \subset [0, 1)$

$\|$

$[0, 1 - \rho^{-1}]$

$\Rightarrow \| A \|^{2t} \le (1 - \rho^{-1})^{2t}$

11/5 FoML 18

Tuesday, November 5, 2024 1:59 PM

Lecture 18 : Optimization (cont'd)

Recap : Analysis of GD on quadratic functions $F(\theta) = \|H\theta - y\|^2 = F(\theta^*) + (\theta - \theta^*)^T K (\theta - \theta^*)$

$K \in \mathbb{R}^{d \times d}$ GD : $\theta_{t+1} = \theta_t - \eta \nabla F(\theta_t)$

→ We saw (when $K \succ 0$) : $\|\theta_t - \theta^*\|^2 \leq (1 - \frac{1}{\rho})^t \|\theta_0 - \theta^*\|^2$

ρ : Condition Number of K : $\frac{\lambda_{max}(K)}{\lambda_{min}(K)} \begin{matrix} \triangleq L \\ \triangleq u \end{matrix}$ ↳ by setting $\eta = \frac{1}{\lambda_{max}(K)}$

Remark :

(1) This is what we call a "linear" convergence (error decays exponentially fast)

(2) The bound $1 - \frac{1}{\rho}$ comes from the operator norm of $A = I - \eta K$

⇒ Any choice of $\boxed{\eta \in (0, \frac{2}{\lambda_{max}(K)})}$ guarantee's exponential convergence

Questions :

① Optimality of GD ?

② What happens when $u = 0$ (in particular when $d > n$)

Answers :

① GD is NOT optimum amongst algorithms that only rely on gradients (first-order method)

Nestewr at 90s : Using "Momentum", one can replace ρ by $\sqrt{\rho}$ on convergence

②: $u = 0 \Rightarrow \rho = +\infty$ ⇒ Previous bound says $\|\theta_t - \theta^*\| \leq \|\theta_0 - \theta^*\|$

→ Rather than tracking $\|\theta_t - \theta^*\|$, now we can track $|F(\theta_t) - F(\theta^*)|$

→ Using again $\eta = \frac{1}{L}$, and recall $F(\theta_t) - F(\theta^*) = (\theta_0 - \theta^*)^T (I - \eta K)^{2t} K (\theta_0 - \theta^*)$

→ Let's again bound the eigenvalues of $(I - \eta K)^{2t} K$

$$\| [I - K/L]^{2t} K \|_{op} \leq \sup_{\lambda \in [0, L]} \left| \lambda (1 - \lambda/L)^{2t} \right| \underset{\lambda = \frac{L}{2t+1}}{=} \underbrace{\frac{L}{2t+1}}_{\downarrow t \uparrow \infty \; 0} \cdot \underbrace{\left(1 - \frac{1}{2t+1}\right)^{2t}}_{\downarrow t \uparrow \infty \; e^{-1}} \leq \frac{L}{2t+1}$$

Therefore, $F(\theta_t) - F(\theta^*) \leq \boxed{\frac{L}{2t+1}} \|\theta_0 - \theta^*\|^2$

Convergence but much slower than $u > 0$

Recap : Till now, we have :

When $K \succ 0$ $(u > 0)$, $\|\theta_t - \theta^*\|^2 \leq (1 - \rho^{-1})^t \|\theta_0 - \theta^*\|^2$

$$F(\theta_t) - F(\theta^*) \leq L (1 - \rho^{-1})^t \|\theta_0 - \theta^*\|^2$$

When $u = 0$, $F(\theta_t) - F(\theta^*) \leq \frac{L}{2t} \|\theta_0 - \theta^*\|^2$

In other words, to reach error ε, we need $\begin{cases} t \simeq \rho \cdot \log(\frac{1}{\varepsilon}) & \text{iterations} \quad (u > 0) \\ t \simeq L/\varepsilon & \text{iterations} \quad (u = 0) \end{cases}$

Remark :

→ We have shown that upper bounds for the loss convergence at a certain rate

↳ This may be pessimistic in practice !

↳ Contrast with "scaling laws", which looks at typical case

Q: What happens when $\eta \to 0$? $[\theta_{t+1} = \theta_t - \eta \nabla F(\theta_t)]$

The sequences $\{\theta_t^{(\eta)}\}_t$ accumulates to a conti. curve $\{\theta(t)\}_{t \in \mathbb{R}_+}$

$\frac{\theta_{t+1} - \theta_t}{\eta} = -\nabla F(\theta_t)$ Say now $\theta_t = \theta(\eta t)$

Then $\dot{\theta}(t_0) = \frac{\theta(t_0 + \eta) - \theta(t_0)}{\eta} = -\nabla F(\theta(t_0))$

having more theoretical significance

Therefore $\{\theta(t)\}_{t \in \mathbb{R}_+}$ satisfies : $\dot{\theta}(t) = -\nabla F(\theta(t))$ which is called $\boxed{\text{Gradient Flow}}$

→ At small η, GD is a time discretisation of the gradient flow ODE

FML Lecture 19: Optimization cont'd: Convex Functions

Recap: Analysis of GD on quadratic functions

$F(\theta) = F(\theta^*) + \frac{1}{2}(\theta-\theta^*)^T K(\theta-\theta^*)$ eigenvalues of K: $[u, L]$ & condition number $\rho = \frac{L}{u}$

→ Choosing $\eta = 1/L$, iteration complexity to reach error ε

$$t \simeq \begin{cases} \rho \cdot \log(1/\varepsilon) & \text{when } u>0 \\ L \cdot \frac{1}{\varepsilon} & \text{when } u=0 \end{cases}$$

→ When $\eta \ll 1/L$ → GD can be analyzed using differential calculus (Gradient Flow ODE: $\dot{\theta}(t) = -\nabla F(\theta(t))$)

→ When $\eta > 2/L$, then GD diverges

Today: Beyond Quadratic Functions

- The Convex case: (Recall F convex if $\forall x, y \in \mathbb{R}^d$, $\forall \alpha \in [0,1]$, $F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y)$)

In quadratic case, $\nabla^2 F(\theta) = K$ as a constant (matrix)

→ In the convex case, $\nabla^2 F(\theta)$ is no longer constant, but it satisfies $\nabla^2 F(\theta) \geq 0$

Def. A function $F \in C^2$ is $\boxed{u\text{-strongly convex}}$ if $\nabla^2 F(\theta) - uI \geq 0$ for $\forall \theta$ & $u > 0$

In other words, for all θ, all eigenvalues of $\nabla^2 F(\theta)$ are $\geq u > 0$

Def. $F \in C^2$ is $\boxed{L\text{-smooth}}$ if ∇F is L-Lipschitz. $(L>0)$

i.e., all eigenvalues of the Hessians $\nabla^2 F(\theta)$ are bounded by $L > 0$

→ $\boxed{\text{“Sandwich” Property}}$: If F is u-strongly convex & L-smooth, then:

$\forall x, y$, $F(x) + \langle \nabla F(x), y-x \rangle + \frac{u}{2}\|x-y\|^2 \leq F(y) \leq F(x) + \langle \nabla F, y-x \rangle + \frac{L}{2}\|x-y\|^2$

→ Strongly Convex Setting (u-sc)

→ Do we have a unique minimizer in this case?

Assume (towards contradiction) that θ_1^*, θ_2^* where $\theta_1^* \neq \theta_2^*$ and both minimize.

Plug $x = \theta_1^*$ & $y = \theta_2^*$ in the sandwich lower bound

$F(\theta_1^*) + \frac{u}{2}\|\theta_1^* - \theta_2^*\|^2 \leq F(\theta_2^*)$!! Contradiction

So minimizer is unique.

