CSE 480/680

1/22/2026

Lecture 5

LS as Max Likelihood

HW 2 posted NOT GRADED

L II

stat Refresher

MC 2035

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

S. V. Machines

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 Regression Problem

1. error in matrix-vector form

$\mathcal{D} = \{(x^i, y^i),\ i = 1:n\}$

$f(x) = \beta^T x$

Want $\beta = ?$

Loss $= L_2 = (y - f(x))^2$ | $\Rightarrow$ want $\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ L_2^{train}$

$$L_2^{train} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{(y^i - \beta^T x^i)}_{e^i}{}^2 = \frac{1}{n}\|e\|_2^2 = \frac{1}{n}$$

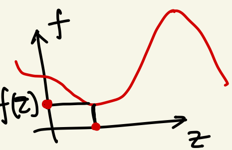
$$e = \begin{bmatrix} y^1 - \beta^T x^1 \\ y^2 - \beta^T x^2 \\ \cdots \end{bmatrix} = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \end{bmatrix} - \begin{bmatrix} (x^1)^T \beta \\ (x^2)^T \beta \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} y^1 \\ y^2 \\ \vdots \end{bmatrix}}_{y} - \underbrace{\begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \cdots \end{bmatrix}}_{X} \beta = y - X\beta$$

(d columns; $\beta \in \mathbb{R}^d$)

$\beta^T x = x^T \beta$

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

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



$f^* = \min_z f(z)$

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

Finding the optimal β

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

Loss in matrix-vector form

$$= y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta$$

$a^T z$ = $z^T A z$

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

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

$a^T z = g(z) = z^T a$

$\nabla g(z) = a$

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

$\nabla h = 2Az$

A symmetric

2. $y^T y$ → 0

$\min L_2^{train} \iff \nabla L_2^{train} = 0$

3. $$\nabla L_2^{train} = 0 - X^T y \cdot 2 + 2 X^T X \beta = 0$$

Find β by solving $\nabla L_2^{train} = 0$

4. L.S. Solution

$$X^T X \beta = X^T y$$

$X^T X$: $d \times d$

$A\beta = b$

If $(X^T X)^{-1}$ exists $\Rightarrow$

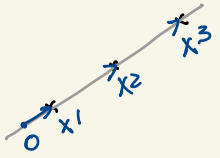
$$\hat{\beta} = (X^T X)^{-1} X^T y = A^{-1} b$$

$$= By$$

B = pseudo-inverse of X

$BX = I_d$ $d \times d$

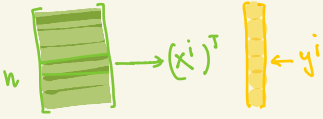
$XB \neq I_n$ $n \times n$



$\mathbb{R}^2$

$d = 2$

rank $X = 1 < d$



$n > d$

$X^T X$ nonsingular $\iff$

X^T has rank d $\iff$

columns linearly indep $\iff$

$\{x^1, x^2, \cdots\}$ —"—

We will assume it always

Suppose $\frac{1}{n}\sum_{i=1}^{n} x^i = 0$

$\frac{1}{n} X^T X = \Sigma$ covariance of the data

(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

$$Y = X\beta + E$$

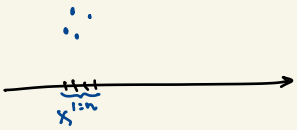
- The solution $\hat{\beta}$ is chosen to minimize the sum of the squared errors $\sum_{i=1}^{d}(\varepsilon^i)^2 = \sum_{i=1}^{d}(y^i - \beta^T x^i)^2 = ||E||^2$

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

- 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.



