STAT 391

2/25/2025

Lecture 15

Linear Regression
- Double Descent

Q2 Thu 2/27

Lecture Notes – Regression

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

May, 2023

Prediction problems ✓

Linear regression ← Examples σ^2

(Linear regression for non-linear f)

Is model ≈ true?

Reading: Ch.

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.

x = covariates regressors attributes features inputs

Data $(x^1, y^1), \cdots (x^n, y^n)$

Model $y = \beta_0 + \beta^T x + \varepsilon$ $\sim$ noise $\sim N(0, \sigma^2)$

$= [\beta_0 \ \beta_1 \cdots \beta_d] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} + \varepsilon$

y = output dependent variable

Wanted params: $\beta_{0:d}^{ML}, \sigma^2$

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 $\mathcal{D}$

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

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

Estimation of σ^2

- Let $\hat{\epsilon}^i = y^i - (x^i)^T \beta^{ML}$, for $i = 1 : n$
- $\hat{\epsilon}^i$ called the residuals

- If we plug in β^{ML} in the log-likelihood equation (12) we obtain

$$l(\sigma^2, \beta^{ML}) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\hat{\epsilon}^i)^2 + \text{constant} \tag{21}$$

- Maximizing this expression w.r.t. σ^2 gives

$$(\sigma^2)^{ML} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\epsilon}^i)^2 \tag{22}$$

- The predictor is $f(x) = x^T \beta^{ML}$
- Hence, for a new $x \in \mathbb{R}^d$, our guess of y is $\hat{y} = f(x)$

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**

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

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

with $(X^T X)^{-1} X^T \equiv X^\dagger$ the pseudoinverse of X

- β^{ML} is linear in y!

$\max_{\sigma^2} -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}[C]$ — Calculus

$$\frac{\partial \ell}{\partial(\sigma^2)} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2}\frac{+1}{(\sigma^2)^2}C = 0 \Rightarrow \frac{C}{\sigma^2} = n \Rightarrow (\sigma^2)^{ML} = \frac{1}{n}\sum_{i=1}^{n}\left(y^i - (x^i)^T\beta^{ML}\right)^2$$

residual $\hat{\varepsilon}^i$

$f(x^i)$ = prediction of y^i

ε^i = estimate of noise ε^i

$= \frac{1}{n}\sum_{i=1}^{n}(\hat{\varepsilon}^i)^2 \simeq$ Variance !!

avg. of $(\text{errors})^2$

"LEAST SQUARES"

Least Sum of Squared Errors

Ex: $s = \sum_{i=1}^{n}\hat{\varepsilon}^i \cdot \frac{1}{n} = 0$

Best predictor (linear) : $\hat{}$ argmin $\sum \text{err}^2$

Model $\varepsilon^i \sim N(0, \sigma^2)$

ML Estimation (Training) ↑

Prediction (Testing) ↓

$\downarrow \beta^{ML}, (\sigma^2)^{ML}$

Predictor $f(x) = (\beta^{ML})^T x$ — linear predictor

new x → $\hat{y}$ prediction of y given x

Statistical properties of β^{ML}

- Assume the true model is $y^i = \beta^T x^i + \epsilon^i$ with $\epsilon \sim N(0, \sigma^2)$.
- Here β, σ^2 are true parameters and we assume we know them.
-
- β^{ML} is a random variable. Let us calculate its mean and standard deviation.
- **Expectation** of β^{ML}

$$
\begin{aligned}
E[\beta^{ML}] &= E[X^\dagger y] = E[X^\dagger (X\beta + \epsilon)] && (15)\\
&= E[X^\dagger X]\beta + E[X^\dagger \epsilon)] && (16)\\
&= \underbrace{X^\dagger X}_{I_d}\beta + X^\dagger \underbrace{E[\epsilon)]}_{0} = \beta && (17)
\end{aligned}
$$

Hence, $E[\beta^{ML}] = \beta$ and we say that β^{ML} is **unbiased**

- **Covariance** of β^{ML}
- By a similar (but longer) calculation, we obtain that

$$Cov(\beta^{ML}) = \sigma^2 (XX^T)^{-1} \in \mathbb{R}^{d \times d} \quad (18)$$

- In fact, β^{ML} has a Normal distribution. Remember from (15)

$$\beta^{ML} = \beta + X^\dagger \epsilon. \quad (19)$$

This is a linear transformation of ϵ, a Gaussian variable. Therefore,

$$\beta^{ML} \sim N(\beta, \sigma^2 (XX^T)^{-1}). \quad (20)$$

- This is a **multivariate Normal** distribution over $\mathbb{R}^d$,

Cov β^{ML}

$$\Sigma_\beta = \begin{bmatrix} \mathrm{Var}\,\beta_0^{ML} & \cdots & \mathrm{Cov}(\beta_i^{ML}, \beta_j^{ML}) \\ & \ddots & \\ & & \mathrm{Var}\,\beta_d^{ML} \end{bmatrix} \qquad i,j = 0:d$$

$$\Sigma_\beta = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{n} \cdot \frac{1}{\hat{\sigma}_x^2}$$

