Lecture Notes I – 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** $\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$
- **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
  - $y \in \{1, 2, \ldots m\}$ multiclass classification

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

Prediction

- **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** $y^{1:n}$

- **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
  - $y \in \{1, 2, \ldots m\}$ multiclass classification

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

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

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.

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 $\mathcal{D}$

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

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

Solution by Maximum Likelihood

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

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**

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

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

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

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}

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

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

$$
Cov(\beta^{ML}) \;=\; \sigma^2(\mathsf{X}\mathsf{X}^T)^{-1} \;\in\; \mathbb{R}^{d\times d} \tag{18}
$$

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

$$
\beta^{ML} \;=\; \beta + X^\dagger \epsilon. \tag{19}
$$

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

$$
\beta^{ML} \;\sim\; N(\beta, \sigma^2(\mathsf{X}\mathsf{X}^T). \tag{20}
$$

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

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 $\hat{y} = f(x)$