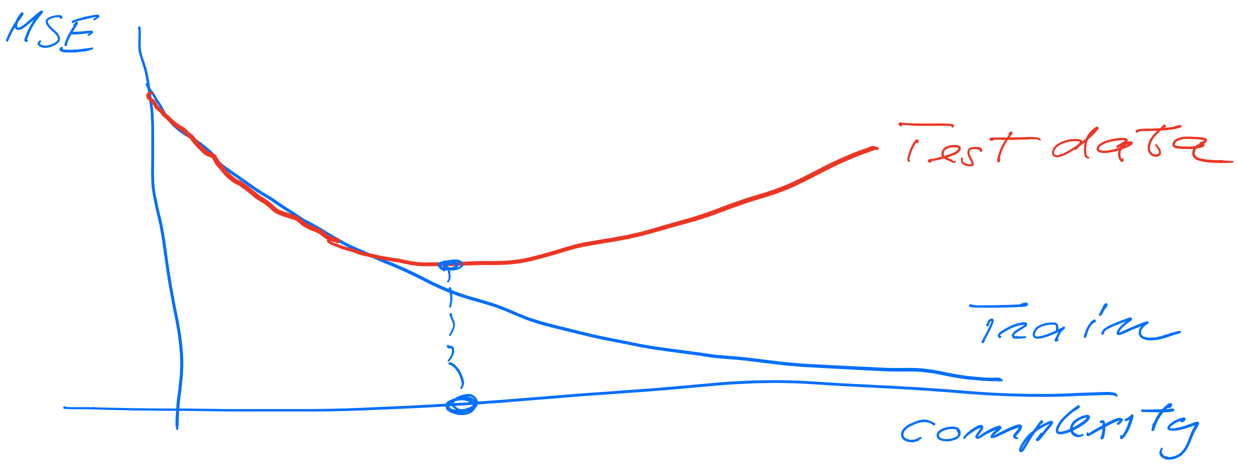
Friday August 28

Review of statistics & Probability Theory



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

$$= \mathbb{E}[(y - \tilde{y})^2]$$

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

$$\|x\|_2 = \sum_i \sqrt{x_i^2}$$

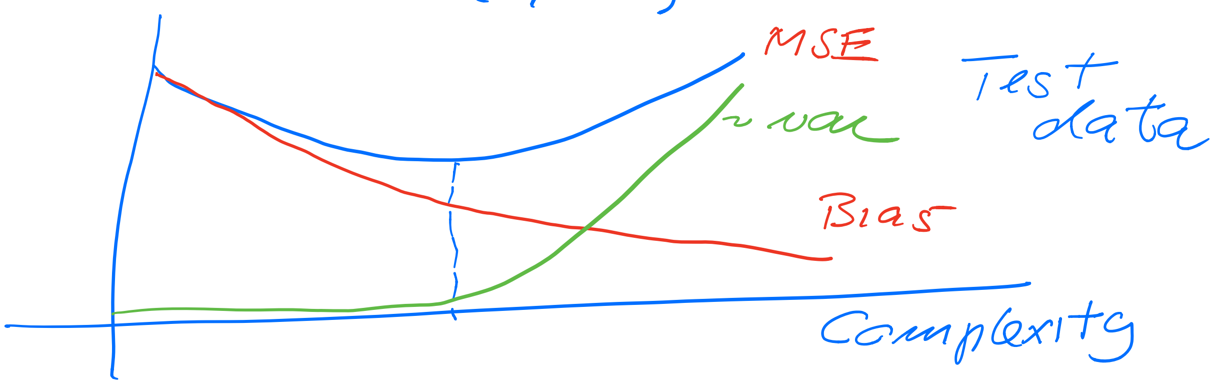
$$\|x\|_2^2 = \sum_i x_i^2$$

Bias-Variance trade-off

$$MSE = \underbrace{\mathbb{E}[(y - E[\tilde{y}])^2]}_{\text{Bias}} + var(\tilde{y}) + \sigma^2$$

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

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



Trade-off between complexity and the # of data.

Domain $D = \{(x_0, y_0), (x_1, y_1) \dots (x_{n-1}, y_{n-1})\}$

$x \subset X \quad y \subset Y$

assume we have a PDF

$p(x)$ and $p(y)$

$0 \leq p(x) \leq 1 \qquad x \in [a, b]$

$$\int_a^b p(x)\,dx = 1$$

Discrete case

$$\sum_{x_i \in X} p(x_i) = 1$$

Expectation values, moments

$$\mathbb{E}[x^n] = \int^b x^n p(x)\,dx$$

$$a$$

$$\sum_{x_i \in X} x_i^{n} \, p(x_i)$$

$$\mu_x = \underline{\mathbb{E}[x]} = \langle x \rangle = \mu$$

variance

$$\mathrm{var}(x) = \sigma_x^2 = \sigma^2 = \int_a^b (x-\mu_x)^2 \, p(x) \, dx$$

$$\left(\sum_{x_i \in X} (x_i - \mu_x)^2 \, p(x_i) \right)$$

sample mean

$$\bar{\mu}_x = \frac{1}{n} \sum_{i=0}^{n-1} x_i \neq \mu_x$$

sample variance

$$\bar{\sigma}_x^2 = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{\mu}_x)^2$$

