3 Estimation Theory

- **Markov Inequality**

If $X \geq 0$, random variable

$$Pr(X \geq \alpha \mathbb{E}[X]) \leq \frac{1}{\alpha}$$

- **Chebyshev Inequality**

$$Pr(|X - \mathbb{E}[X]| > \alpha\sigma) \leq \frac{1}{\alpha^2} \quad \text{if} \quad Var(X) = \sigma^2$$

- **Moment Generating function**

<u>def</u> $M_X(t) = \mathbb{E}[e^{tX}]$

$$= \int e^{tx} f_X(x)\,dx$$

<u>Prop</u> $\left.\frac{\partial}{\partial t} M_X(t)\right|_{t=0} = \mathbb{E}[X]$

- **Conditional Expectation**

<u>def</u> $\mathbb{E}[X|Y=y]$ = expectation of X given $Y=y$

$*$ **Tower property** $\mathbb{E}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}_X[X]$

• Maximum Likelihood Estimate (MLE)

def $\hat{Y} = \underset{y}{\operatorname{argmax}}\, P_{X|Y}(x|y)$

when y maximizes the likelihood which y makes X the most probable

ex) $f_{X|Y}(160|A)$ vs $f_{X|Y}(160|B)$

$$\frac{1}{\sqrt{2\pi \cdot 10^2}} e^{-\frac{(160-170)^2}{2 \cdot 10^2}} \quad \text{vs} \quad \frac{1}{\sqrt{2\pi \cdot 15^2}} e^{-\frac{(160-180)^2}{2 \cdot 15^2}}$$

$>$ $\Rightarrow$ A is probable

$<$ $\Rightarrow$ B is probable

• Maximum A Poster (MAP)

def $\hat{Y}_{MAP} = \underset{y}{\operatorname{argmax}} \; P_{Y|X}(y|x)$

$$P_{Y|X}(y|x) = \frac{P_{X|Y}(x|y)\, P_Y(y)}{P_X(x)}$$

$$P_X(x) = \sum P_{X|Y}(x|y)\, P_Y(y)$$

ex> $Pr(Y=A) = \frac{2}{3}$ $\quad Pr(Y=B) = \frac{1}{3}$

$$P_{Y|X}(A|160) \quad vs \quad P_{Y|X}(B|170)$$

$$\frac{f_{X|Y}(160|A)\, Pr(Y=A)}{f_X(160)} \quad vs \quad \frac{f_{X|Y}(160|B)\, Pr(Y=B)}{f_X(160)}$$

$$\frac{1}{\sqrt{2\pi \cdot 10^2}} e^{-\frac{(160-170)^2}{2 \cdot 10^2}} \underbrace{\times \frac{2}{3} \; vs \; \frac{1}{\sqrt{2\pi \cdot 15^2}} e^{-\frac{(160-180)^2}{2 \cdot 15^2}} \times \frac{1}{3}}_{\text{weight}}$$

pick random : A is more probable

* MAP is Bayes Optimal

= minimize error

- **Fano's Inequality**

X: input Y: true label $\in \mathcal{Y} = \{1, 2, \cdots, k\}$

estimator $\hat{Y}(X)$

$$\Pr(Y \neq \hat{Y}(X)) \geq \frac{H(Y|X) - 1}{\log k}$$

- **Parameter Estimation**

$X_1, X_2, \cdots, X_n \sim \text{iid } P_\theta$

$$P_{X^n}(x^n) = \prod_{i=1}^{n} P_\theta(x_i)$$

$$= \prod_{i=1}^{n} P_X(x_i|\theta)$$

• Naïve Bayes

X^n : input Y : label

$$\hat{Y}_{MLE} = \underset{y}{\operatorname{argmax}}\ P_{X^n|Y}(x^n|y)$$

Namely, all features $x_1, \dots, x_n$ are independent

$$\underline{\text{def}}\quad P_{X^n|Y}(x^n|y) = \prod_{i=1}^{n} P_{X_i|Y}(x_i|y)$$

• Gaussian descriminant → Σ

$$X|_A \sim N(\mu_0, \Sigma_0),\quad X|_B \sim N(\mu_1, \Sigma_1)$$

MLE:

$$\frac{1}{\sqrt{(2\pi)^n|\Sigma_0|}}\exp\left(-\frac{1}{2}(x^n-\mu_0)^T\Sigma_0^{-1}(x^n-\mu_0)\right)$$

vs

$$\frac{1}{\sqrt{(2\pi)^n|\Sigma_1|}}\exp\left(-\frac{1}{2}(x^n-\mu_1)^T\Sigma_1^{-1}(x^n-\mu_1)\right)$$

MAP: $\frac{1}{\sqrt{(2\pi)^n |\Sigma_0|}} \exp\left(-\frac{1}{2}(x^n-\mu_0)^T \Sigma_0^{-1} (x^n-\mu_0)\right) \times P_Y(0)$

vs

$\frac{1}{\sqrt{(2\pi)^n |\Sigma_1|}} \exp\left(-\frac{1}{2}(x^n-\mu_1)^T \Sigma_1^{-1} (x^n-\mu_1)\right) \times P_Y(1)$

$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ & & \\ Cov(X_n, X_1) & & Cov(X_n, X_n) \end{bmatrix}$$

$$Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

- **positive definite matrix** (p.d)

<u>def</u> X is p.d if $v^T X v \geq 0 \quad \forall v \in \mathbb{R}^n$

$v^T X v > 0 \quad \forall v_{\neq 0} \in \mathbb{R}^n$

(all eigenvalues are positive)

- **Positive semi-definite matrix (psd)**

<u>def</u> X is p.s.d if $v^T X v \geq 0 \quad \forall v \in \mathbb{R}^n$

(all eigenvalues are nonnegative)

ex) covariance matrix

- **Minimize MSE**

$\mathbb{E}[(X - \hat{X}(Y))^2]$: want to minimize

$\rightarrow$ minimum achiever $\Rightarrow \mathbb{E}[X|Y]$ as $\hat{X}(Y)$)

• Bias & Variance

$$MSE = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

$$= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta)^2]$$

$$= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta)^2]$$

$$= Var(\hat{\theta}_n) + Bias(\hat{\theta}_n)^2$$

$$Bias[\hat{\theta}_n] = \mathbb{E}(\hat{\theta}_n - \theta)$$