σ^2: noise; $\hat{\sigma}_x^2$: input data "spread"

$$X = \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix} \Rightarrow X^T X = \sum_{i=1}^{n} (x^i)^2 = n \cdot \mathrm{Var}\, x^i$$

sample variance $\hat{\sigma}_x^2$

$d=1$, $\beta_0 = 0$ $\Rightarrow$ $y = \beta x + \varepsilon$

$x, \beta \in (-\infty, \infty)$

Assume $\sum x^i = 0 \Leftrightarrow \frac{1}{n} \sum x^i = 0$

$$\sigma_\beta^2 = \frac{1}{n} \frac{\sigma^2}{\hat{\sigma}_x^2}$$

σ^2: $\propto$ noise; $\hat{\sigma}_x^2$: $\propto \frac{1}{x \text{ data "spread"}}$

$\to 0$ with n

$$\frac{\hat{\sigma}_x^2}{\sigma_{noise}^2} = \mathrm{SNR} = \text{Signal-to-Noise Ratio}$$

SNR ↑ $\Rightarrow$ σ_β^2 ↘

When model $y = \beta_0 + \beta_1 x_1 + \dots \beta_d x_d + \varepsilon$ NOT true

① $\hat{\beta}_{ML}$ Estimation

~in reg how used ?

② Test it →

Interpret coefficients $\beta_0, \beta_1, \dots$ (Hypothesis testing, Confi Intervals)

"Interpolation"

inference ⇒ $\beta_j \simeq 0$: x_j not important for y

$\beta_j > 0$, $\beta_j < 0$: x_j has significant effect on y

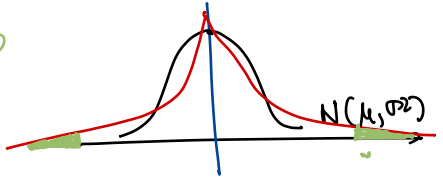
diagnosis

1. I Model ≈ true?
2. approx true but outlier
   - wrong data (not from P data source)
   - from P when P has heavy tails

HAVE LARGE ~~EFFECT~~ ON $\hat{\beta}_{ML}$!! INFLUENCE

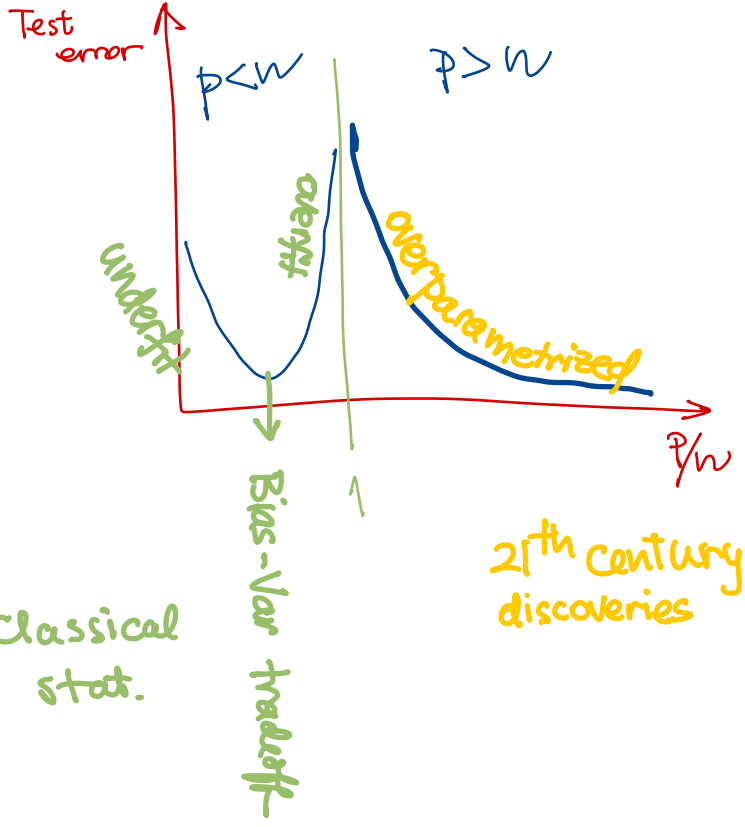
statistically $P(y^{out} | X) \approx 0$ by your model

$P(\text{tail}) \not\approx 0$



Bias-Variance Revisited

$\frac{1}{n^{test}} \sum (\hat{\varepsilon}_i)^2$ on test data



$n = 300$ $n^{test} = 500$

$p < n$

$p > n$

$y = \alpha^T x + \varepsilon$ $x \in \mathbb{R}^{100}$

True

↓ Neural net

$\tilde{x} \in \mathbb{R}^p$

$p/n \in (0, 5)$

Model $y = \beta^T \tilde{x} + \tilde{\varepsilon}$