Subject PrePoc2 2/4 - 07 Date /

Attendants

DL Book 5.2

capacity, overfitting, underfitting

訓練集合のみ観測できるとき、テスト集合にどう影響を与えるか → statistical learning theory.

訓練データとテストデータは

data generating process と

呼ばれるデータ集合の確率分布から生成される

iid assumption

各データ集合の事例が互いに独立であり (independent)

訓練とテストの集合が同一の分布に従う

ことを仮定するのが一般的 (identically distributed)

機械学習アルゴリズムがどの程度うまくいくか?

1. train loss を小さく

2. train loss と test loss を小さく.

Subject Date 2

Attendants

過小適合 = 未学習 (underfitting)

過剰適合 = 過学習 (overfitting)

↑ train loss が十分下がらない

model の capacity で control できる

↳ 勾配の関数に適合する能力

この control あたりは hypothesis space を

選ぶ、ということ

The model specifies which family of functions

the learning algorithm can choose

from when varying the parameters

in order to reduce a training objective,

Subject Date 3

Attendants

関代ではないので

(truly objectiveをいうために paramが変わるところに)

whichみたいに考える

The model specifies which family of function the learning algorithm can choose from

which という関代 which はここ

e.g. learning algo が
↳ early stopping あり

H w/ early

H w/o early

なし

Hにかかることを表す

family

選択肢などの of function

Subject Date 4

Attendants

モデルの表現能力 (representation capacity)

関数集合から最適な関数

を発見するのは難しい

↓

早に訓練損失を大幅に減少させる

関数を発見

Early Stopping

とかで制限される

学習アルゴリズムの effective capacity が

モデルの表現力 (representational capacity) に等る

= 表現容量

可能性がある.

Occam's razor

統計的学習 (SLT: Statistical Learning Theory) は

representation capacity を定量化する方法が

いくつもあり e.g. VC次元 (Vapnik-Chervonenkis Dim.)

