STAT 391

2/20/2025

Lecture 14

Linear Regression

- Sol HW 4
- HW 5 due +24 h
- LVII Regression + Notes + figures

fig-toy-linreg.png

Lecture Notes VII – Regression

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

~~May, 2023~~

Prediction problems

Linear regression

Linear regression for non-linear f

Reading: Ch.

Prediction

- **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \dots (x^n, y^n)\}$
- **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- **Outputs** $y^{1:n}$

- **Goal** Learn/estimate $f(x)$ **predictor** for y

- By type of output
  - **Classification** if $y \in S_Y$ discrete
  - $y \in \{0, 1\}$ (or $\{\pm 1\}$) binary classification
  - $y \in \{1, 2, \dots m\}$ multiclass classification

  - **Regression** if $y \in \mathbb{R}$ continuous

Part I Estimating a P on S $\quad X \sim P$ on S

+ Model Selection

Part II Prediction $\quad X \sim P_X$ on S_X

$Y \sim P_{Y|X}$ on S_Y ← dependent on X

Prediction problems by S_Y

- S_Y discrete — classification
- $S_Y \subset (-\infty, \infty)$ continuous — regression

Model of prediction

- $P_{Y|X=x}$ uncertainty in Y
  ↑ depends on input !!
- $X \sim P_X \quad S_X$ ← random X ← ignoring it / independently of it
  $Y \sim P_{Y|X} \quad S_Y$ ← random Y given X uncertain ← LEARN IT

$(x, y) \in S_X \times S_Y$

P_{XY} = joint

P_X = marginal of X

P_Y = — " — of Y

$P_{Y|X}$ = conditional distribution of Y given X

Prediction Problem

features
covariates
x = input(s)
attributes

- **Data** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$
- **Inputs** $x^{1:n} \in \mathbb{R}$, or $\mathbb{R}^d$
- **Outputs** $y^{1:n}$ output

TRAINING
(Learning)
(Estimation)

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = x \in \mathbb{R}^d$$

- **Goal** Learn/estimate $f(x)$ **predictor** for y

deterministic

$$x \xrightarrow{f(x)} y$$

HAS TO BE DETERMINISTIC

- By type of output
  - **Classification** if $y \in S_Y$ discrete
  - $y \in \{0, 1\}$ (or $\{\pm 1\}$) binary classification
  - $y \in \{1, 2, \ldots m\}$ multiclass classification
  - **Regression** if $y \in \mathbb{R}$ continuous

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

- **Model family** $\mathcal{F} = \{f_\theta : \mathbb{R}^d \to S_Y, \theta \in \Theta\}$

parametric

  - $\mathcal{F} = \{$ linear functions $\}$ linear regression/classification
  - $\mathcal{F} = \{$ polynomes of degree 2, 3, $\ldots \}$ polynomial regression/classification
  - $\mathcal{F} = \{\frac{1}{1+e^{-\beta^T x + \beta_0}}\}$ logistic regression

non-param.

  - neural network regression/classification
  - kernel regression/classification
  - $\mathcal{F} = \{$ monotonic functions $\}$ isotonic regression
  - support vector regression/classification
  - regression/classification *trees* (and random forests)

Linear regression

(see also last 3 pages)

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

Model

- $y^i = \beta_0 + \beta_1 x^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}$ (univariate regression)
- $y^i = \beta_0 + \beta_1 x^i_1 + \beta_2 x^i_2 + \ldots \beta_d x^i_d + \epsilon^i$ for $x^{1:n} \in \mathbb{R}^d$ (multivariate regression)

  $y \in \mathbb{R}$ is output/response

  $x^{1:n}_j$ for $j = 1:d$ are input(s)/covariates/features/attributes/...

  $\beta_{1:d}$ are **regression coefficients**, β_0 is intercept

  $\epsilon^{1:n} \in \mathbb{R}$ is noise, $\epsilon^{1:n} \sim N(0, \sigma^2)$ i.i.d.

Model

- $S_x = \mathbb{R}^d$ $\quad X \sim P_X$ on $\mathbb{R}^d$

  $S_y = \mathbb{R}$ $\quad Y \sim P_{Y|X}$ $\quad Y = f_\beta(X) + \varepsilon$

sample spaces

deterministic

$N(0,\sigma^2)$ noise

- independent of X *

** $f_b(x) = \beta_0 + \beta_1 x_1 + \cdots \beta_d x_d$

$$= \underbrace{[\beta_0 \; \beta_1 \cdots \beta_d]}_{\mathbb{R}^{d+1}} \cdot \underbrace{\begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}}_{d+1 \text{ dim.}}$$

Want: parameters $\underline{\beta_{0:d}, \sigma^2}$ ← Estimate by ML.