→ Strongly convex functions satisfy a P-L inequality:

$\|\nabla F(\theta)\|^2 \geq 2u(F(\theta) - F(\theta^*))$ where θ^* is the unique minimizer of F

Pf. Recall Sandwich Property:

$$\underbrace{G_x}_{\text{parameter}}(\overbrace{y}^{\text{variable}}) \triangleq F(x) + \langle \nabla F(x), y-x \rangle + \frac{u}{2}\|y-x\|^2 \leq F(\theta) \quad (*)$$

$\nabla_\theta G_x(y) = \nabla F(x) + u(y-x)$

$\Rightarrow y_x^* = x - \frac{1}{u}\nabla F(x)$

So $G_x(y_x^*) \leq F(\theta^*)$ [taking $\min(\cdot)$ on both sides of $(*)$]

$\|$

$F(x) + \langle \nabla F(x), \frac{-1}{u}\nabla F(x) \rangle + \frac{u}{2}\|\frac{1}{u}\nabla F(x)\|^2$

$= F(x) - \frac{1}{2u}\|\nabla F(x)\|^2$

$\Rightarrow \|\nabla F(\theta)\|^2 \geq 2u(F(\theta) - F(\theta^*))$

#

→ This PL inequality allows us to establish linear convergence

Prop. Choose $\eta = 1/L$. The iterates of GD satisfy $F(\theta_t) - F(\theta^*) \leq (1-\rho^{-1})^t (F(\theta_0) - F(\theta_t))$ where $\rho = L/u$

Pf. From $\theta_t = \theta_{t-1} - \nabla F(\theta_{t-1})/L$

$F(\theta_t) = F(\theta_{t-1} - \nabla F(\theta_{t-1})/L) \leq F(\theta_{t-1}) + \langle \nabla F(\theta_{t-1}), -\nabla F(\theta_{t-1})/L \rangle + \frac{L}{2}\|\nabla F(\theta_{t-1})/L\|^2$

$\Rightarrow F(\theta_t) \leq F(\theta_{t-1}) - \frac{1}{2L}\|\nabla F(\theta_{t-1})\|^2$ [Descent Lemma]

So $F(\theta_t) - F(\theta^*) \leq F(\theta_{t-1}) - F(\theta^*) - \frac{1}{2L}\|\nabla F(\theta_{t-1})\|^2$

$\overset{\text{(previous lem.)}}{\leq} F(\theta_{t-1}) - F(\theta^*) - \frac{u}{L}(F(\theta_{t-1}) - F(\theta^*))$

$= (1-\rho^{-1})(F(\theta_{t-1}) - F(\theta^*))$

$\leq (1-\rho^{-1})^t (F(\theta_0) - F(\theta^*))$

#

→ As in the quadratic setting, condition number of Hessians $\rho = L/u$ determines speed of convergence

Q: Continuous-time Analysis?

Recall: PL Inequality: $\|\nabla F(\theta)\|^2 \geq 2u(F(\theta) - F(\theta^*))$

Gradient Flow: $\dot{\theta}(t) = -\nabla F(\theta(t))$

Track $F(\theta(t)) - F(\theta^*) \triangleq f(t) \geq 0$

$f'(t) = \langle \nabla F(\theta(t)), \dot{\theta}(t) \rangle = -\|\nabla F(\theta(t))\|^2 \leq -2u \cdot f(t)$

$\Rightarrow f(t) \leq f(0)e^{-2ut}$ Gronwall's lemma

So the loss decreases exponentially.

Rmk. In the continuous setting, we don't see L appears.

11/12 FoML 20

Tuesday, November 12, 2024 1:58 PM

FML Lecture 20 : Optimization : Surrogate Loss , SGD

Recap : Analysis of GD on convex functions

Strongly convex setting : $F(\theta_t) - F(\theta^*) \leq \left(1 - \frac{\mu}{L}\right)^t \left(F(\theta_0) - F(\theta^*)\right)$

Today : * Analysis of GD in (vanilla) convex setting.

* Discuss examples where convex functions appear in ML : linear classification
* Stochastic Gradient Descent

Reminder : from quadratic setting

When we lost strong convexity ($\mu = 0$), we went from a $O((1-\rho^{-1})^t)$ rate to a $O(1/t)$ rate

Q. Same thing in the general convex setting ?

A: Yes !

Focus on the continuous time : $\dot{\theta}(t) = -\nabla F(\theta(t))$

Consider the function : $L(t) = t \cdot \left(F(\theta(t)) - F(\theta^*)\right) + \frac{1}{2}\|\theta(t) - \theta^*\|^2$: Lyapunov Function (dynamical system)

where $\theta_* \in \underset{\theta}{\arg\min}\, F(\theta)$ (global minimiser)

(represents stability)

(in general, it always decreases, i.e., the system converges to stationary)

Let's compute $L'(t) = \left(F(\theta(t)) - F(\theta_*)\right) + t < \nabla F(\theta(t)) , \underbrace{-\nabla F(\theta(t))}_{\dot{\theta}(t)} > + < \dot{\theta}(t) , \theta(t) - \theta_* >$

$= F(\theta(t)) - F(\theta_*) - t\|\nabla F(\theta(t))\|^2 - < \nabla F(\theta(t)) , \theta(t) - \theta_* >$

$= \underbrace{F(\theta(t)) - F(\theta_*) + < \nabla F(\theta(t)) , \theta_* - \theta(t) >}_{\leq 0 \text{ as } F \text{ is convex}} - t\|\nabla F(\theta(t))\|^2$

≤ 0

Therefore $L(t) \leq L(0)$

$\Rightarrow t \cdot \left(F(\theta(t)) - F(\theta_*)\right) \leq L(t) \leq L(0) = \frac{1}{2}\|\theta(0) - \theta_*\|^2$

$\Rightarrow F(\theta(t)) - F(\theta_*) \leq \frac{1}{2t}\|\theta(0) - \theta_*\|^2$

- Beyond Gradient Descent

↳ Momentum and acceleration : Use memory to improve convergence : $O(1/t) \to O(1/t^2)$

$O((1-\rho^{-1})^t) \to O((1-\rho^{-\frac{1}{2}})^t)$

↳ Normalisation / Adaptive Learning rates (Adam, Adagrad, ⋯)

↳ Second - Order Methods : use gradient ∇F but also Hessian information $\nabla^2 F(\theta)$

(eg. guass - Newton : $\theta_{t+1} = \theta_t - \nabla^2 F(\theta_t)^{-1} \nabla F(\theta_t)$)

(very fast but very expensive)

↳ Stochastic Gradient Descent : See next !

Examples of Convex ERM :

→ Ex 0 : Linear Regression : $\hat{R}(\theta) = \|\hat{H}^T\theta - \hat{y}\|^2$, $\hat{R}$ is convex & quadratic

→ Linear Classification : $\{(x, y)\}$ where $y \in \{1, \cdots, K\}$ (y is discrete label)

→ Simplest instance : $K = 2 \longrightarrow$ Binary Classification

eg. Spam filter / Fraud detection / text is AI - generated

→ Natural Loss $\ell(y, \hat{y}) = \mathbb{1}_{\{y \neq \hat{y}\}}$, or ,

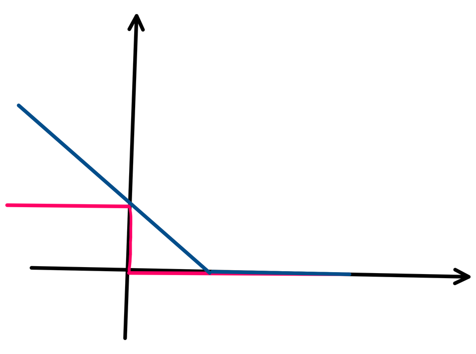
$y \backslash \hat{y}$	-1	+1
-1	0	1
+1	1	0

, $\mathbb{1}_{\{y\hat{y} < 0\}}$

→ Associated ERM : $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) \longrightarrow$ counts average # of mistakes

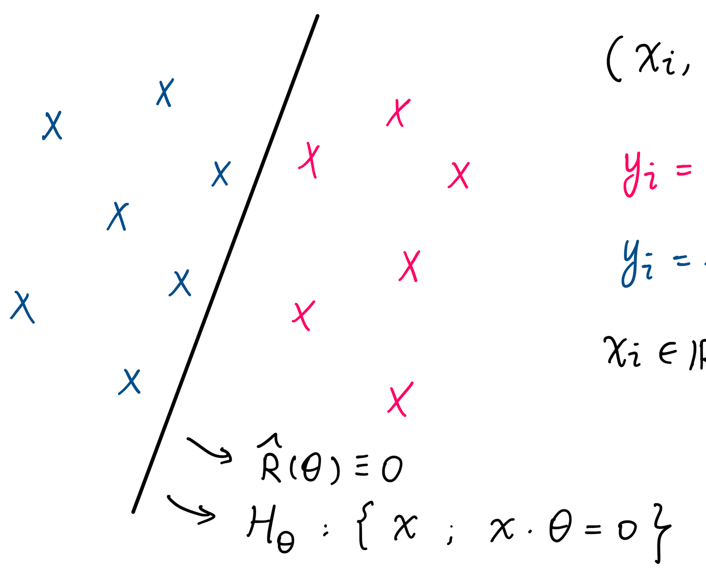
→ Q : Can we use gradient descent methods to solve this ERM ?

