EE2211 Lecture 7:

Overfitting, Model Complexity, Feature Selection / Regularization, Bias-Variance Tradeoff. (Math)

Reading: Lec_7.pdf

Review of Linear & Polynomial Regression

Goal: Given features $\underline{x} \in \mathbb{R}^d$, we want to predict $y \in \mathbb{R}$.

- $\underline{x}$: one-dim or d-dim $(d \geq 1)$
- y: one-dim

Input: Training data $\{(\underline{x}_i, y_i)\}_{i=1}^{m}$

Test set $\{\underline{x}_j\}_{j=m+1}^{n} = \{\underline{x}_{m+1}, \underline{x}_{m+2}, \ldots, \underline{x}_{m+n}\}$

In prev lectures, we called one test sample $\underline{x}_{new}$.

Learning/Training: Learn regression coefficients $\bar{w}^* = \begin{bmatrix} b^* \\ \underline{w}^* \end{bmatrix}$

Testing/Prediction: Prediction of $\hat{y}_{m+1}, \ldots, \hat{y}_{m+n}$ corresponding to $\{\underline{x}_{m+1}, \underline{x}_{m+2}, \ldots, \underline{x}_{m+n}\}$

Affine / 1-dim case $\quad d=1$

$m=4 \quad \{(x_1,y_1), (x_2,y_2), (x_3,y_3), (x_4,y_4)\}$: training set.

$$y_{new} = b^* + w^* x_{new}$$

Design matrix

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} \in \mathbb{R}^{m \times (d+1)} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \in \mathbb{R}^m$$

Training: $\bar{w}^* = (X^T X)^{-1} X^T y \in \mathbb{R}^{d+1}$

($X^T X$ has an inverse) $\nearrow$

Testing: Given x_{new}, $y_{new} = \begin{bmatrix} 1 \\ x_{new} \end{bmatrix}^T \bar{w}^*$

$1 \times 2 \qquad 2 \times 1$

Polynomial / 1-dim case ($d=1$)

$m=4$ trg samples

$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$: training set.

Quadratic relationship: $y = b + w_1 x + w_2 x^2$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix} \in \mathbb{R}^{m \times (d+2)} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \in \mathbb{R}^m$$

Training: $\bar{w}^* = (X^T X)^{-1} X^T y \in \mathbb{R}^{d+2}$

($X^T X$ has an inverse) $\nearrow$

Testing: Given x_{new}, $y_{new} = \begin{bmatrix} 1 \\ x_{new} \\ x_{new}^2 \end{bmatrix}^T \underline{w}^*$

1x3 3x1

Notation: Sometimes, when we do polynomial regression, the design matrix is P instead of X.

Training: $\bar{w}^* = (P^T P)^{-1} P^T y$

Testing: $$y_{new} = \underbrace{\begin{bmatrix} 1 \\ x_{new} \\ x_{new}^2 \end{bmatrix}^T}_{P_{new}^T} \bar{w}^*$$

<u>Note on Training & Test Sets</u>

Affine is a special case of polynomial

$\Rightarrow$ use P instead of X from now on.

Training: $\bar{w}^* = (P^T P)^{-1} P^T y \in \mathbb{R}^{d'}$

Test: $x_{mt1}, x_{mt2}, \cdots, x_{mtn}$ test samples.

$$\underline{y_{new}} = \underbrace{\underset{n \times d'}{P_{new}} \; \underset{d' \times 1}{\bar{w}^*}} \in \mathbb{R}^n$$

$\boxed{\text{If } P^T P \text{ does not have an inverse, then } P^T P + \lambda I \text{; this has an inverse}}$

$$P_{new} = \begin{bmatrix} p_{m+1}^T \\ \vdots \\ p_{m+n}^T \end{bmatrix} \in \mathbb{R}^{n \times d'}$$

Note on Training & Test Sets

There should be zero overlap between trg and test sets.

$$\downarrow$$
$$\overline{w}^*$$

Goal of regression: Do well / Predict well on new, unseen data.

Test set: Evaluate how good is our learning / trg procedure!

Bias-Variance Tradeoff

$$\underbrace{\text{Test Error}}_{\text{mean-squared error}} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error.}$$

Suppose $y = f(x) + \varepsilon$

f: deterministic

$f(x) = w_0 + w_1 x + w_2 x^2$

ε: random. mean 0 & var σ^2.

Repeat training 5 times.

Each time picked 10 trg samples

Each training set D is random. $p(D)$

Using each training set D, learn $\hat{f}_D(x)$

depends on D
will change as D changes.

In prev ex. average predictions over 5 trials (5 diff training sets)
Perform ∞ trials (∞ # of training sets)

$$\hat{f}_{avg}(x) = E_D[\hat{f}_D(x)]$$

expectation w.r.t. $p(D)$
averaging over ∞ trials.

Thm: Bias-Variance Decomposition Thm.

your prediction

Test Error $= Bias(\hat{f})^2 + Var(\hat{f}) + \sigma^2$

$$E[(y(x) - \hat{f}_D(x))^2] = (\hat{f}_{avg}(x) - f(x))^2 + E_D[(\hat{f}_D(x) - f_{avg}(x))^2] + \sigma^2$$

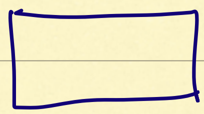
true target associated to x

$$(X^TX+\lambda I)^{-1}X^Ty = X^T(XX^T+\lambda I)^{-1}y$$

primal dual

Least norm solution: $X^T(XX^T)^{-1}y$

underdetermined



$d>m$

$1\ 2\ 1$

$R_2 - 5R_1$ ~~rank~~ $\text{rank}\left(\begin{pmatrix} 1 & 2 & 1 \\ 5 & 10 & 5 \\ 3 & 7 & c \end{pmatrix}\right)$ $\text{rank}\begin{pmatrix} 1 & 2 & 1 \\ 5 & 10 & 5 \end{pmatrix} = 1.$

can be at most 2.

rank = # of linearly independent rows.

last row is linearly indep of $\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$.

$\text{rank}(3\times 3) = 2.$

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 3 & 7 & c \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 1 \\ 3 & 7 & c \\ 0 & 0 & 0 \end{pmatrix}$$

no pivot

~10 T/F

~10 MCQ. (a), (b), (c) ... (g)

~ 3 FIB

$\downarrow$

~ 3 parts

$d > m \Rightarrow$ dual $\quad w = P^T \underbrace{(PP^T + \lambda I)^{-1}}_{\sim d^3} y$