Linear regression

Model

- $y^i = \beta_0 + \beta_1 x^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}$ (univariate regression)
- $y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots \beta_d x_d^i + \epsilon^i$ for $x^{1:n} \in \mathbb{R}^d$ (multivariate regression)
  $y \in \mathbb{R}$ is output/response
  $x_j^{1:n}$ for $j = 1:d$ are input(s)/covariates/features/attributes/...
  $\beta_{1:d}$ are **regression coefficients**, β_0 is intercept
  $\epsilon^{1:n} \in \mathbb{R}$ is noise, $\epsilon^{1:n} \sim N(0, \sigma^2)$ i.i.d.

deterministic

$$Y = f_\beta(x) + \varepsilon \sim N(f_\beta(x), \sigma^2) \equiv N(\mu_x, \sigma^2)$$

$N(0,\sigma^2)$ noise

- independent of X

μ_x

deterministic fction of X

Model class $\mathcal{F} = \{ f_\beta,\ \beta_{0:d} \in \mathbb{R}^{d+1},\ \sigma^2 > 0 \}$

Linear functions of X

Solution by Maximum Likelihood

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

- **Likelihood** Probability $y \mid X$, parameters
- Parameters $\beta_0, \beta_{1:d}, \sigma^2$
- $\beta_0^{ML}, \beta_{1:d}^{ML}, (\sigma^2)^{ML} = \underset{\beta_0, \beta_{1:d}, \sigma^2}{\text{argmax}}\, l(y|X, \beta_0, \beta_{1:d}, \sigma^2)$

Max log Likelihood $(\underbrace{y^{1:n} \mid x^{1:n}}_{\text{want } P_{Y|X} \text{ not also } P_X}, \beta_0, \beta_{1:d}, \sigma^2)$

Before (Ch1): $\ell(\text{data} \mid \text{params})$

1. Likelihood

$y^i \sim N(f_\beta(x^i), \sigma^2)$

$$L(y^{1:n} | x^{1:n}, \beta_{0:d}, \sigma^2) = \prod_{i=1}^{n} e^{-\frac{(y^i - f_\beta(x_i))^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}}$$

μ^i

1'. log-L

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

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

STAT

CALCULUS

2.

$$\underset{(\beta,\sigma^2)}{\ell} = -\sum_{i=1}^{n} \overbrace{(y^i - \beta^T x^i)^2}^{(r^i)^2} \cdot \frac{1}{2\sigma^2} - \frac{1}{2}\cdot n \ln(2\pi\sigma^2)$$

max β,σ^2

Variables σ^2

$\beta_0, \beta_1, \dots \beta_d$

$$f_\beta(x^i) = \beta_0 \cdot \underset{x_0^i}{1} + \beta_1 x_1^i + \dots \beta_d x_d^i$$

$$\frac{\partial}{\partial \beta_j} f_\beta(x^i) = x_j^i$$

$$\frac{\partial \ell}{\partial \beta_j} = -2 \cdot \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^i - \beta^T x^i) \cdot (-x_j^i) = 0$$

$j = 0, 1, \dots d$

$$\left\{ \sum_{i=1}^{n} y^i x_j^i = \beta^T \sum_{i=1}^{n} (x^i) x_j^i \right. \quad \text{for all } j$$

Linear system

d+1 unknowns

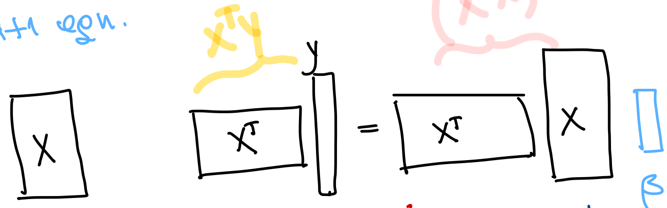
d+1 eqn.