A : No ! Gradients are zero a.e. !



Sol. Replace this loss by a smoother one (with nonzero gradient) : $\boxed{\text{Surrogate Loss}}$

$z = y\hat{y}$, $\ell(z) = \max(0, 1 - z)$ $\boxed{\text{Hinge Loss}}$



- Geometric Interpretation of Hinge Loss

We want to find $\theta \in \mathbb{R}^d$ s.t. $\begin{cases} \theta \cdot x_i > 0 & \text{if } y_i = +1 \\ \theta \cdot x_i < 0 & \text{if } y_i = -1 \end{cases}$

Assume that data is linearly separable : ∃ such hyperplane

x X

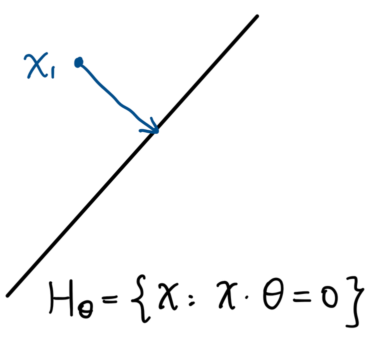
X x

(X or problem) [not linearly separable]

Q : Which hyperplane to pick amongst those that separate data ?

↳ We may want to pick a hyperplane as far as possible from the data

→ Maximise the margin



$\text{dist}(x_i, H_\theta) = \frac{|< x_i , \theta >|}{\|\theta\|}$

Several Options :

(i) Find hyperplane with largest margin : $\max_\theta \min_i \frac{y_i < x_i, \theta >}{\|\theta\|} \Leftrightarrow \min_\theta \|\theta\|$ subject to $y_i < x_i , \theta > \geq 1$ for $\forall i = 1, \cdots, n$

$\boxed{\text{Support Vector Machine (Vapnik)}}$

FML Lecture 21 : Stochastic Gradient Descent

Recap : • Binary Classification : Error measure $\ell(y,\hat{y}) = 1_{\{y\hat{y}<0\}}$

defines a loss with no gradients !

• Introduce a surrogate loss $\tilde{\ell}(y\hat{y})$ with "useful" gradients

• Margin : first example of surrogate loss

for SVM , $\text{Margin}(\theta) = \min_i \frac{y_i \langle x_i, \theta\rangle}{\|\theta\|}$

$\rightarrow \max_\theta \text{Margin}(\theta)$

$\rightarrow$ penalize small margins : $\tilde{\ell}(y\hat{y}) = \max(1-y\hat{y}, 0)$

$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \max(1-y_i\langle x_i,\theta\rangle, 0) + \frac{\lambda}{2}\|\theta\|^2$$

$\rightarrow$ This ERM is associated with the perceptron [McCulloch & Pitts , 1943 , Rosenblatt' 50s]

• $\hat{R}$ is convex w.r.t. θ (in fact it is λ- strongly convex)

$\Leftrightarrow$ Logistic Loss : $\tilde{\ell}(y\hat{y}) = \log(1+e^{-y\hat{y}})$

$\rightarrow$ Probability Interpretation

Model : $y|x \sim \text{Bern}\left(\frac{e^{\frac{1}{2}\langle x,\theta\rangle}}{e^{\frac{1}{2}\langle x,\theta\rangle}+e^{-\frac{1}{2}\langle x,\theta\rangle}}\right)$

$$\mathbb{P}_\theta(y=+1|x) = \frac{e^{\frac{1}{2}\langle x,\theta\rangle}}{e^{\frac{1}{2}\langle x,\theta\rangle}+e^{-\frac{1}{2}\langle x,\theta\rangle}} = \frac{1}{1+e^{-\langle x,\theta\rangle}} = \frac{1}{1+e^{-y\langle x,\theta\rangle}}$$

$$\mathbb{P}_\theta(y=-1|x) = 1-\mathbb{P}_\theta(y=+1|x) = \frac{1}{1+e^{\langle x,\theta\rangle}} = \frac{1}{1+e^{-y\langle x,\theta\rangle}}$$

$\rightarrow$ Consider the MLE :

$$\max_\theta \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_\theta(y_i|x_i) \Leftrightarrow \min_\theta \frac{1}{n}\sum_{i=1}^n \log(1+e^{-y_i\langle x_i,\theta\rangle}) = \hat{R}(\theta) \text{ using logistic loss}$$

$\rightarrow \hat{R}$ is also convex (as $t \mapsto \log(1+e^{-t})$ is convex)

$\rightarrow$ All these surrogate losses can be optimized by GD (thanks to convexity)

$\rightarrow$ Big caveat : any ERM of the form $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i, \theta)$ has a gradient of the form : $\nabla_\theta \hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell(y_i, x_i, \theta)$ $\rightarrow$ need to use all the data all the time

unfeasible in large scale ML !

• Stochastic Gradient Descent [Robbins & Munro 50s]

We can view the training loss $\hat{R}(\theta)$ as an expectation over the data :

$$\hat{R}(\theta) = \mathbb{E}_{(x_i,y_i)\sim T}[\ell(x_i,y_i,\theta)] \quad (+\lambda H(\theta))$$ optional regularization

$$\nabla_\theta \hat{R}(\theta) = \mathbb{E}_{(x_i,y_i)\sim T}[\nabla_\theta \ell(x_i,y_i,\theta)]$$

$i_t \sim \text{Unif}\{1,\cdots,n\}$

$\rightarrow$ At each iteration t, we draw a point $i_t \sim T$ and define

$$\theta_t = \theta_{t-1} - \eta_t \underbrace{\nabla \ell(x_{i_t}, y_{i_t}, \theta_{t-1})}_{g_t(\theta_{t-1})}$$

↳ (or also a "minibatch" of $k \ll n$ points)

Stochastic approximation of the gradient

Q : Is SGD a descent method ?

A : No ! Updates might increase the loss , but should decrease "on average"

Some key questions :

(*) Underlying assumptions that make SGD valid ?

(*) Role of the learning rate η_t ?

(*) Performance in convex functions ?

Key Assumptions :

(i) Unbiased Gradient Descent : $\mathbb{E}[g_t(\theta_{t-1})|\theta_{t-1}] = \nabla F(\theta_{t-1})$

eg. $g_t(\theta_{t-1}) = \nabla \ell(x_{i_t}, y_{i_t}, \theta_{t-1})$

F : objective function

$$\mathbb{E}[g_t(\theta_{t-1})|\theta_{t-1}] = \frac{1}{n}\sum_{i=1}^n \nabla \ell(x_i, y_i, \theta_{t-1}) = \nabla F(\theta_{t-1}) \quad ✓$$

(ii) Variance Control : $\|g_t(\theta_{t-1})\|^2 \le B^2$ a.s.

