CSE 480/680

1/22/2026

Lecture 5

HW2 posted
NOT GRADED

LII
Stat Refresher
MC2035
9:30, 10:30

Predictors

- K-Nearest-Neighbor
- Linear - for regression
  - for classification

Algorithms

LS Regression

Concepts

- Decision Region, Dec. Boundary
- Training error, Test error
  - Expected error
- Variance, Bias
- Loss functions - training / test / expected loss
- Max Likelihood

Lecture II: Linear regression and classification. Loss functions

Marina Meilă
mmp@uwaterloo.ca

With Thanks to Pascal Poupart & Gautam Kamath
Cheriton School of Computer Science
University of Waterloo

January 12, 2026

Linear predictors generalities ✓

Loss functions ✓

Least squares linear regression

- Linear regresssion as minimizing L_{LS}
- Linear regresssion as maximizing likelihood

Linear Discriminant Analysis (LDA)

QDA (Quadratic Discriminant Analysis)

Logistic Regression

The PERCEPTRON algorithm

SVMachines

Reading HTF Ch.: 2.1–5,2.9, 7.1–4 bias-variance tradeoff, Murphy Ch.: 1., 8.6[1], Bach Ch.:

[1] Neither textbook is close to these notes except in a few places; take them as alternative perspectives or related reading

LS Linear Regression

1. $e = y - X\beta$

2. Loss in matrix vector form

$$n \cdot L_2^{train}(\beta) = \|e\|_2^2 = (y - X\beta)^T (y - X\beta)$$
$$= y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta$$

(symmetric; $-y^T X\beta = a^T$)

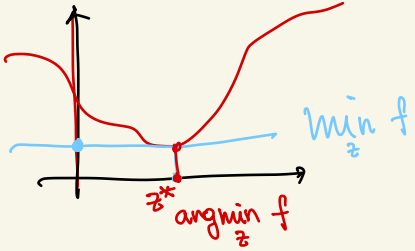
3. Find $\hat{\beta} = \arg\min_{\beta} L_2^{train}(\beta)$

by solving $\nabla L_2^{train}(\beta) = 0$

3.1.
$$\nabla L_2^{train}(\beta) = 0 - 2X^T y + 2X^T X\beta = 0$$

3.2 solve linear system
$$(X^T X)\beta = X^T y$$
A b

$A = d \times d$
$\beta \in \mathbb{R}^d$
$b = \mathbb{R}^d$



$\|e\|_2^2 = e^T e$

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

$g(z) = a^T z \in \mathbb{R}$

$\nabla g = a$

quadratic

$h(z) = z^T A z$

$\nabla h = 2Az$

A symmetric

$Az = b$

$\beta = A^{-1}b$

3.2 solve linear system

$$(X^TX)\beta = X^Ty$$

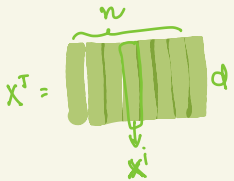
A $\quad$ b

$A = d \times d$

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

$b = \mathbb{R}^d$

$$\boxed{\hat{\beta} = (X^TX)^{-1}X^Ty}$$ LS Solution



$n > d$

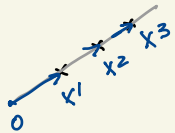
1. X^TX invertible? $\Leftrightarrow$ X^T full rank

$\Leftrightarrow$ X^T rank d $\Leftrightarrow$ $x^1, x^2, \dots x^n$ colums of X^T $\supset$ base of $\mathbb{R}^d$

Will assume X^T full rank d always

Ex: rank $X^T < d$

$d = 2$
$n = 3$

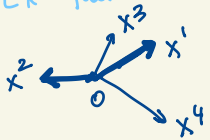


rank $[x^1\ x^2\ x^3] = 1 < d$

$x^2 = \alpha_2 \cdot x^1$

$x^3 = \alpha_3 x^1$

Ex full rank



2. $B = (X^TX)^{-1}X^T \Rightarrow \hat{\beta} = By$ linear in y !!

X

prediction $\hat{y}(x) = \beta^T x =$

$= y^TB^Tx = X^TBy$

tr. set y

new x $\quad$ X training set

B = pseudo inverse of X

$BX = I_d \quad d \times d$

$XB \neq I_n \quad n \times n$

Categorical attributes

$x_j \in \{\underbrace{red}_{a_1}, \underbrace{white}_{a_2}, \ldots\}$ $j = A$ categorical

$|A| = m$

1-hot encoding / "sparse" coding / indicator vars

$$x_j \in A \Rightarrow \begin{cases} x_{j1} = 1 & x_j = a_1 \\ x_{j2} = 1 & x_j = a_2 \\ \vdots & \\ x_{jm} = 1 & x_j = a_m \end{cases}$$ and 0 otherwise

x_{j1} ← $x_j = a_1$

x_{j2} ← $x_j = a_2$

x_{jm} ← $x_j = a_m$ redundant

Ex: why?