$\text{Cov}(X) = \sigma_x^2$ small

$\hat{\beta} = \frac{1}{\sigma_x^2} X^T y$

Exercise: . $d=1$ $\hat{\beta}=?$

- $x^i \leftarrow \alpha x^i$, $\alpha > 0$
- $y^i \leftarrow \alpha y^i$, $\alpha > 0$
- $y^i \leftarrow y^i + \alpha$

Categorical attributes

$x_j \in \{$ red, white, ... $\}$

$\in \{$ normal, no gas, battery dead, ... $\}$ ←

1-hot encoding

"sparse —"—"

$x_j \rightarrow \begin{cases} x_{j1} = 1 \text{ if } x_j = a_1 \text{ and } 0 \text{ otherwise} \\ x_{j2} = 1 \text{ if } x_j = a_2 \quad —"— \\ \vdots \\ x_{jm} \quad \ldots \end{cases}$

$x_j \in A$, $|A| = m$

↳ indicator variables

x_1 = time of day $\in \mathbb{R}$

$x_2 \in A$, $m = 3$

y = late to work $\in \mathbb{R}$

$x^1 = (7\text{am}, \text{normal}) \rightarrow (7, 1, 0, 0)$ (normal)

$x^2 = (10\text{pm}, \text{no gas}) \rightarrow (22, 0, 1, 0)$

...

$m-1$ variables sufficient

Transforming categorical inputs into real values

- if X_j takes two values (e.g "yes", "no"), map it to $\{\pm 1\}$ or $\{0, 1\} \subset \mathbb{R}$.
- discrete multivariate inputs
  - Let X_j take values in $\Omega_j = \{0, \ldots r-1\}$.
  - One defines the $r-1$ binary variables $\tilde{X}_{jk} = \mathbf{1}_{[]} X_j = k,\ k = 1 : r-1$. The variable X_j is replaced with $\tilde{X}_{jk},\ k = 1 : r-1$
  - the parameter β_j with $r-1$ parameters $\beta_{j1} \ldots \beta j, r-1$ representing the coefficients of $\tilde{X}_{j1}, \ldots \tilde{X}_{j,r-1}$.

  This substitution is widely used to parametrize any function of a discrete variable as a linear function

Example: The demographic variable Race takes values in { African, Asian, Caucasian, ...}; the corresponding parameters in the model will be $\hat{\beta}_{\text{Asian}}, \hat{\beta}_{\text{Caucasian}}, \ldots$.

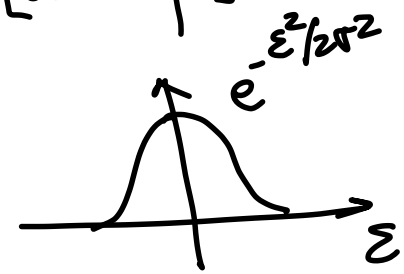
The statistical view of Machine Learning: Likelihood = Pr[data | model]

$D = \{(x^i, y^i),\ i = 1:n\}$

$L(\beta) = \Pr[\text{data} \mid \beta, \cdot]$

Our Model of the data source

$= \Pr[y \mid x, \text{parameters}]$



Gaussian

Linear predictor Model $y = \beta^T x + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

σ: noise "amplitude"

Write the likelihood

$n = 1 \quad (x, y)$

What is random? $\varepsilon = y - x^T\beta$

$y - x^T\beta$ → data and parameters

$p(\varepsilon) = e^{-\frac{\varepsilon^2}{2\sigma^2}} \cdot \frac{1}{\sigma\sqrt{2\pi}} = L(\beta, \sigma)$ likelihood

$n > 1 \quad \varepsilon^1, \varepsilon^2, \ldots \varepsilon^n$ iid $\Rightarrow L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\varepsilon^i)^2}{2\sigma^2}}$

STATISTICS

The statistical view of Machine Learning: Likelihood

CALCULUS

- $L(\beta,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y^i-(x^i)^T\beta)^2/2\sigma^2}$

- Log Likelihood

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

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

want $\operatorname{argmax}_{\beta,\sigma^2} \ell(\beta,\sigma^2)$

PRINCIPLE (Max likelihood)