FML Lecture 22 : SGD

Recap : → Stochastic Gradient Descent

$\theta_t = \theta_{t-1} - \eta_t\, g(\theta_{t-1})$, $g(\theta_{t-1})$: estimator of gradient of $F(\theta_{t-1})$ at θ_{t-1}

→ Main Example : $F(\theta) = \mathbb{E}_X[\ell(\theta, X)]$

and $g(\theta_t) = \nabla_\theta\, \ell(\theta_t, X_t)$ (gradient w.r.t. a single sample)

→ Today : understand the role learning rate η_t

$\boxed{\bullet \text{ Problem Set-up :}}$

→ $X \sim P$ in $\mathbb{R}^d$ s.t. $\mathbb{E}_P(X) = \theta^*$, $\mathbb{E}_P\left[\|X - \theta^*\|^2\right] = \sigma^2 < +\infty$

→ Define $F(\theta) = \frac{1}{2}\mathbb{E}_P\left[\|X - \theta\|^2\right]$ Global Min $\theta^* = \mathbb{E}_P[X]$

→ Goal : Minimize F using SGD :

At iteration t, we draw $X_t \sim P$ (indep. from all previous data)

$$\theta_t = \theta_{t-1} - \eta_t \underbrace{\nabla_\theta\left[\tfrac{1}{2}\|X_t - \theta\|^2\Big|_{\theta = \theta_{t-1}}\right]}_{\theta_{t-1} - X_t}$$

$$= (1 - \eta_t)\,\theta_{t-1} + \eta_t X_t$$

$\theta_1 = X_1$

$\theta_2 = \frac{1}{2}X_1 + \frac{1}{2}X_2$

$\theta_3 = \frac{2}{3}\theta_2 + \frac{1}{3}X_3 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$

$\vdots$

Q : How to pick η_t ?

Idea 1 : If $\eta_t = 1/t$, then $\theta_t = \frac{1}{t}\sum_{j=1}^{t} X_j$

We can see this by induction :

$$\theta_t = (1 - \eta_t)\,\theta_{t-1} + \tfrac{1}{t}X_t = \left(1 - \tfrac{1}{t}\right)\tfrac{1}{t-1}\sum_{j=1}^{t-1} X_j + \tfrac{1}{t}X_t = \tfrac{1}{t}\sum_{j=1}^{t} X_j$$

Idea 2 : If $\eta_t = \frac{2}{t+1}$, then $\theta_t = \frac{1}{t(t+1)}\sum_{j=1}^{t} j\cdot X_j$

↳ ex : Check recurrence

→ Q : Principled way to select learning rate ?

→ From $\theta_t = (1 - \eta_t)\,\theta_{t-1} + \eta_t X_t$

We have a recurrence error : $\theta_t - \theta^* = (1 - \eta_t)(\theta_{t-1} - \theta^*) + \eta_t (X_t - \theta^*)$

$$\Rightarrow \theta_t - \theta^* = (1 - \eta_t)\left[(1 - \eta_{t-1})(\theta_{t-2} - \theta^*) + \eta_{t-1}(X_{t-1} - \theta^*)\right] + \eta_t (X_t - \theta^*)$$

$= \quad \vdots$ where $\prod_{k=t+1}^{t}(1 - \eta_k) \triangleq 1$

$$= \prod_{j=1}^{t}(1 - \eta_j)(\theta_0 - \theta^*) + \sum_{j=1}^{t}\left(\prod_{k=j+1}^{t}(1 - \eta_k)\right)\eta_j\cdot(X_j - \theta^*) \quad [\text{random}]$$

Using that $X_1, \cdots, X_t$ are i.i.d. ,

$$\mathbb{E}\left[\|\theta_t - \theta^*\|^2\right] = \prod_{j=1}^{t}(1 - \eta_j)^2(\theta_0 - \theta^*)^2 + \underbrace{\sum_{j=1}^{t}\mathrm{Var}\left(\left(\prod_{k=j+1}^{t}(1 - \eta_k)\right)\eta_j\cdot(X_j - \theta^*)\right)}_{\sum_{j=1}^{t}\eta_j^2\prod_{k=j+1}^{t}(1-\eta_k)^2\,\underbrace{\mathrm{Var}(X_j - \theta^*)}_{\sigma^2}}$$

$$= \prod_{j=1}^{t}(1 - \eta_j)^2(\theta_0 - \theta^*)^2 + \sigma^2\sum_{j=1}^{t}\eta_j^2\prod_{k=j+1}^{t}(1 - \eta_k)^2 \quad [\text{deterministic}]$$

Rmk. We can see that as long as we have $\prod_{j=1}^{t}(1 - \eta_j)^2(\theta_0 - \theta^*)^2 \xrightarrow{t \nearrow \infty} 0$ we have $\prod_{j=1}^{t}(1 - \eta_j)(\theta_0 - \theta^*) \xrightarrow{t \nearrow \infty} 0$, So $\theta_t \xrightarrow{a.s.} \theta^*$

So it is asympotically unbiased

Q : Do we have any idea on controlling it to be unbiased non-asymptotically ?

→ To get smaller error as t increases , we need :

(i) forget initial conditions : we need $\prod_{j=1}^{t}(1 - \eta_j)^2 \to 0$ as $t \nearrow \infty$

(ii) Control of the variance : $\sum_{j=1}^{t}\eta_j^2\prod_{k=j+1}^{t}(1 - \eta_k)^2 \to 0$ as $t \nearrow +\infty$

A : (i) Assume $\eta_t \to 0$ as $t \nearrow +\infty$

$$\log\prod_{j=1}^{t}(1 - \eta_j)^2 = 2\sum_{j=1}^{t}\log(1 - \eta_j) \leq -\sum_{j=1}^{t}\eta_j \quad \text{by} \quad \log(1 - \eta_j) \approx -\eta_j$$

$\Rightarrow$ We need $\sum_{j=1}^{t}\eta_j \nearrow +\infty$ as $t \nearrow +\infty$

(ii) Decomposition of Variance term : assume $\eta_t \geq 0$ & is non-increasing & $\eta_1 \leq 1$.

Let $m \in [t]$.

$$\sum_{j=1}^{t}\eta_j^2\prod_{k>j}^{t}(1 - \eta_k)^2 \leq \sum_{j=1}^{t}\eta_j^2\prod_{k>j}^{t}(1 - \eta_k)$$

$$= \sum_{j=1}^{m}\eta_j^2\prod_{k>j}^{t}(1 - \eta_k) + \sum_{j=m+1}^{t}\eta_j^2\prod_{k>j}^{t}(1 - \eta_k)$$

$$\leq \prod_{k>m}^{t}(1 - \eta_k)\sum_{j=1}^{m}\eta_j^2 + \eta_m\sum_{j=m+1}^{t}\eta_j\prod_{k>j}^{t}(1 - \eta_k) = (*)$$

↳ $\prod_{k>j}^{t}(1 - \eta_k) \leq \prod_{k>m}^{t}(1 - \eta_k)$ for $m \geq j$ ($(1 - \eta_k) \in [0, 1]$)

& $\eta_j^2 \leq \eta_m\cdot\eta_j$ for $m < j$. (non-decreasing)

Since $\prod_{k>j}^{t}(1 - \eta_k) = \exp\left(\log\prod_{k>j}^{t}(1 - \eta_k)\right) = \exp\left(\sum_{k>j}^{t}\log(1 - \eta_k)\right) \leq \exp\left(-\sum_{k>j}^{t}\eta_k\right)$,

Then $(*) \leq \exp\left(-\sum_{k>m}^{t}\eta_k\right)\sum_{j=1}^{m}\eta_j^2 + \eta_m\sum_{j=m+1}^{t}\underbrace{\left(1 - (1 - \eta_j)\right)\prod_{k>j}^{t}(1 - \eta_k)}_{\underbrace{\prod_{k>j}^{t}(1-\eta_k)}_{a_j} - \underbrace{\prod_{k\geq j}^{t}(1-\eta_k)}_{a_{j-1}}}$

By observation , $\sum_{j=m+1}^{t}(a_j - a_{j-1}) = \overbrace{a_t}^{1} - a_m$

$$= \exp\left(-\sum_{k>m}^{t}\eta_k\right)\sum_{j=1}^{m}\eta_j^2 + \eta_m\left(1 - \prod_{k\geq m}^{t}(1 - \eta_k)\right)$$

$$\leq \exp\left(-\sum_{k>m}^{t}\eta_k\right)\sum_{j=1}^{m}\eta_j^2 + \eta_m \quad \text{for } m \leq t$$

Recall that $\sum_{j=1}^{+\infty}\eta_j = +\infty$ from the bias term ,

$\Rightarrow$ In particular , if we consider $\sum_{j=1}^{+\infty}\eta_j^2 < +\infty$ & $\sum_{j=1}^{+\infty}\eta_j = +\infty$

By picking $m = t/2$, we have variance $\searrow 0$ as $t \nearrow +\infty$

→ Remark : ① It's a careful balance between forgetting I.C. ($\sum_j \eta_j = +\infty$) & controlling overall variance ($\sum_j \eta_j^2 < +\infty$)

② Our previous examples of choosing $\eta_t = 1/t$ or $2/t+1$ make sense

③ $\sum_{j=1}^{+\infty}\eta_j^2 < +\infty$ is sufficient but not necessary .