Covariance

$$\mathrm{cov}(x,y) = \iint dx\,dy \; p(x,y) \times (x-\mu_x)(y-\mu_y)$$

$$= \underline{\iint dx\,dy \; p(x,y) \, xy} - \mu_x \mu_y$$

$$= \mathbb{E}[xy] - \mu_x \mu_y$$

i.i.d = independent and identically distributed

$$P(x\,y) = \underline{p}(x)\,\underline{p}(y)$$

same distributions

$$\mathbb{E}[xy] = \underbrace{\int dx\, p(x)\, x}_{\mu} \underbrace{\int dy\, p(y)\, y}_{\mu}$$

$$\mu_x \cdot \mu_y = \mu^2$$

$$\underline{\mathrm{cov}(x,y) = 0}$$

$$\underline{\iint dx\, dy\; x \cdot y\; P(x\,y)}$$

$$\mu_x = \int dx\, p(x)\, x = \mu_y = \mu$$

$$P(x\,y) = \underline{p(x)}\,p(y)$$

if i.i.d. then $\mathrm{cov}(x,y) = 0$

Covariance matrix (2 features)

$$C[x,y] = \begin{bmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) \end{bmatrix}$$

$$\mathrm{cov}(x,x) = \mathrm{var}(x)$$

$$C[x,y] = \begin{bmatrix} \mathrm{var}(x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \mathrm{var}(y) \end{bmatrix}$$

Introduce a scaling function

correlation function

$$K[x,y] = \begin{bmatrix} \dfrac{\mathrm{cov}(x,x)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(x)}} & \dfrac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} \\ \dfrac{\mathrm{cov}(y,x)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} & \dfrac{\mathrm{cov}(y,y)}{\sqrt{\mathrm{var}(y)\,\mathrm{var}(y)}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$$

take values between -1 and 1.

More Probability ...

X or Y

$$P(X \vee Y) = P(x) + P(y) - \underline{P(X \wedge Y)}$$

Joint probability

$$P(X, Y) = P(X \cap Y) = P(X|Y)\,P(Y)$$
$$= P(Y|X)\,P(X)$$

Marginal (sum rule) distribution

$$P(x) = \sum_{y_i \in Y} P(x \mid Y = y_i)\, P(Y = y_i)$$

Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)} \qquad P(y) > 0$$

posterior, likelihood, prior

$$= \frac{P(y|x)\,P(x)}{\sum_{x_i \in X} P(y \mid X = x_i)\, P(X = x_i)}$$

Baye's theorem

Example: mammogram

sensitivity: $X = 1$ positive event

$x = 0$ negative

$y = 1$ Breast cancer

$y = 0$ no breast cancer

$P(x = 1 | y = 1)$ $0.8 = 80\%$

$P(x = 1 \mid y = 1) = 0.8$ 80%

$P(y=1 \mid x=1) = ?$

$P(y=1) = 0.004$

$P(y=0) = 1 - P(y=1)$

$P(x=1 \mid y=0) = 0.1$

Baye's theorem

$P(y=1 \mid x=1) =$

$$\frac{P(x=1 \mid y=1)\,P(y=1)}{P(x=1 \mid y=1)\,P(y=1) + P(x=1 \mid y=0)\,P(y=0)}$$

$= 0.031 \Rightarrow 3\%$

Maximum Likelihood Estimator

$P(y \mid x) = ?$

Model $P(Y \mid X, \beta) \longrightarrow$

$P(D \mid \beta) \longrightarrow$ iid (x_i, y_i)

$n-1$

$$P(D|\beta) \propto \underline{\prod_{i=0} P(y_i, x_i | \beta)}$$

β which maximizes $P(D|\beta)$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max}\ P(D|\beta)$$

$$= \underset{\beta \in \mathbb{R}^p}{\arg\max} \sum_{i=0}^{n-1} \log P(y_i x_i | \beta)$$

instead of maximizing, we can minimize

$$C(\beta) = -\sum_{i=0}^{n-1} \log \underline{P(y_i x_i | \beta)}$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\ C(\beta)$$

$$P(y_i x_i | \beta) = ??$$

$$\boxed{y_i = \underline{f(x_i)} + \underline{\varepsilon_i}} \qquad \boxed{X_{i*}\beta}$$

$\in \mathbb{R}^{n \times p}$ $\quad$ $\beta \in \mathbb{R}^p$

$$\varepsilon_i \sim N(0, \underline{\sigma^2})$$

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

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

$$= x_{i*}\beta + 0$$

$$= x_{i*}\beta$$

$$\mathrm{var}[y_i] = \mathbb{E}\left[\left(y_i - \mathbb{E}[y_i]\right)^2\right]$$

$$= \mathbb{E}[y_i^2] - (x_{i*}\beta)^2$$

$$= \mathbb{E}\Big[\underline{(x_{i*}\beta)^2} + 2\underbrace{\varepsilon_i}_{=0} x_{i*}\beta + \underline{\underline{\varepsilon_i^2}}\Big] - \underline{(x_{i*}\beta)^2}$$

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

Assumption is that

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

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_i - x_{i*}\beta)^2}{2\sigma^2}\right]$$

$$= p(y_i x_i \mid \beta)$$

$$C(\beta) = \frac{n}{2}\log(2\pi\sigma^2)$$

$$+ \sum_{i=0} \frac{\left(\left(y_i - x_{i*}\beta\right)\right)}{2\sigma^2}$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0 \Rightarrow$$

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

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

$$y_i = X_{i*}\beta + \varepsilon_i$$

$$\begin{array}{cc} & \\ i \rightarrow & \\ & \end{array}
\begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0p-1} \\ x_{10} & & & \\ \vdots & & & \\ x_{n-1\,0} & - & \cdots & \end{bmatrix}
\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$