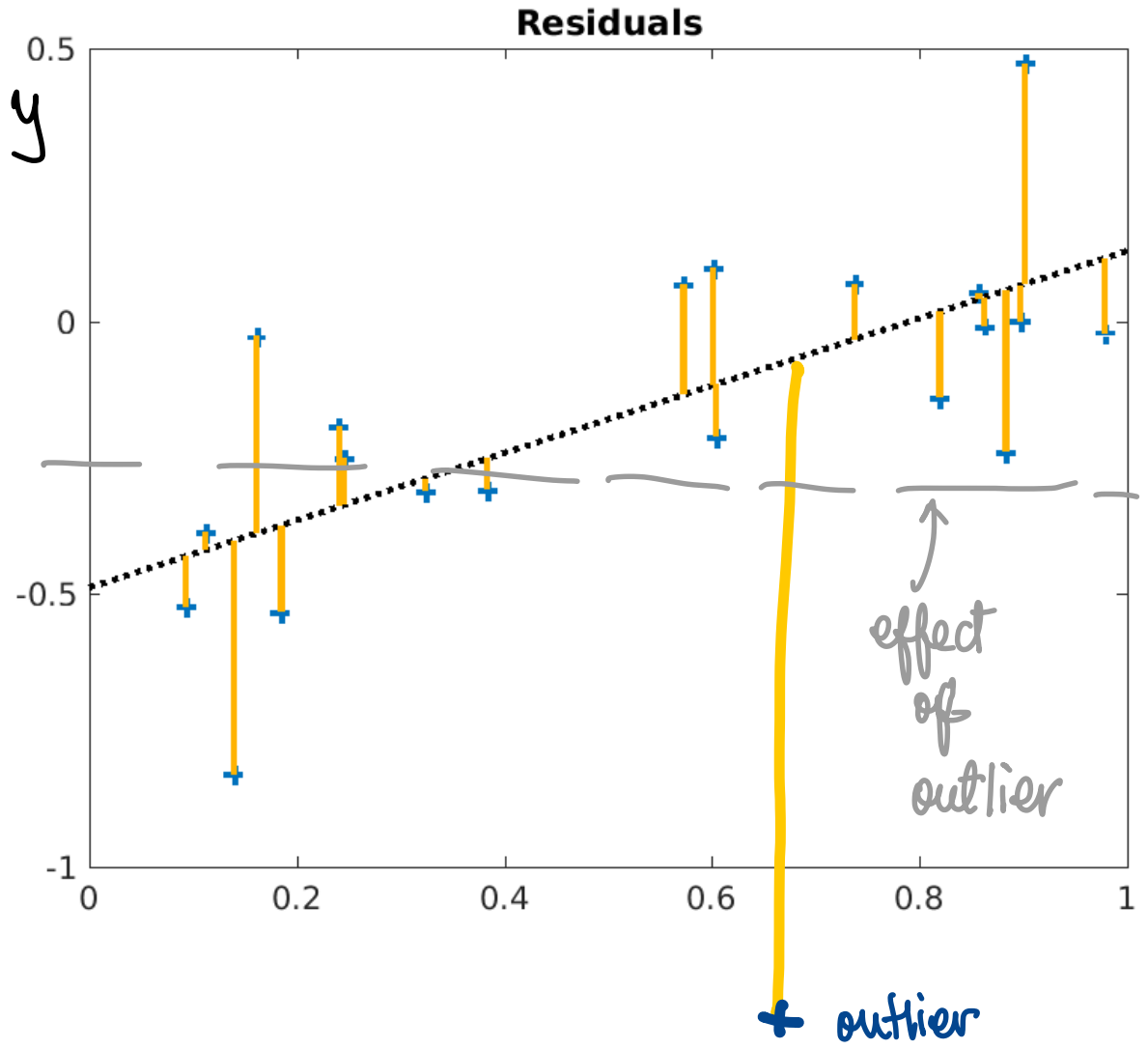
"best" parameters = argmax ℓ

$$\max_{\beta} \ell(\beta,\sigma^2) \Leftrightarrow \min_{\beta} \sum_{i=1}^{n}(y^i-\beta^T x^i)^2$$

train L_2

LSquares !!

The statisticaal view of Machine Learning: Likelihood



fit to data ≡
≡ train ≡ learn

Ex

1. $\sigma^2 \leftarrow \dots$
noise level

2. $\hat{\sigma}^2, \hat{\beta} \rightarrow \varepsilon^{1:n}$
Gaussian?
stat. test

3. outliers
ε^i large

4. outliers have large influence

LS not robust to outliers

The statistical 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**

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

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

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

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!