Even constant learning rate is valid (if we perform averaging of iterates) ?

§ SGD in action : The Perceptron

Consider a dataset $\{(x_i, y_i)\}_{i\in[n]}$ with $x_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$

We want to train a linear classifier : $\hat{y}(x) = \mathrm{sign}(\langle x, \theta\rangle)$

Perceptron Algorithm : [Rosenblatt 1950 s]

(·) Start from $\theta_0 = 0$

(·) At each step $t = 0, 1, 2, \cdots$

↳ Select a random example $i \in [n]$

→ If $y_i\langle x_i, \theta_t\rangle < 1$ → mistake

$\theta_{t+1} = \theta_t + y_i x_i$

Otherwise $\theta_{t+1} = \theta_t$

Rmk. If we made a mistake (wrong side / too small margin) , we push it to the right side by adding $y_i\langle x_i, y_i x_t\rangle = y_i^2\|x_i\|^2 = \|x_i\|^2$

FML Lecture 23 : The perceptron and SGD

Recap : Given a dataset : $S = \{(x_i, y_i)\}_{i\in[n]}$ with $x_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$

We train a linear classifier : $x \mapsto \text{sign}(\langle x, \theta\rangle)$, $\theta \in \mathbb{R}^d$

Using a perceptron :

(i) Initialize $\theta_0 = 0$

(ii) At each iteration t, select a sample i_t

if $y_{i_t} \cdot \langle x_{i_t}, \theta_t\rangle < 1$, then $\theta_{t+1} = \theta_t + y_i x_i$

Otherwise $\theta_{t+1} = \theta_t$

Q : Link between perceptron & SGD ?

Recall : The hinge loss : $\ell(y\hat{y}) = \max(1 - y\hat{y}, 0)$

$\to$ This defines the empirical loss : $L(\theta) = \sum_{i=1}^{n} \ell(y_i \langle x_i, \theta\rangle)$

$\to$ Consider SGD on this empirical loss : $\theta_{t+1} = \theta_t - \gamma_t \nabla_\theta \ell(y_{i_t}\langle x_{i_t}, \theta\rangle)$

$$= \begin{cases} -y_{i_t} x_{i_t} & \text{if } 1 - y_{i_t}\langle x_{i_t}, \theta_t\rangle > 0 \\ 0 & \text{otherwise} \end{cases}$$

$\to$ If pick $\gamma_t = 1$ for $\forall t$, then we get perceptron !!!

Q : Does the perceptron learn ?

Q1 : Can it fit the training set ?

Q2 : Will it correctly classify a test data point ?

Assumption : Dataset is linearly separable

$\to$ Recall : notion of margin of a seperating hyperplane

$H_\theta = \{x ; \langle x, \theta\rangle = 0\}$ and a dataset $S = \{\underbrace{(x_i, y_i)}_{z_i}\}_{i\in[n]}$

Define the margin : $\gamma(S, \theta) = \min_{i\in[n]} \dfrac{y_i\langle x_i, \theta\rangle}{\|\theta\|} > 0$

$\gamma(S) = \max_\theta \gamma(S, \theta)$ & $\theta^* = \arg\max_\theta \gamma(S, \theta)$

$\to$ Define $D(S) = \max_{i\in[n]} \|x_i\|$

(For Q1)

$\to$ Thm. The perceptron algorithm makes at most $\dfrac{2 + D(S)^2}{\gamma(S)}$ margin mistakes on any linearly-separable dataset S

pf. Main Idea : Controlling # of mistakes ~ Controlling how much can θ_t change, in fact, $\|\theta_t\|$ suffices

Upper bound : Sps. we made a mistake at iteration t :

$$\|\theta_{t+1}\|^2 = \|\theta_t + y_{i_t} x_{i_t}\|^2 = \|\theta_t\|^2 + \underbrace{\|x_{i_t}\|^2}_{\le D(S)^2} + \underbrace{2y_{i_t}\langle x_{i_t}, \theta_t\rangle}_{2}$$

So, m_t = # of margin mistakes after t iterations

$\|\theta_t\|^2 \le m_t (D(S)^2 + 2)$

Lower bound : Let θ be any unit vector s.t. H_θ is a seperating hyperplane

If we make a mistake at step t :

$\langle \theta, \theta_{t+1} - \theta_t\rangle = \langle \theta, y_{i_t} x_{i_t}\rangle \ge \gamma(S, \theta)$

In particular, $\langle \theta^*, \theta_{t+1} - \theta_t\rangle \ge \gamma(S)$ & $\|\theta^*\| = 1$

$$\|\theta_t\| \ge \langle \theta_t, \theta^*\rangle \overset{\text{telescope \& } \theta_0 = 0}{=} \sum_{j=1}^{t} \langle \theta_j - \theta_{j-1}, \theta^*\rangle \ge m_t\, \gamma(S)$$

Therefore, $m_t^2 \cdot \gamma(S)^2 \le \|\theta_t\|^2 \le m_t \cdot (2 + D(S)^2)$

$$\Rightarrow m_t \le \frac{2 + D(S)^2}{\gamma(S)^2}$$

#

$\to$ Thus, perceptron eventually correctly classifies all training points

For (Q2) :

Assume datapoints $z_i = (x_i, y_i)$ are drawn i.i.d. from D and test point $z \sim D$ (i.i.d.)

Thm. [Vapnik, Chovronekis]

Assume dataset $S = \{z_1, \cdots, z_n\}$ is linearly separable. Let $\theta(S)$ be the output of the perceptron on S (run until convergence)

Then the prob. of making a margin mistake on $z = (x, y)$ satisfies

$$\mathbb{P}(y\langle \theta(S), x\rangle < 1) \le \frac{1}{n+1}\mathbb{E}\left[\frac{2 + D(\bar{S})}{\gamma(\bar{S})}\right] \quad \text{where } \bar{S} = S \cup \{z\}$$

pf. We exploit the exchangeability of the data $\{z_i\} = \{(x_i, y_i)\}_{i\in[n]}$ and $z = (x, y)$

Joint distribution of $x_1, \cdots, x_n$ does not depend on the order

(1) $\mathbb{P}[y\langle \theta(S), x\rangle < 1] = \mathbb{E}[\mathbb{1}_{\{y\langle\theta(S), x\rangle < 1\}}]$

- Define $S^{-k} \triangleq \{z_1, \cdots, z_{k-1}, z_{k+1}, \cdots, z_n, z\}$

Exchangeability : the order of these R.V.'s does not affect the test error

i.e., Running perceptron on S^{-k} and testing on z_k gives the same prob. error for each k

So $\mathbb{P}[y\langle\theta(S), x\rangle < 1] = \frac{1}{n+1}\sum_{k=1}^{n+1}\mathbb{E}[\mathbb{1}_{\{y_k\langle\theta(S^{-k}), x_k\rangle < 1\}}]$

$\to$ Recall that running perceptron on $\bar{S}$ makes at most $m = \dfrac{2 + D(S)^2}{\gamma(S)^2}$ mistakes.