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

$$X = \begin{bmatrix} x_0^i \cdots\cdots x_d^i \\ \cdots\cdots \end{bmatrix}$$

n rows x (d+1) columns

$$y = \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix}$$

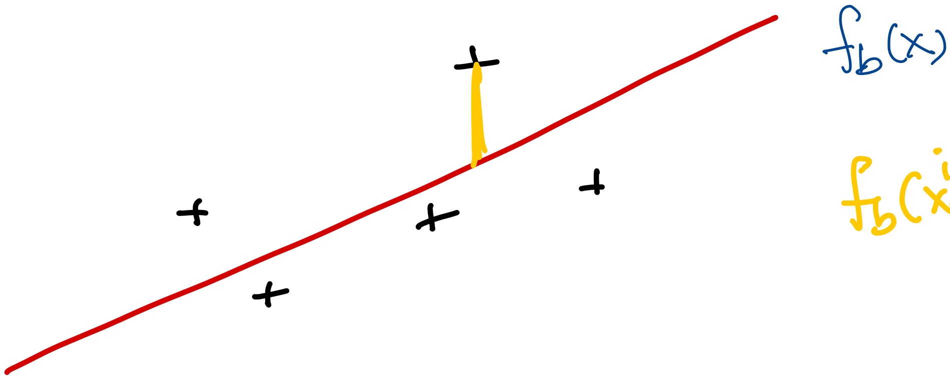


LEAST SQ SOLUTION

solve $\Rightarrow$ $\beta^{ML} = \underbrace{(X^T X)^{-1} X^T}_{X^+} y \in \mathbb{R}^{d+1}$

$$X^+ X = (X^T X)^{-1} (X^T X) = I$$

pseudo-inverse of X



$f_b(x^i) - y^i$ = residual of x^i

$$\min_{\beta} \sum_i (r^i)^2$$

More detailed derivation of Regression formula

$$\sum_{i=1}^{n} (y^i - \beta^T x^i)\cdot(-x^i_j) = 0 \qquad j = 0:d$$

$$\sum_{i=1}^{n} \beta^T x^i \, x^i_j = \sum_i \boxed{y^i x^i_j}$$

$0 \quad 1 \quad \cdots \quad d \qquad x^i$

$n \left[\quad X \quad \right] \leftarrow \boxed{y^i}$

$y^T X \in \mathbb{R}^{d+1}$ row vector

$$\beta^T \sum \left((x^i)^T \, (0 \quad 1 \quad \cdots \quad d) \right)$$

$$\beta^T \left(\sum (x^i)^T \, x^i \right) \rightarrow X^T X$$

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

$(d+1)\times(d+1)$

Linear regression model in matrix form

For a single data point (x^i, y^i)

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_d \end{bmatrix} \in \mathbb{R}^d \qquad x^i = \begin{bmatrix} x_1^i \\ x_2^i \\ \dots \\ x_d^i \end{bmatrix} \in \mathbb{R}^d \tag{1}$$

Then,

$$y^i = \beta_0 + (x^i)^T \beta + \epsilon^i \tag{2}$$

For all $\mathcal{D}$

$$y = \begin{bmatrix} y^1 \\ y^2 \\ \dots \\ y^n \end{bmatrix} \in \mathbb{R}^n \qquad \mathsf{X} = \begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \dots \\ (x^n)^T \end{bmatrix} \in \mathbb{R}^{n \times d} \qquad \epsilon = \begin{bmatrix} \epsilon^1 \\ \epsilon^2 \\ \dots \\ \epsilon^n \end{bmatrix} \tag{3}$$

$$y = \beta_0 1 + \mathsf{X}\beta + \epsilon \qquad \text{with } Cov(\epsilon) = \sigma^2 I_d \tag{4}$$

The (log)-likelihood

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

$$\epsilon^i \;=\; y^i - \beta_0 - \beta^T x^i \quad \sim N(0, \sigma^2) \tag{5}$$

- 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

$$L(\beta_0, \beta_{1:d}, \sigma^2) \;=\; \prod_{i=1}^{n} p_{0,\sigma^2}(\epsilon^i) \tag{6}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\epsilon^i)^2}{2\sigma^2}} \tag{7}$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y^i - \beta_0 - \beta^T x^i)^2}{2\sigma^2}} \tag{8}$$

- **log-likelihood**

$$l(\beta_0, \beta_{1:d}, \sigma^2) \;=\; \tag{9}$$

$$= \sum_{i=1}^{n} \left\{ -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) - \frac{1}{2} \left(y^i - \beta_0 - \beta^T x^i \right)^2 \frac{1}{2\sigma^2} \right\} \tag{10}$$

$$= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \ln(2\pi) + \text{constant} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^i - \beta_0 - \beta^T x^i \right)^2 \tag{11}$$

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

$$l(\beta_{1:d}, \sigma^2) \;=\; -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^i - \beta_0 - \beta^T x^i\right)^2 + \text{constant} \tag{12}$$

- For any σ^2,

$$\operatorname*{argmax}_{\beta} l(\sigma^2, \beta) \;=\; \operatorname*{argmin}_{\beta} \sum_{i=1}^{n}\left(y^i - \beta_0 - \beta^T x^i\right)^2 \tag{13}$$

  a **Least Squares Problem**

vector calculus

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

$$\boxed{\beta^{ML} \;=\; (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T y} \tag{14}$$

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