Subject Date 5

Attendants

representation capacity で定量化

↓

汎化 gap は モデル容量 が増えることで 定量化

train sample が増えることで

汎化 gap は減少する量によって

上に有界である

この Bound が ML が機能するための根拠

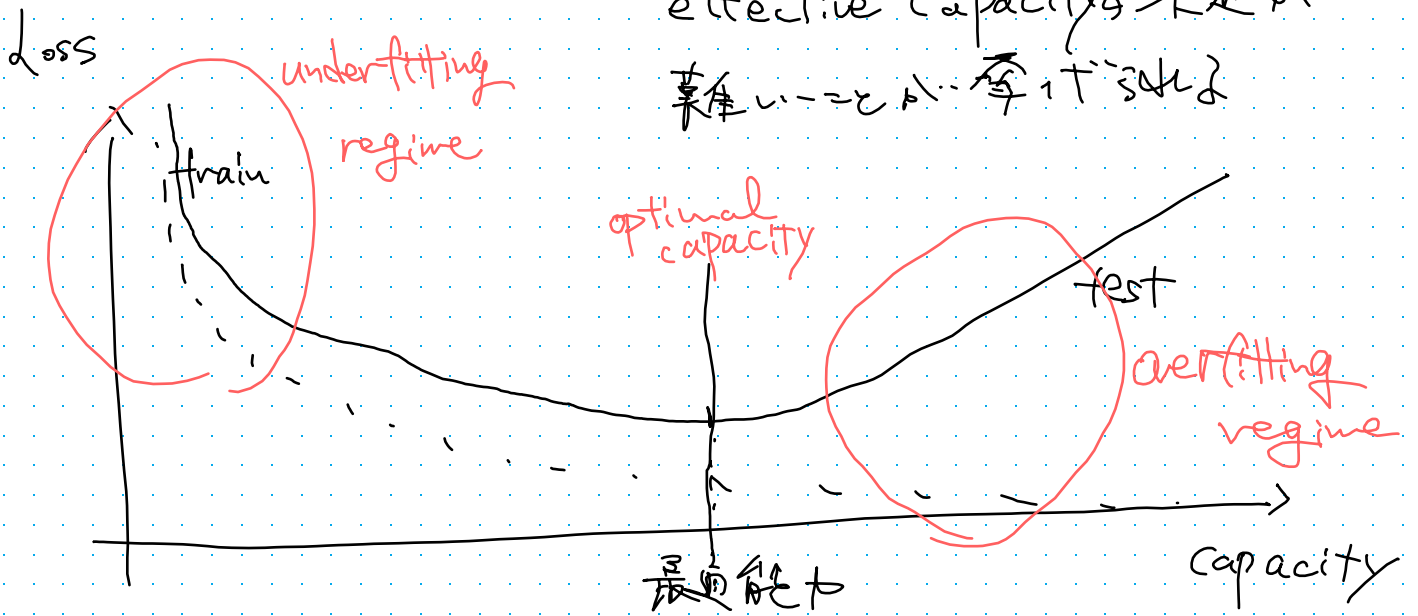
↑

Deep の場合は使われない

∵ Bound が strict でない & DL algo での

effective capacity の決定が

難しいことが等しくされる



Subject Date 6

Attendants

任意の高い capacity という最も極端な状態に達するには

non-parametric モデルの概念を導入する

↓

実際には実装できない理論的な抽象化になるも、

↔ 複雑さを訓練集合のサイズの関数にすることで

実用的な non-parametric モデルを設計することもできる

e.g. nearest neighbor regression

$\hat{y} = y_i$ である $i = \underset{i}{\operatorname{argmin}} \| X_{i,:} - x \|$ のときの

最近傍にひもづけられている y によって

すべての $X_{i,:}$ に対する y_i の平均を取ることによって

アルゴリズムをひもづけられる

Xとyを保存

テスト点 x を分類する場合，モデルは訓練集合の中の最も近い要素を調べ

$X_{i,:}$ の意味 Pythonのsliceと同じ

$A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 4 & 5 & 6 & 7 \\ 8 & 9 & 10 & 11 \end{pmatrix}$ $A_{1,:} = (4\ 5\ 6\ 7)$

Subject Date 7

Attendants

$e_j = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}^T$ ← j番目

$e_j^T W = W_{j,:}$

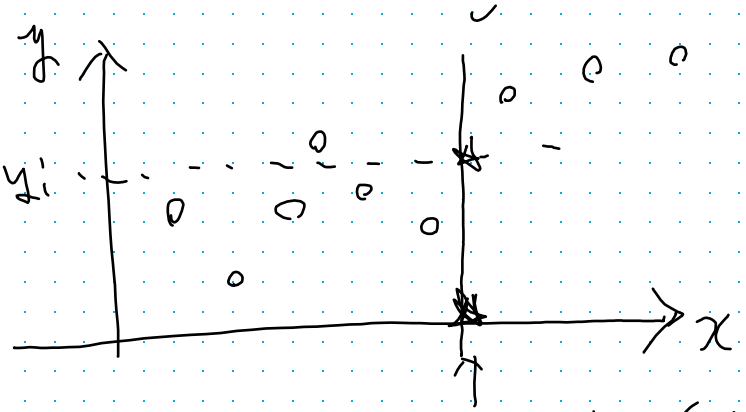
j番目の W の row

$W = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$

$e_1^T W = (0\ 1)\begin{pmatrix} a & b \\ c & d \end{pmatrix} = (0\times a + 1\times c,\ 0\times b + 1\times d)$

$= (c\ \ d)$

$i = \operatorname{argmin} \|X_{i,:} - x\|_2^2$ のとき $\hat{y} = y_i$ とする



入力 x_i は含まれている

k個のtrain sampleの近い方のyの平均で $\hat{y}$ を推定する

Subject Date

Attendants

理想的なモデルは，
データ生成する真の確率分布を事前に知っている oracle

pcxayっぷる Oracle が予測した際に生じる誤差を
ベイズ誤差（Bayes Error）という

train sample ↑ 汎化が悪くなることはない
ノンパラでは良くなる
パラメトリックはいずれも Bayes Error
に漸近する

訓練集合のサイズが大きくなると
train loss が増える（適合が難しくなるから）
test loss は 訓練データとつじつまの合う仮説が減るため

最適な容量でのテスト誤差は Bayes Error に漸近

訓練誤差は memorizing できる場合があるので
ベイズ誤差を下回ることもある

Subject Date 9

Attendants

train sample が 増えるほど

optimal capacity が大きくなるが

後にサチル.

no free lunch theorem

どの分類アルゴリズムも

(データ生成分布すべてを平均すると)

過去に観測されていない点を分類する際の

Error Rate は同じ

言い換えると

他の ML アルゴリズムよりも ~~全般的に~~ よいと

言えるアルゴリズムは存在しない

データを生成する全ての分布

を平均した場合に成立する

Subject

Date 10

Attendants

e.g.