There are at most m indices $i_1, \cdots, i_m \in [n]$ where we have made mistakes $(m \le n)$

If $k \notin \{i_1, \cdots, i_m\}$, then $\theta(\bar{S}) = \theta(S^{-k})$

$\Rightarrow y^k\langle\theta(S^{-k}), x_k\rangle > 1$

Other terms contribute at most 1. So

$$\mathbb{P}[y\langle\theta(S), x\rangle < 1] \le \frac{1}{n+1}\mathbb{E}[m] = \frac{1}{n+1}\mathbb{E}\left[\frac{2 + D(\bar{S})^2}{\gamma(\bar{S})^2}\right]$$

#

Rmk. ① Unlike SVM, perceptron does not necessarily converges to a unique hyperplane

it stops as soon as it makes no mistake

② From the perspective of SGD, the reason we can converge by just choosing a constant convergence rate can be boiled down to the simplicity of our hinge loss

12/3 FoML 25

Tuesday, December 3, 2024   2:03 PM

FML Lecture 25 : Geometric Deep Learning

→ Basic Supervised Learning Set-up :

Input Space $\mathcal{X}$ (high - dimensional)

Output Space $\mathcal{Y}$ (low - dimensional , e.g. , $\mathcal{Y} = \mathbb{R}$)

Hypothesis Class , $\mathcal{F} = \{ f : \mathcal{X} \to \mathcal{Y} \}$, often indexed by a complexity parameter $\mathcal{F}_\delta = \{ f \in \mathcal{F} , \gamma(f) \le \delta \}$

Goal : Approximate unknown target f^* via ERM

$\hat{f} \in \underset{f \in \mathcal{F}_\delta}{\operatorname{argmin}} \frac{1}{n} \sum_i \ell(f(x_i), y_i)$ where we assume $y_i = f^*(x_i) + \varepsilon$

Recall Decomposition of error :

$R(\hat{f}) \le \varepsilon_{approx}(\delta) + \varepsilon_{stat}(\varepsilon)$

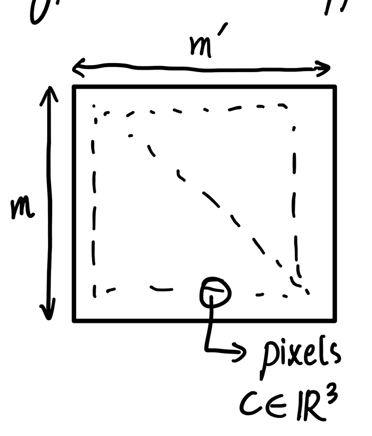
Conclusion : To efficiently learn , we need accurate (ε_{approx} small) yet " small " hypothesis $\mathcal{F}_\delta$ (ε_{stat} small)

⇒ Need to exploit any prior information on target f^*

• Learning in the Physical World

High - dimensional input space $\mathcal{X}$ in typical ML applications ?

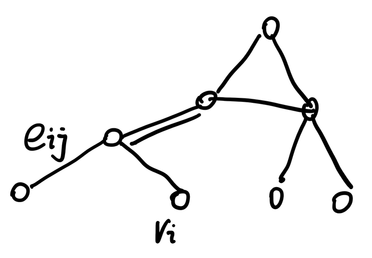
→ $\mathcal{X} = \{ \text{images} \}$ represented as



$x \in \mathbb{R}^{3 \times m \times m'}$

$\boxed{\text{2D - grid}}$ encoding RGB

→ $\mathcal{X} = \{ \text{molecules} \}$ represented as



$\boxed{\text{graphs}}$, encoding the atoms and their chemical bounds

$V_i \in \{ O, C, H, N, \cdots \}$, $e_{ij} \in \mathbb{R}^5$

→ $\mathcal{X} = \{ \text{text / language} \}$, represented as a $\boxed{\text{sequence}}$ $\{ w_1, w_2, \cdots, w_T \} \in$ Dictionary

• $\mathcal{X}$ is in fact a space of signals that live on a physical domain Ω : $\mathcal{X} = \{ x : \Omega \to C \}$ (2D-grid, graphs, sequence) channels

$u \mapsto x(u)$

→ We can add signals or scale them , $(\alpha x + \beta y)(u) = \alpha x(u) + \beta y(u)$ ⇒ $\mathcal{X}$ is a vector space

→ Inner Product Structure in C " upgrades " to inner product structure in $\mathcal{X}$: $\langle x, y \rangle_{\mathcal{X}} = \int_\Omega \langle x(u), y(u) \rangle_C \, du$

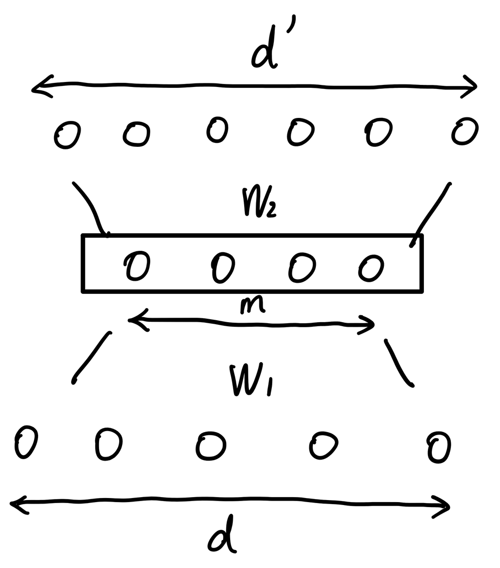
Q : Why is this physical domain useful ?

Symmetry : A symmetry of an object is a transformation that leaves the object unchanged

Ex. 1 :

finite # of symmetries   infinite # of symmetries

Ex 2 : $f(x ; W_1, W_2) = W_2 \rho(W_1 x)$



A symmetry of this architecture is a transformation of parameters $W = \{ W_1, W_2 \}$

s.t. $f(x ; g(W)) = f(x ; W)$ for $\forall W$

$W_1 \in \mathbb{R}^{m \times d}$

$W_2 \in \mathbb{R}^{d' \times m}$

$\sigma : \{1, m\} \to \{1, m\}$ a permutation

Given σ , we define perm. matrix $\Pi_\sigma \in \{0, 1\}^{m \times m}$ wher $(\Pi_\sigma)_{ij} = \begin{cases} 1 & \text{if } \sigma(i) = j \\ 0 & \text{otherwise} \end{cases}$

So $g_\sigma(\{ W_1, W_2 \}) = \{ \Pi_\sigma W_1 , W_2 \Pi_\sigma^T \}$ is a symmetry of f s.t. $f(x ; g_\sigma(W)) = f(x, W)$ for $\forall \sigma, \forall x, \forall W$

Rmk. ① Permutation Symmetry is indep. of the form σ !! (So we at least have $m!$ symmetries for one - layer NN)

② e.g. $\rho(t) = t$, then $f(x ; g(W)) = W_2 \Pi_\sigma^T \Pi_\sigma W_1 x = W_2 W_1 x = f(x ; W)$

So , it's natural to think : if we have some assumptions on ρ , we can explore more symmetries (like orthogonal)

(consider homogeneous functions like ReLU)

→ Most importantly , we are interested in symmetries of the target $f^* : \mathcal{X} \to \mathcal{Y}$

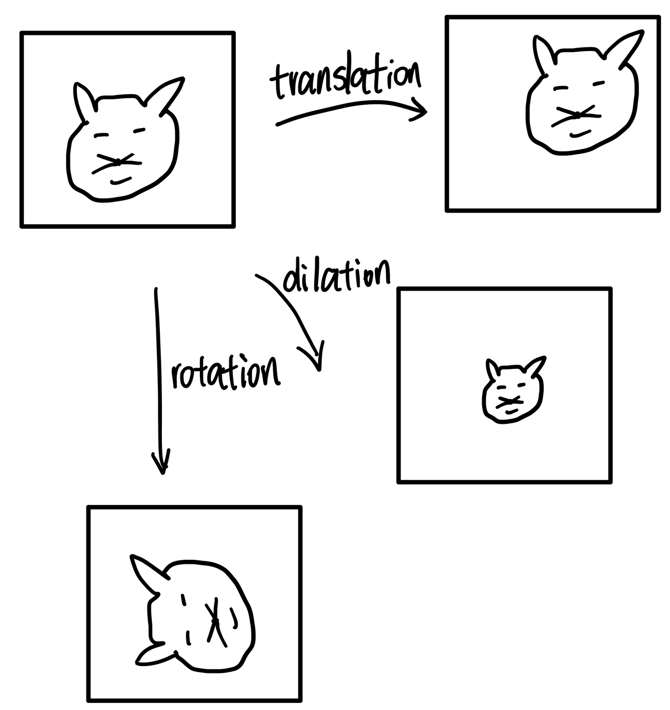
↳ Transformations $g : \mathcal{X} \to \mathcal{X}$ s.t. $f^*(g(x)) = f^*(x)$ for $\forall x \in \mathcal{X}$

↳ Challenging in genearic high - dimension !