$$x_{j1} + x_{j2} + \ldots x_{jm} = 1$$

$$x_j = 1 - \sum_{q<m} x_{jq}$$

(Linear) least squares regression

- define **data matrix** or (transpose) **design matrix**

$$\mathbf{X} = \begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \cdots \\ (x^i)^T \\ \cdots \\ (x^n)^T \end{bmatrix} \in \mathbb{R}^{N \times n} \quad \text{and} \quad Y = \begin{bmatrix} y^1 \\ y^2 \\ \cdots \\ y^n \end{bmatrix}, \quad E = \begin{bmatrix} \varepsilon^1 \\ \varepsilon^2 \\ \cdots \\ \varepsilon^d \end{bmatrix} \in \mathbb{R}^d$$

- Then we can write

$$Y = X\beta + E$$

- The solution $\hat{\beta}$ is chosen to minimize the sum of the squared errors $\sum_{i=1}^{d}(\varepsilon^i)^2 = \sum_{i=1}^{d}(y^i - \beta^T x^i)^2 = ||E||^2$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{d}(y^i - \beta^T x_i)^2$$

- This optimization problem is called a **least squares** problem. Its solution is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \tag{8}$$

- Underlying statistical model $y = \beta^T x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$ (and i.i.d. sampling of $(x^{1:N}, y^{1:N})$ of course).
  Then $\hat{\beta}$ from (8) is the Maximum Likelihood (ML) estimator of the parameter β.

The statistical view of Machine Learning: Likelihood $= Pr[\text{data} \mid \text{model}]$

What is observed

my model (≠ true model)

In prediction

Model: $Pr[y \mid \underline{x}, \text{parameters}] = L(\text{params})$

y: random; $\underline{x}$: given

Max Likelihood Principle

Choose params $=$ argmax $L(\ )$

Linear regression

1. $y = \underline{\beta}^T \underline{x} + \varepsilon$ ↷ noise
2. $\varepsilon \sim N(0, \sigma^2)$ Gaussian
3. ε^i iid

Max L. Estimation

1. Write $L(\beta, \sigma^2)$

$n = 1$ data $= (x, y)$

$\varepsilon = y - \beta^T x \Rightarrow p(\varepsilon) = \frac{e^{-\varepsilon^2/2}}{\sigma\sqrt{2\pi}} =$

$p(y \mid \underline{x}, \beta, \sigma^2) = L(\beta, \sigma^2)$

The statistical view of Machine Learning: Likelihood

$n > 1$ $(x^{1:n}, y^{1:n})$ data $\Rightarrow \varepsilon^i = y^i - \beta^T x^i$

$$p(\varepsilon^{1:n} \mid x^{1:n}, \beta, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-(\varepsilon^i)^2/2\sigma^2} \right) = L(\beta, \sigma^2)$$