Model	MNIST	WILDS	PACS
A	90%	40%	10%
B	50%	30%	60%
C	50%	40%	50%

(こういう感じで 考えるデータセット全てに対し
良い性能とるようなモデルはなく、

test error は全て同じになる

from Dist (全部の平均)

現実世界では確率分布の種類に仮定を設けることで、
その分布に対して良い性能で発揮する学習アルゴを設計できる

Subject Date 11

Attendants

正則化 ノーフリーランチ theorem は特定のタスクに対してうまく機能するMLアルゴリズムを設計しなきゃいけないことを意味してる.

e.g.
weight decay を含めて重みの優先度をつけることができる

↓

一般化 (regularizer) : ペナルティコスト

optimization

この2つが大事.

解きたいタスクに適した正則化の形式を選択する必要がある

Subject Date 12

Attendants

1. (a) Explain the notations of underfitting, overfitting bias, variance, model capacity

underfitting

when model has low capacity,

(both of representation and effective ← num of param, $\sin(x)$? ...)

error of prediction of train sample and unseen samples is still high. in this case.

model is underfitting.

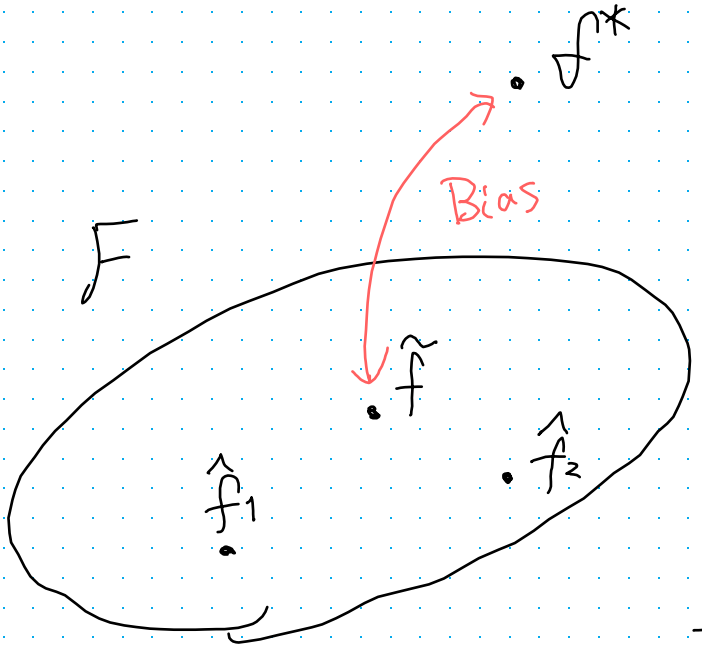
Overfitting A statistical model fits to training sample itself.

(when model capacity is enough to memorize training samples.)

It leads to degrade test error.

---

bias variance



$$f^* = \underset{f}{\operatorname{argmin}} R(f)$$

$$\tilde{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} R(f)$$

$$\hat{f}_1 = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}(f, Q_1)$$

$$R(f) = \mathbb{E}_{x,y \sim Q} P(f(x), y)$$

$$\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} P(f(x_i), y_i)$$

Subject Date 14

Attendants

$$\text{Bias} = \mathbb{E}[\hat{f}(x)] - f(x)$$

$$\text{Variance} = \mathbb{E}[\hat{f}(x)^2] - \mathbb{E}[\hat{f}(x)]^2$$

Bias-Variance Decomposition

汎化誤差の期待値を bias-variance-noise の3つの和に分解.

$$\mathbb{E}[(y - \hat{f}(x))^2]$$

$$= (\text{Bias})^2 + \text{Var} + \sigma^2$$

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

$$f(x) = \underset{f}{\operatorname{argmin}} R$$

Subject Date 15

Attendants PRML (上) p(46

最適な予測

$$h(x) = \mathbb{E}[t|x] = \int t\, p(t|x)\, dt.$$

ここで t はラベルのようなものと考える!?

