Lecture November 1

assumptions :

$$y = f(x) + \varepsilon$$

$$\varepsilon \sim N(\underbrace{0}_{\text{mean of } \varepsilon}, \sigma^2)$$

variance
$\sigma^2 = 1$

ideally there is a continuous function $f(x)$ which descri-bes our output/target y.

we make a model for

$$f(x) \simeq \tilde{y} = X\beta$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; C(\beta)$$

X is assumed to not be a stochastic matrix

$$\mathbb{E}[y_i] = \frac{1}{n}\sum_{i=0}^{n-1} y_i$$

$$y_i = \underbrace{\sum_{j=0}^{p-1} x_{ij}\beta_j}_{X_{i*}\beta} + \varepsilon_i$$

$$X_{i*}\beta \quad + \varepsilon_i$$

$$\mathbb{E}[y_i] = \mathbb{E}[x_{i*}\beta] + \mathbb{E}[\varepsilon_i]$$

$$= \underset{\substack{\simeq \\ f(x_i)}}{x_{i*}\beta} + 0$$

$$\mathbb{E}[(y_i - \mu_{y_i})^2] = \operatorname{var}[y_i]$$

$$= \frac{1}{n}\sum_{i=0}^{n-1} (y_i - \mu_{y_i})^2$$

$$\mathbb{E}[y_i]$$

$$\Rightarrow \operatorname{var}[y_i] = \sigma^2$$

Since $\varepsilon \sim N(0, \sigma^2) \Rightarrow$

$$y \sim N(X\beta, \sigma^2)$$

$$y_i \sim N(x_{i*}\beta, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2}{2\sigma^2}\right]$$

Maximum likelihood

y_i are independent and identically distributed

$$P(y, x \mid \beta) = \prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_{i*}\beta)^2}{2\sigma^2}\right)$$

$$= \prod_{i=0}^{n-1} p(y_i, x \mid \beta)$$

We want to maximize the probability.

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} P(y, x \mid \beta)$$

$$-\log p(y, x \mid \beta)$$

$$= -\sum_{i=0}^{n-1} \log p(y_i, x \mid \beta)$$

$$= \left(\frac{n}{2} \log(2\pi\sigma^2) + \frac{\|(y - X\beta)\|_2^2}{2\sigma^2} \right)$$

$$\frac{\partial}{\partial \beta}\left(-\log p(y x \mid \beta) \right) = 0$$

$$X^T(y - X\beta) = 0$$

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

Maximum likelihood Estimate (MLE)