STATS

CALCULUS

2. Log-likelihood

$$\ell(\beta, \sigma^2) = \ln L(\beta, \sigma^2)$$

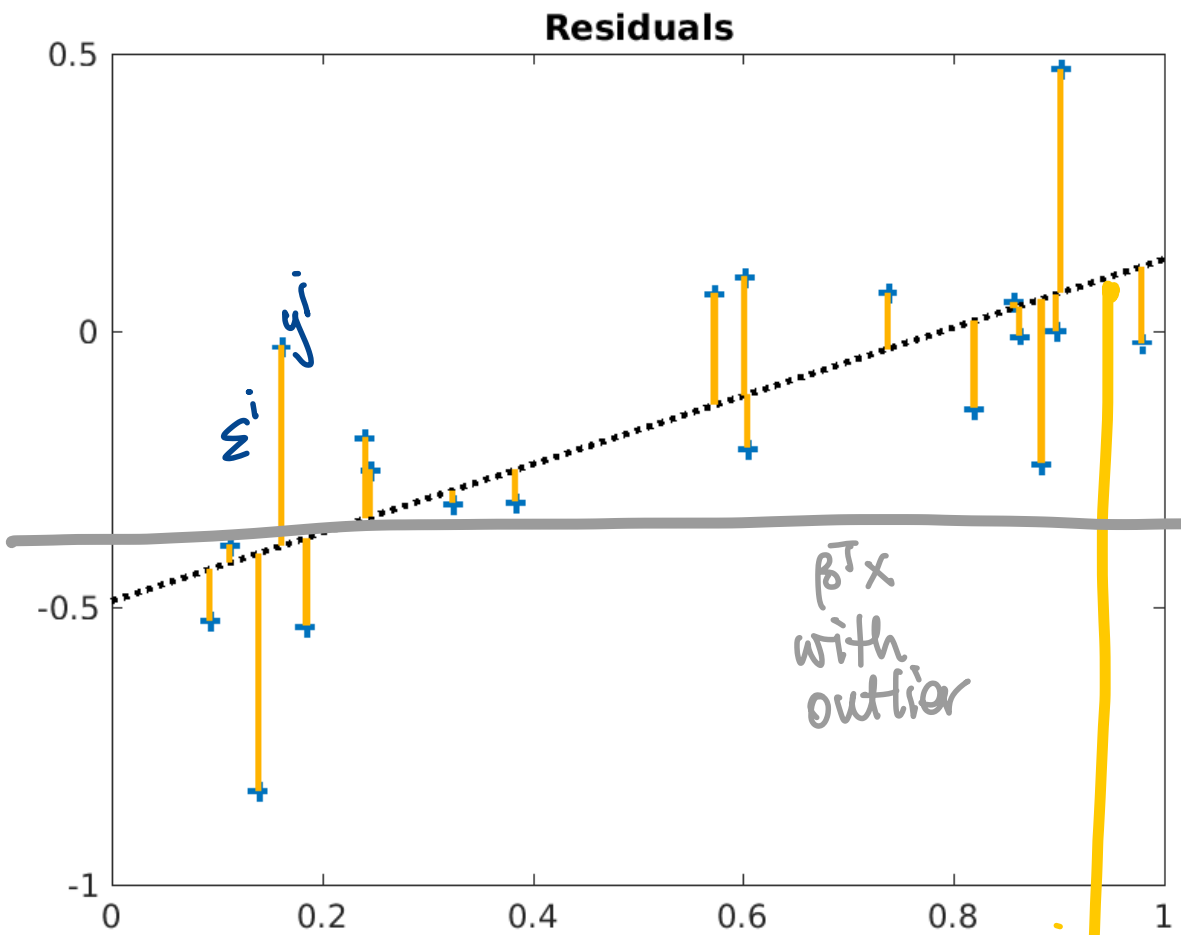
$$= \sum_{i=1}^{n} \left\{ -\ln \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2} (y^i - \beta^T x^i)^2 \right\}$$

3. Maximize ℓ over β, σ^2

$$\Leftrightarrow \left[\min_{\beta} \sum_{i=1}^{n} (y^i - \beta^T x^i)^2 \right] \equiv \text{LS estimation}$$

$$\max_{\sigma^2} \ell(\hat{\beta}, \sigma^2)$$

The statistical view of Machine Learning: Likelihood



1. Prediction
$\hat{y}(x) = \hat{\beta}^T x$
estimated from $\mathcal{D}$
↑ new x
Ex: ↑ confidence

2. $\varepsilon^{1,2,\ldots n}$ → estimate σ^2

3. Noise Gaussian?
Statistical test$(\varepsilon^{1:n})$

4. Outliers
$|\varepsilon^i|$ large

The statisticaal view of Machine Learning: Likelihood

- What is random? **the noise** $\epsilon^{1:n}$
- Express noise as function of $(x^{1:n}, y^{1:n})$

$$\epsilon^i = y^i - \beta_0 - \beta^T x^i \quad \sim N(0, \sigma^2) \tag{9}$$

- Likelihood

- Let $p_{0,\sigma^2}(\epsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\epsilon^2}{2\sigma^2}} = N(\epsilon; 0, \sigma^2)$
- Then

$$L(\beta_0, \beta_{1:d}, \sigma^2) = \prod_{i=1}^{n} p_{0,\sigma^2}(\epsilon^i) \tag{10}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\epsilon^i)^2}{2\sigma^2}} \tag{11}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y^i - \beta_0 - \beta^T x^i)^2}{2\sigma^2}} \tag{12}$$

- **log-likelihood**

$$l(\beta_0, \beta_{1:d}, \sigma^2) = \tag{13}$$

$$= \sum_{i=1}^{n} \left\{ -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) - \frac{1}{2} \left(y^i - \beta_0 - \beta^T x^i \right)^2 \frac{1}{2\sigma^2} \right\} \tag{14}$$

$$= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) + \text{constant} - \frac{1}{2} \sum_{i=1}^{n} \left(y^i - \beta_0 - \beta^T x^i \right)^2 \tag{15}$$

Maximizing the log-likelihood w.r.t β

- For simplicity, let $\beta_0 = 0$; hence $y^i = \beta^T x^i + \epsilon^i$
- log-likelihood

$$l(\beta_{1:d}, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^i - \beta_0 - \beta^T x^i \right)^2 + \text{constant} \tag{16}$$

- For any σ^2,

$$\operatorname*{argmax}_{\beta} l(\sigma^2, \beta) = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} \left(y^i - \beta_0 - \beta^T x^i \right)^2 \tag{17}$$

  a **Least Squares Problem**

- In matrix form $\min_\beta \|y - \mathbf{X}\beta\|^2$
- Solution

$$\beta^{\mathrm{ML}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{18}$$

  with $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \equiv \mathbf{X}^\dagger$ the pseudoinverse of $\mathbf{X}$
- β^{ML} is linear in y!