期待二乗損失は

$$\mathbb{E}[L] = \int \{y(x) - h(x)\}^2 p(x)\, dx$$

$$+ \iint \{h(x) - t\}^2 p(x,t)\, dx\, dt$$

この第二項は $y(x)$ と t 上でデータに含まれる本質的なノイズのみに依存する

この第一項を最小にする関数 $y(x)$ を求める必要がある

分布 $p(t,x)$ に従う 多数のデータ集合 D に対し
学習アルゴリズムを実行
⇒ 予測func $y(x;D)$ をもとめられる
と仮定する

このとき 異なるデータ集合は異なるfuncを学ぶため、
二乗損失も異なる

学習アルゴリズムの性能は
"データ集合のとり方に関する予測の意味で評価される"

$$\{y(x;D)-h(x)\}^2 = \{y(x;D)-\mathbb{E}_D[y(x;D)]+\mathbb{E}_D[y(x;D)]-h(x)\}^2$$

これに対して D で期待値をとると

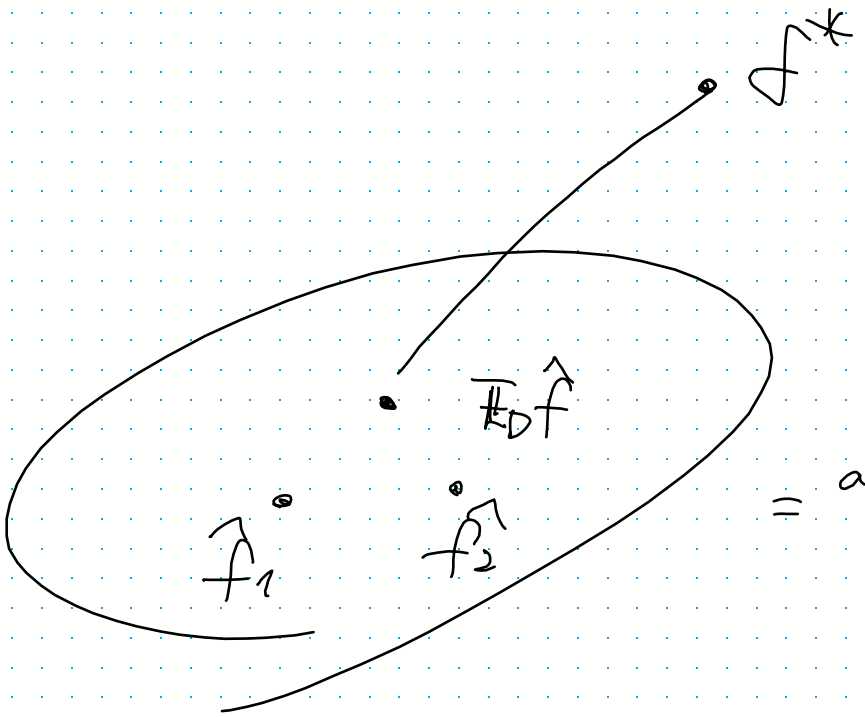
$$\mathbb{E}_D\left[\{y(x;D)-h(x)\}^2\right] = \{\mathbb{E}_D[y(x;D)]-h(x)\}^2 + \mathbb{E}_D\left[\{y(x;D)-\mathbb{E}_D[y(x;D)]\}^2\right]$$

となる

Subject

Date 17

Attendants



$$f^* = \underset{f}{\operatorname{argmin}} \; R(f)$$

$$= \underset{f}{\operatorname{argmin}} \; \mathbb{E}_{x,y \sim D}\left[p(f(x), y) \right]$$

$$\tilde{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; R(f)$$

$$= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathbb{E}_{x,y \sim D}\left[p(f(x), y) \right]$$

$$\hat{f}_1 = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \hat{R}(f, D_1)$$

$$= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathbb{E}_{x,y \sim D_1}\left[p(f(x), y) \right]$$

$$= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \frac{1}{M} \sum_{m=1}^{M} p(f(x), y)$$

Subject Date 18

Attendants

$$\mathbb{E}_D\left[\{y(x;D)-h(x)\}^2\right] = \{\mathbb{E}_D[y(x;D)]-h(x)\}^2 \leftarrow \text{bias}^2$$

$$+\mathbb{E}_D\left[\{y(x;D)-\mathbb{E}_D[y(x;D)]\}^2\right] \leftarrow \text{var}$$

Because

$$\{y(x;D)-h(x)\}^2 = \{y(x;D)-\mathbb{E}_D[y(x;D)]$$

$$+\mathbb{E}_D[y(x;D)]-h(x)\}^2$$

$$= \{y(x;D)-\mathbb{E}_D[y(x;D)]\}^2 + \{\mathbb{E}_D[y(x;D)-h(x)\}^2$$

$$-2\{y(x;D)-\mathbb{E}_D[y(x;D)]\}\{\mathbb{E}_D[y(x;D)-h(x)]\}^2$$

take expectation over D → 0

$$\mathbb{E}_D\left[\{y(x;D)-h(x)\}^2\right] = \{\text{Bias}\}^2 + \{\text{Variance}\}$$