↳ Instead , use physical domain Ω to describe symmetries !

• Symmetries of Ω

→ Images $f^*(x)$ = Is there a cat in x ?



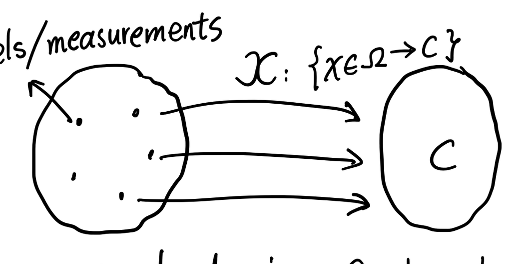
12/5 FoML 26

Thursday, December 5, 2024 2:01 PM

FML Lecture 26 : Learning with Symmetries

Recap : Learning in Physical World

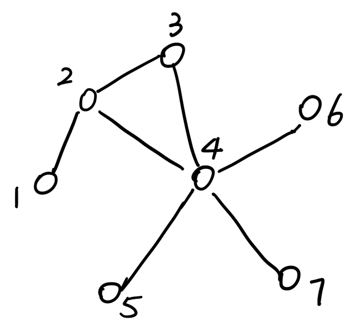
pixels/measurements $\mathcal{X} : \{x \in \Omega \to C\}$



Ω : signal domain C : channels ($\subseteq \mathbb{R}^n$)

Physical domain enables a description of symmetries

Symmetries arising on graphs (motivation : molecules , traffic network)



(Undirected) Graph : $G = (V, E)$

$V = \{1, \cdots, n\}$, n = size of the graph

$E = \{ (i,j) , i, j \in V \}$

Represented with Adjacency Matrix :

$A \in \{0,1\}^{n \times n}$, where $A_{ij} = \begin{cases} 1 , & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$

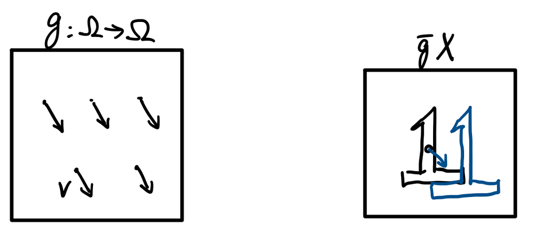
G of size $n \xrightarrow{\text{encoding}} A$

↳ We can relabel the vertices in $n!$ ways , while all the adjacency matrices are related by : $\bar{A} = \Pi A \Pi^T \iff \bar{A}$ obtained by permuting rows & columns of A

• From Symmetries of Ω to the symmetries of $\mathcal{X}$

A domain transformation $g : \Omega \to \Omega$ defines a transformation $\bar{g} : \mathcal{X} \to \mathcal{X}$ by : $(\bar{g}X)(u) = X(g^{-1}u)$ for $\forall u \in \Omega$

→ $\bar{g}$ defines a linear transformation $\bar{g}(\alpha X + \beta Y) = \alpha \bar{g}(X) + \beta \bar{g}(Y)$



$(\bar{g}X)(u) = X(u - v)$

• Symmetries & Groups

We observe that (i) $g = Id$ is a symmetry

(ii) g and h are symmetries , then $g \circ h$ and $h \circ g$ are also symmetries

(iii) If g is a symmetry , then its inverse g^{-1} is also a symmetry

Symmetries form a $\boxed{\text{group}}$ using composition

↳ Groups can either be discrete (finite elements) or continuous

Ex. → $\mathbb{Z}_q$: cyclic group of integers modulo q

→ Rubik's Cube

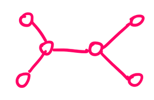
→ $(\mathbb{R}, +)$, $(\mathbb{R} \setminus \{0\}, \times)$

• Summary so far : (i) We use physical domain Ω to define a group G of transformations

(ii) This group defines a symmetry group of the target function f^* : $f^*(g \cdot x) = f^*(x)$ for $\forall g \in G$, $x \in \mathcal{X}$

→ In many ML applications , we have prior knowledge of (some) symmetries of the target

Examples :

Arithmetic	Symmetry
$x_1 + x_2 = ?$	Commutative Structure
Image Classification	Orthogonal Group O_n
Proteins / Biology  (n atoms)	Permutation Group S_n & Orthogonal Group O_n

Q1 : Why symmetries are useful for learning ?

Q2 : How to leverage them in practice ?

• Invariant Learning

→ Let $f^* : \mathcal{X} \to \mathcal{Y}$ be the target function

→ we consider symmetry group G acting on $\mathcal{X}$

→ We say that f^* $\boxed{\text{is invariant to } G \text{ (or } G\text{-invariant)}}$ if $f^*(g \cdot x) = f^*(x)$ for $\forall g \in G$, $x \in \mathcal{X}$

→ assume w.l.o.g. , G is discrete

→ Given any $f : \mathcal{X} \to \mathcal{Y}$, we define the average w.r.t. G as $Sf : \mathcal{X} \to \mathcal{Y}$ s.t. $\boxed{Sf(x) = \frac{1}{|G|} \sum_{g \in G} f(g \cdot x)}$ for $\forall x \in \mathcal{X}$

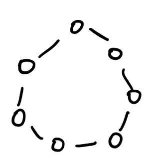
$\{g \cdot x , g \in G\} = O(x)$: Orbit of G passing through $\mathcal{X}$

↳ So S is thus averaging over group orbits

→ $Sf^* = f^*$ as $f^*(g \cdot x) = f^*(x)$ for $\forall x \in \mathcal{X}$, $\forall g \in G$

→ Given a hypothesis class $\mathcal{F}$, we can make it G-invariant : $S\mathcal{F} = \{ Sf , f \in \mathcal{F} \}$ (as long as g & g^{-1} acts transitively on $\mathcal{F}$)

Eg. $\Omega = \{1, \cdots, q\}$ for $q > 2$ & q is prime . $G = \mathbb{Z}_q$



$\mathcal{F} = \{$ polynomials $p(x_1, \cdots, x_q)$ of degree $k \}$ $(x_i : i \to C)$

Eg. for $k = 2$, $\mathcal{F}$ could contain x_1^2, $x_1^2 + x_2 x_3$, $\cdots$

$S\mathcal{F} = \{ Sp(x_1, \cdots, x_q) = \frac{1}{q} \sum_{j=1}^{q} p(x_{1+j}, x_{2+j}, \cdots, x_{q+j}) , p \text{ polynomials} \}$

Essential Idea : $\mathcal{X} = G \otimes \underline{\underline{\mathcal{X}/G}}$

↳ and we only need this information

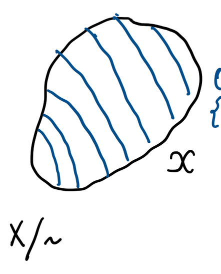
we pick Sf as the representative of each equivalence class $[f]$