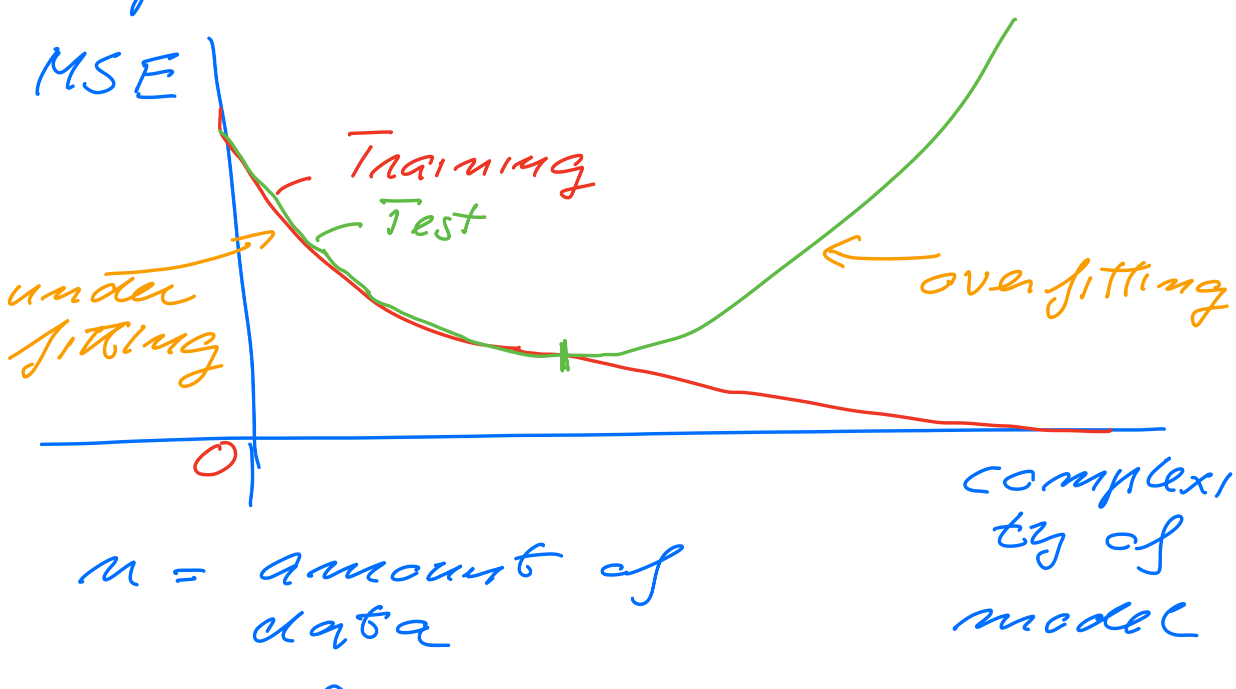
— Resampling Methods —

- Best possible estimate of MSE, R2, accuracy etc

$$MSE\text{-}OCS = \mathbb{E}\left[(y_i - \tilde{y}_i)^2\right]$$

$$= \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2$$

MSE as a function of model complexity, in order to select the best possible model



n = amount of data

$$D = \{(x_0, y_0), \ldots (x_{n-1}, y_{n-1})\}$$

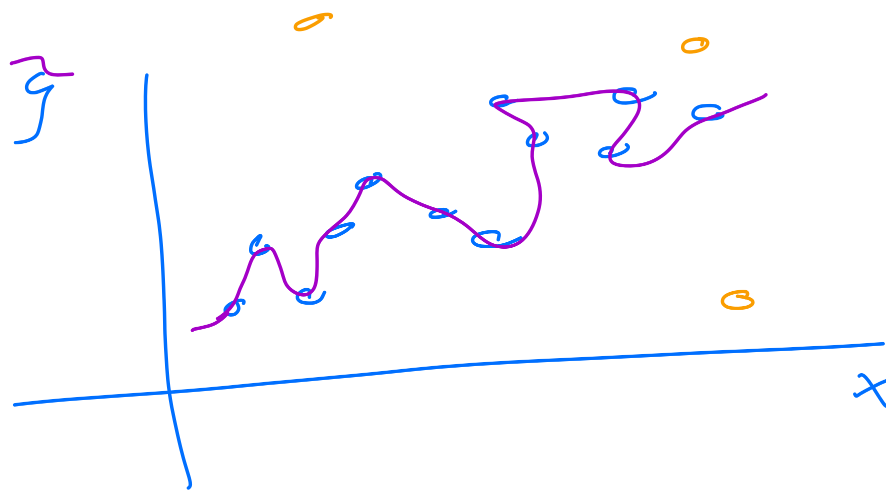
$$\tilde{y} = X\beta \quad \text{(our model)}$$

example

$$\tilde{y}_i = \sum_{j=0}^{p-1} \beta_j x_i^j$$

The polynomial order is the complexity.

simple data



How do we know if we are overfitting?

1) Need Reliable (statistically reliable estimate) estimate of our score

(MSE, R2, ...)

MSE as example:

$$\mathbb{E}[(y-\tilde{y})^2] = \frac{1}{n}\|(y-\tilde{y})\|_2^2$$

Our problem is that we do not know a priori the distribution of β (parameters) and the Domain D.

- statistical resampling
  - Bootstrap (few data)
  - Cross-validation
  - Jackknife
  - Blocking

($\mu_x = \int p(x)dx \cdot x$

$$\sum_{i \in D} p(x_i) x_i$$

sample mean

$$\bar{\mu}_x = \frac{1}{n}\sum x_i \neq \mu_x$$

MSE

test

2) These results depend on the amount of data.

3) Signal for overfitting is an increasing score as function of complexity

$\Rightarrow$ Bias-variance tradeoff.