Subject

Date 19

Attendants

$$\mathbb{E}_D\left[\{y(x;D) - h(x)\}^2\right] = \{\mathbb{E}_D[y(x;D)] - h(x)\}^2 \leftarrow \text{bias}^2$$

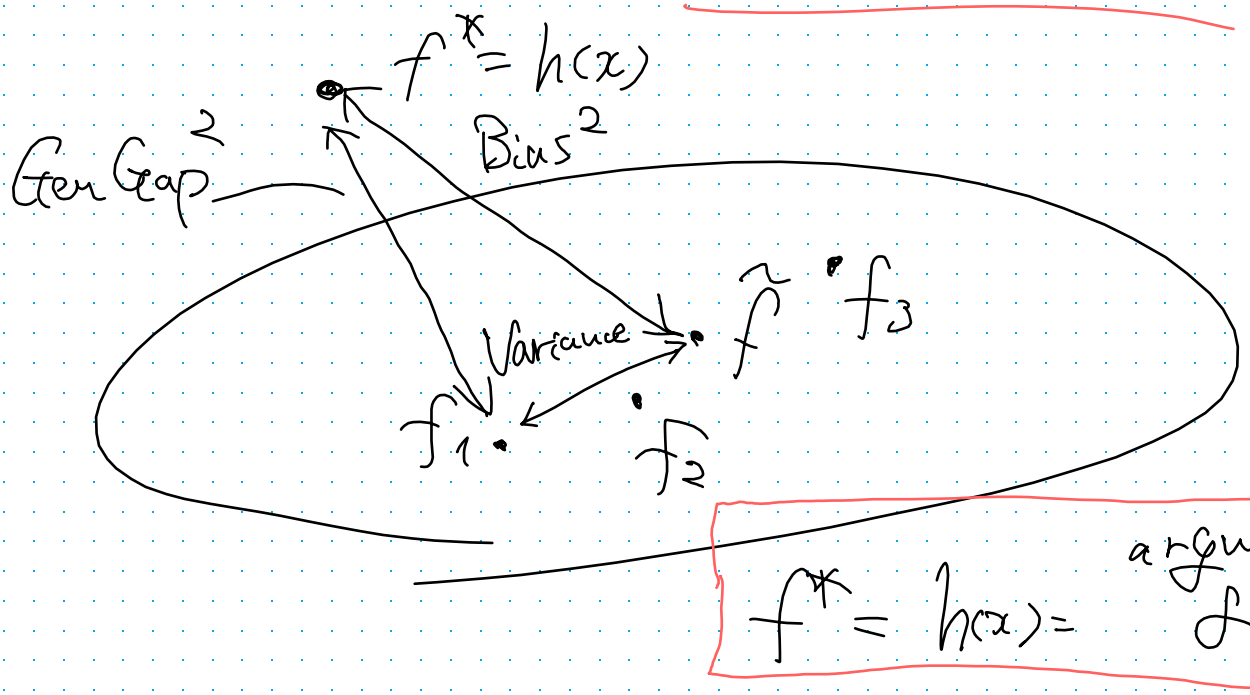
$$+ \mathbb{E}_D\left[\{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2\right] \leftarrow \text{var}$$

$$f_1 = y(x;D_1) = \boxed{\underset{f \in \mathcal{F}}{\operatorname{argmin}}\; y(x;D_1)}$$