12/10 FoML 27

Tuesday, December 10, 2024 1:59 PM

FML Lecture 27 : The Geometric DL Blueprint

Recap : Decomposition of input space into orbits



G : group of transition acting on $\mathcal{X}$

$O(x) := \{g\cdot x,\ g\in G\}$ group orbit ; $x\sim x'$ iff they are on the same orbit

↳ If the target $f^*: \mathcal{X}\to\mathcal{Y}$ is G - invariant , then f^* can be viewed as a function $X/\sim \ \longrightarrow \mathcal{Y}$ instead

→ Averaging operator $Sf(x) = \frac{1}{|G|}\sum_{g\in G} f(g\cdot x)$ maps arbitary hypothesis space $\mathcal{F} = \{f: \mathcal{X}\to\mathcal{Y}\}$ into an invariant hypothesis class : $S\mathcal{F} = \{Sf,\ f\in\mathcal{F}\}$

↳ Verify that Sf is G - invariant : $Sf(g\cdot x) = Sf(x)$ for $\forall x\in\mathcal{X}$, $\forall g\in G$

$Sf(g\cdot x) = \frac{1}{|G|}\sum_{g'\in G} f(g'\cdot(g\cdot x)) = \frac{1}{|G|}\sum_{\tilde g\in G} f(\tilde g\cdot x) = Sf(x)$ #

→ When f^* is G - invariant , should we use $\mathcal{F}$ or $S\mathcal{F}$?

• Approximation Error : $\inf_{f\in\mathcal{F}} \|f^*-f\|^2$ v.s. $\inf_{\tilde f\in S\mathcal{F}} \|f^*-\tilde f\|^2$

Fact : The averaging operator S is an orthogonal projection

pf. for $\forall h: \mathcal{X}\to\mathcal{Y}$: $\|h\|^2 = \|Sh + (I-S)h\|^2$

$= \|Sh\|^2 + \|(I-S)h\|^2 + 2\langle Sh, (I-S)h\rangle$

As $\langle Sh, (I-S)h\rangle = \int_{\mathcal{X}\backslash\sim}\Big(\int_{O(x)} Sh(\bar x)\cdot(I-S)h(\bar x)\, \underset{\substack{\downarrow \\ \text{average dummy variable}}}{d\bar x}\Big)dx = \int_{X/\sim} h(x)\underbrace{\Big[\int_{O(x)} h(\bar x)\,d\bar x - h(x)\Big]}_{0} d\bar x = 0$ (*)

So S is an orthogonal projection.

#

By this fact, we have

$$\|f^* - f\|^2 \overset{\text{Fact}}{=} \|S\overset{f^*}{\overbrace{f^*}} - Sf\|^2 + \|\overset{0}{\overbrace{(I-S)f^*}} - (I-S)f\|^2$$

$$= \|f^* - Sf\|^2 + \|(I-S)f\|^2$$

$$\geq \|f^* - Sf\|^2$$

$$\Rightarrow \inf_{\tilde f\in S\mathcal{F}} \|f^* - \tilde f\| \leq \inf_{f\in\mathcal{F}} \|f^* - f\|^2$$

→ So Approximation error is not degraded (if $S\mathcal{F}\subseteq\mathcal{F}$, then they're equal)

→ Statistical Error ?

$S\mathcal{F}$ is defined over smaller space $X/\sim$, so stat. error is not degraded either

→ Using Symmetries helps the learning task

→ The larger the symmetry group , the smaller the quotient space $X/\sim$

• Big caveat so far : Computing Sf is expensive , especially as $|G|$ is large , even $|G|$ is ∞ !

↳ Q : Efficient Algorithm ?

• The Geometric DL Blueprint

Consider a linear hypothesis f :

$Sf(x) = \frac{1}{|G|}\sum_{g\in G} f(g\cdot x) = f\Big(\Big[\frac{1}{|G|}\sum_{g\in G} g\Big]\cdot x\Big) = f(\bar x)$ where $\bar x$: group average of x

→ The group average can be computed efficiently in our cases of interest

Eg 1. $\mathcal{X} = \{x: \Omega\to\mathbb{R}\}$, $\Omega = \{1,\cdots,m\}$. $G = S_m$

$\simeq \mathbb{R}^m$

$\bar X\in\mathcal{X}\simeq\mathbb{R}^m$ and $\bar X_j = \frac{1}{m!}\sum_{\sigma\in S_m}(\sigma\cdot X)_j = \frac{1}{m}\sum_{i=1}^m X_j$ → simple average over all cordinates !

easy ~

Eg 2. $\mathcal{X} = \{x: \mathbb{R}^2\to\mathbb{R}\}$, G = Translation Group .

$\bar X(u) = \int_G (g\cdot x)(u) = \int_G x(u-v)\,dv = \int_G x(l)\,dl$ → average of image

Problem : The averaging loses too much information !

How to complement the linear invariant ?

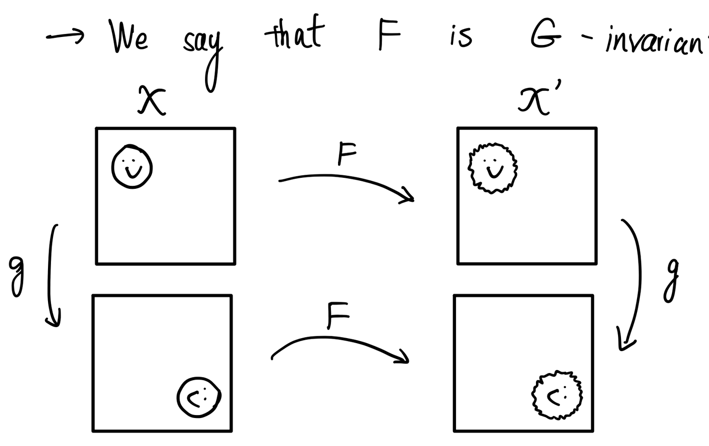
• From invariance to equivariance

→ Sps. we have $F: \mathcal{X}\to\mathcal{X}'$ s.t. G acts on both $\mathcal{X}$ and $\mathcal{X}'$

Eg. $X' = X$ = image , $F(x) = x$ with Van Gogh style

Rmk. X, X' can differ , Eg. $X' = \Omega$, $F(x)$ = location of certain object

→ We say that F is G - invariant if $F(g\cdot x) = g\cdot F(x)$ for $\forall x\in\mathcal{X}$ and $\forall g\in G$



(Commutative diagram)

Q : How to compute linear equivariants ?

We start with G : Translation Group in $\Omega = \mathbb{R}^2$, $X = \{x: \Omega\to\mathbb{R}\}$, $F: X\to X$ and F is linear and commutes with translations

$\Rightarrow$ F is a convolution : $(Fx)(u) = \int_\Omega x(v)\,h(u-v)\,dv$ where $h: \Omega\to\mathbb{R}$ is a filter

Verify linearity & commutativity w/ translation

$F(\alpha x + \beta x') = \alpha F(x) + \beta F(x')$ for $\forall x, x'\in\mathcal{X}$ & $\alpha, \beta\in\mathbb{R}$

$F(g\cdot x)(u) = \int_\Omega (g\cdot x)(v)\,h(u-v)\,dv$

↳ translation by u_0

$= \int_\Omega x(v-u_0)\,h(u-v)\,dv$

change of variables

$= \int_\Omega x(v')\,h(u-u_0-v')\,dv' = g\cdot(Fx)(u)$

Eg 2. $F(x)(u) = \sigma(x(u))$ pointwise transformation for $\forall x$, $\forall u$

$F(g\cdot x)(u) = \sigma(x(g^{-1}u))$

$[g\cdot F(x)](u) = g\cdot\sigma(x(u)) = \sigma(x(g^{-1}u))$

$F_1 : X\to X'$
$F_2 : X'\to X''$ are both equivariant
F_3 is invariant

$\Bigg\}\Rightarrow$ $F_2\circ F_1$ is also equivariant ; $F_3\circ F_1$ is invariant