$$\vdots$$

$$f_N = y(x;D_N)$$

$$\mathbb{E}_D[y(x;D)] = \boxed{\tilde{f} = \frac{1}{N}\sum_{i=1}^{N} f_i}$$

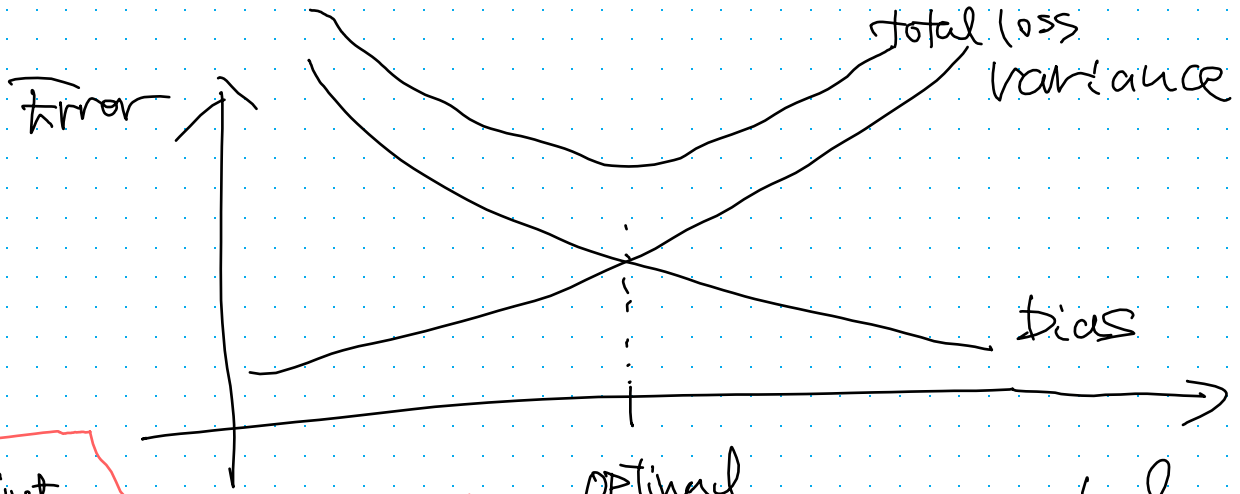


$$\boxed{f^* = h(x) = \underset{f}{\operatorname{argmin}}\; y(x,D)}$$

Subject Date 20

Attendants

7 (b) How do these related to one another



at first, in general. The more increased, the more the bias decreased

if model capacity is lower than optimal capacity, model is not ~~fit~~ train samples. It is called underfitting regime.

when optimal capacity is capacity when min total loss is achieved.

Subject Date 2/

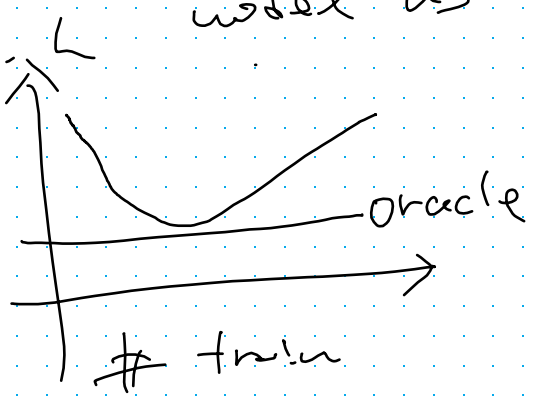
Attendants

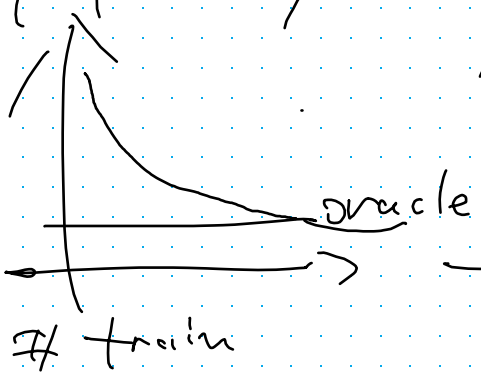
The more model capacity increase, the more bias increase, it leads to overfitting.

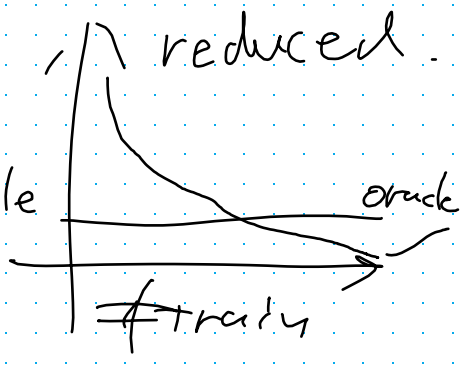
1(c) How does the size of trainset affect the bias - variance trade-off

For nonparametric model, the more #sample, the better generalization and resolve this trade-off

For parametric model, the more #truly sample increase, the more variance reduce and if we use model as asymptotically unbiased, the bias reduced.







optimal capacity

Capacity

Subject

Date 22

Attendants

DL Book 5.4

Bias

$\mathbb{E}$ はデータに対する期待値

$$bias(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta^*$$

$\mathbb{E}(\hat{\theta}_m)$：推定値の期待値　θ^*：真のθ

推定量 $\hat{\theta}_m$ は $bias(\hat{\theta}_m) = 0$ である場合に不偏 (unbiased)

と呼ばれ $\mathbb{E}(\hat{\theta}_m) = \theta^*$ であると意味する

$\lim_{m \to \infty} bias(\hat{\theta}_m) = 0$ のとき

asymptotically unbiased

サンプル分散（sample variance）

$$\hat{\sigma}_m^2 = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \hat{\mu}_m)^2$$

これは偏りのある式

Subject Date 23

Attendants

unbiased sample variance

$$\tilde{\sigma}_m^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x^{(i)} - \hat{\mu}_m)^2$$

$$\mathbb{E}[\tilde{\sigma}_m^2] = \sigma^2$$

Variance は推定量の分散はデータ集合内の事例数 m の関数として減少する

$$MSE = \mathbb{E}[(\hat{\theta}_m - \theta^*)^2] = Bias(\hat{\theta}_m)^2 + Var(\hat{\theta}_m)$$

$$P(|\hat{\theta}_m - \theta^*| > \varepsilon) \to 0$$ consistency

$$P(\lim_{m \to \infty} x^m = m) = 1$$ のとき概収束

almost sure convergence

Consistency によって，推定量の Bias はデータの事例が増えるほど減少する

Subject Date 24

Attendants

$m \uparrow$ bias $\downarrow$ cus $\lim_{m\to\infty} \hat{\theta}_m = \theta^*$

$\mathbb{E}[\hat{\theta}_m] = \theta^* \Leftrightarrow \hat{\theta}_m$ は不偏推定量 (unbiased)

$\lim_{m\to\infty} \mathbb{E}[\hat{\theta}_m] = \theta^* \Leftrightarrow \hat{\theta}_m$ は漸近不偏 (asymptotically unbiased)

分散はデータ数 m の関数として減少する

2(a) Diff of Hyperparam & Param.

⇒ which is trained by using training data.

Hyperparam affect dynamics of training as a training configuration and used for model selection with validation data.

Subject Date 2/7

Attendants

2(b) Two ML Model clearly distinguishing the two categories.

We assume x is input, y is prediction.

① MLP

$\hat{y} = f(x;\theta)$

this θ is model parameter

$L = loss(f, y)$

$\theta_{t+1} \leftarrow \theta_t - \eta \frac{\partial}{\partial \theta} L$

η is called learning rate

this is also hyperparameter

②

Linear Regression

$\hat{y} = W^T x + b$

$W \in \mathbb{R}^{N \times K}$ $x \in \mathbb{R}^{K \times 1}$, $b \in \mathbb{R}^N$

$\left(N \begin{bmatrix} \in \end{bmatrix} \times \begin{bmatrix} 1 \end{bmatrix} = N \begin{bmatrix} 1 \end{bmatrix} \right)$: is

W and b model param

$L = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{n} W_i^2$

this λ is regularization term as hyperparam.

Subject Date 26

Attendants

2(c) How can you choose hyperparam.

if statistical model is regular

we can use information criterion such as AIC (Akaikes Information Criterion)

if statistical model is singular such as deep neural networks.

choose hyperparam configuration where best validation loss is recorded.

At first, we assume to use hold-out strategy that divides train data into train, val, and test data

Subject Modify March 29, 2022 Date 27

Attendants Thanks to Joanis

3. Give two example for each of these problem.

- news text
  identify groups of articles that are similar, e.g. similar Topic.

unsupervised clustering problem

(a) k-means as ML model nonparametric

(b) hierarchical clustering method by using measure of single linkage method or ward method.

Subject Modify 29 March, 2022 Date 2f

Attendants thanks to Ioannis!

- Predict the likelihood of a certain having a given disease given the presence/absence of symptom and results of medical test

supervised classification problem

(a) Using logistic regression.

classification의 경우에는 Linear Regression은 confusing하다.

(b) Using NN model w/ softmax in the last layer.

Subject Modify March 29th 2022. Date 29

Attendants Thanks to Joanis.

height

① Predict whether a flower will have an ~~eigfat~~ higher than 20cm 30 days after it has been sown given characteristics/measurements of soil and environment it will grow in.

種をまく (sow) の下

生育する soil (土壌) や環境の特徴や測定値をもとに、種をまいてから 30日後に花の高さが 20cm以上になるか予測できる。

supervised (height がいくつになるか?)

classification problem (20cm 以上かそうじゃないか?)

(a) decision tree

(b) SVM

そうじゃないなら問題成立しない。

Subject Date 30

Attendants

Q4 (a) 最尤推定

モデルを固定したとき，事例集合が発現する probability が max となる param を求める

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, P(X;\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i;\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\, \log p(x_1;\theta)\, p(x_2;\theta) \cdots p(x_n;\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_i;\theta)$$

---

at first, let θ as model param

and x as input

Subject Date 3/

Attendants

MLE is a method to choose model θ where the likelihood based on model θ is maximized.

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(x|\theta)$$

(ii) Explain the concept of identically and independently distributed (i.i.d) random variables

1. identically means distributed from same distribution

e.g. $x_1 \sim p(x)$ $x_2 \sim p(x)$

2. independently distributed means

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) P(X_2 = x_2)$$

Subject Date 5/2

Attendants

確率変数が同分布とは

random variable $X_1 \cdots X_n$ が同分布とは

任意の実数 x に対し

$$P(X_1 = x) = P(X_2 = x) = \cdots P(X_n = x)$$

が成り立つことをいう

4(c) Let $\theta > 0$ $x_1 \cdots x_n \in \mathbb{R}$ are

drawn i.i.d.

from the dist w/ density function

$$f_\theta(x) = \frac{1}{2\theta} \exp\left(-\frac{|x|}{\theta}\right)$$

consider negative log ... ood

$$-\log f_\theta = -\log(2\theta) + \frac{|x|}{\theta}$$

Subject Date 22

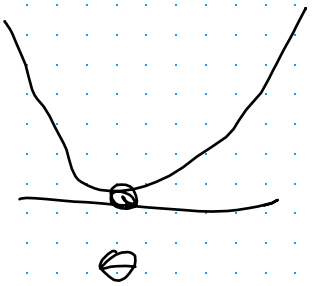
Attendants

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ f_\theta(x)$$

$$= \underset{\theta}{\operatorname{argmin}} - \log f_\theta(x)$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \frac{|x|}{\theta} - \log 2\theta \right\}$$

$$- \log f_\theta(x)$$



$$\frac{\partial - \log f_\theta(x)}{\partial \theta} = -|x|\theta^{-2} - \log 2$$

$$-(\log\theta)'$$

$$\Leftrightarrow \frac{\partial}{\partial\theta} NLL = -\frac{|x|}{\theta^2} - \frac{1}{\theta} - \log 2$$

$$\frac{\partial}{\partial\theta} NLL = 0 \Leftrightarrow$$

Subject Date 34

Attendants

$$f_\theta(x) = \frac{1}{2\theta} \exp\left(\frac{-|x|}{\theta}\right)$$

$$= f(x;\theta)$$

We first obtain the likelihood by multiplying the probability density function for each X_i:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{2\theta} \exp\left(\frac{-|x_i|}{\theta}\right)$$

$$= \left(\frac{1}{2\theta}\right)^n \exp\left(\frac{-\sum_{i=1}^{n}|x_i|}{\theta}\right)$$

Instead of directly maximizing the likelihood, we instead maximize the log likelihood

$$\log L(\theta) = -n\log(2\theta) - \frac{\sum_{i=1}^{n}|x_i|}{\theta}$$

Subject Date 3/4

Attendants

$$\log L(\theta) = -n \log(2\theta) - \frac{\sum_{i=1}^{n} |x_i|}{\theta}$$

To maximize this function, we take a derivative wrt θ

$$\frac{d}{d\theta} L(\theta) = \frac{d}{d\theta} \left(\log 2 + \log \theta \right) \times -n$$

$$\frac{\partial}{\partial \theta} \left(- \frac{\sum_{i=1}^{n} |x_i|}{\theta} \right)$$

$$= \frac{-n}{\theta} + \frac{\sum_{i=1}^{n} |x_i|}{\theta^2}$$

we set this derivative equal to zero then solve for θ

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} |x_i|$$

Subject Date 30

Attendants

(d) $x_1 \cdots x_n \in \mathbb{R}$ would have been identically distributed but not independent would that have changed your answer to the previous question? why?

Ans

Yes, it changed.

Cu2. we cannot use following equation

$$p(x_1, x_2, x_3 \cdots x_n) = \prod_{i=1}^{n} p(x_i)$$