Classification task

Pattern $\longrightarrow$ $\boxed{\text{Alg}}$ $\longrightarrow$ label

Notation :

① $\mathcal{X}$: Feature space = Set of all feature vectors
For e.g. $\mathcal{X} = \mathbb{R}^d$

② Classifier $h : \mathcal{X} \to \{1, ---, M\}$

Simple Case : $M=2$, $\{0,1\}$ or $\{-1,+1\}$

How to design classifiers & how to judge their performance?

Input : $\{ (x_i, y_i), i=1 --- n \}$ Training dataset

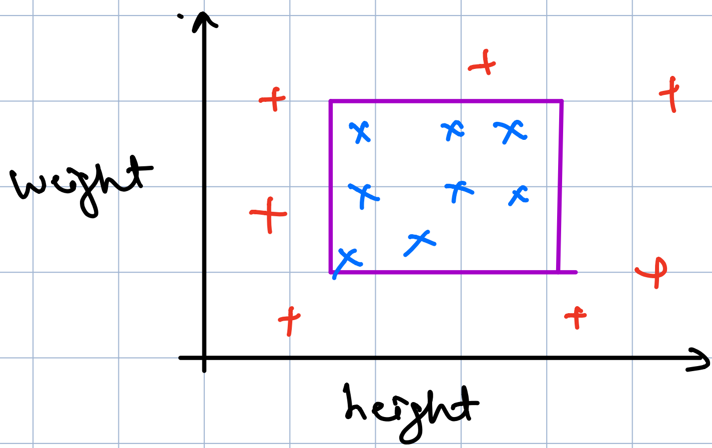
feature vector class label

Using training data, learn an appropriate classifier h

Test : Test & validate the h on "new" data

PTO

Example: Medium-build



Regression task:

Training data: $\{ (x_i, y_i), i = 1 \ldots n \}$

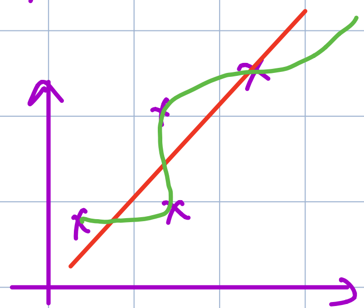
$x_i \in \mathbb{R}^d, \; y_i \in \mathbb{R}$

Need is to learn a function $h: \mathcal{X} \to \mathbb{R}$

Examples: prediction of stock prices, etc

Curve-fitting: $\{ (x_1, y_1), \ldots\ldots (x_n, y_n) \}$

Find a function f s.t. $y = f(x)$



Generalization:- Obtain a classifier/regression function using a training dataset s.t. the error on "unseen" (test) data is low.

Assumption: There is a distribution underlying the data.

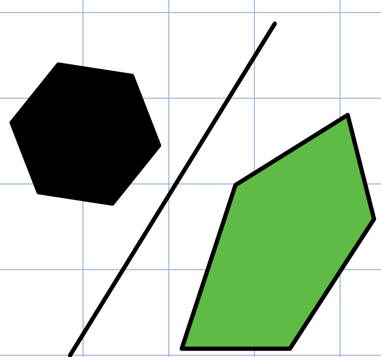
Formally,

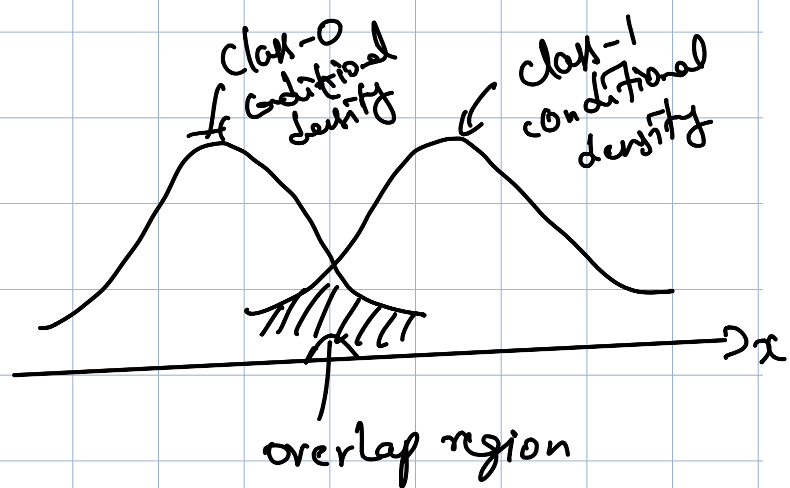
f_i : conditional density of features from class - i

Let $x = (x_1, \ldots x_d) \in \mathbb{R}^d$ represent a feature vector

$f_i(x)$: joint density of $(x_1, \ldots x_d)$ given that x is in class - i

Note: A feature vector x can belong to different classes with different probabilities





PTO

$\mathcal{X}$: feature space $\qquad \mathcal{Y} = \{0, 1\}$ class-labels

$\mathcal{H}$ = set of classifiers = $\{ h \mid h : \mathcal{X} \to \mathcal{Y} \}$

Performance metric : $F(h) = P(\underset{\substack{\nearrow \\ \text{label assigned} \\ \text{by classifier-}h}}{h(x)} \neq \underset{\substack{\uparrow \\ \text{true class} \\ \text{label}}}{y(x)})$

$\downarrow$

Probability of mis-classification

Goal: $h^* = \underset{h \in \mathcal{H}}{\arg\min}\ F(h)$

Bayes classifier

Let $\quad P_i = P(y(X) = i) \leftarrow$ prior probabilities

$q_i(x) = P(y(X) = i \mid X = x)$

Bayes classifier $\quad h_B(x) = \begin{cases} 0 & \text{if } \dfrac{q_0(x)}{q_1(x)} > 1 \\ 1 & \text{else} \end{cases}$

PTO

From Bayes theorem,

posterior $\propto$ prior $\times$ likelihood

$$q_i(x) = \frac{p_i f_i(x)}{p_0 f_0(x) + p_1 f_1(x)}$$

Normalizing constant

Bayes classifier:

Classify x as belonging to class "0"

if $q_0(x) > q_1(x)$

$\iff$ $p_0 f_0(x) > p_1 f_1(x)$

Optimality of Bayes classifier:

Fix a classifier h

Define $R_i(h) = \{ x \in \mathcal{X} \mid h(x) = i \}$, $i = 0, 1$

$F(h) = P(h(X) \neq y(X))$

$= P(X \in R_1(h), X \in \text{Class-}0)$

$$+ P(X \in R_0(h),\ X \in \text{Class-1})$$

$$= p_0\, P(X \in R_1(h) \mid X \in \text{Class-0}) + p_1\, P(X \in R_0(h) \mid X \in \text{Class-1})$$

$$= p_0 \int_{R_1(h)} f_0(x)\, dx + p_1 \int_{R_0(h)} f_1(x)\, dx$$

For the Bayes classifier h_B:

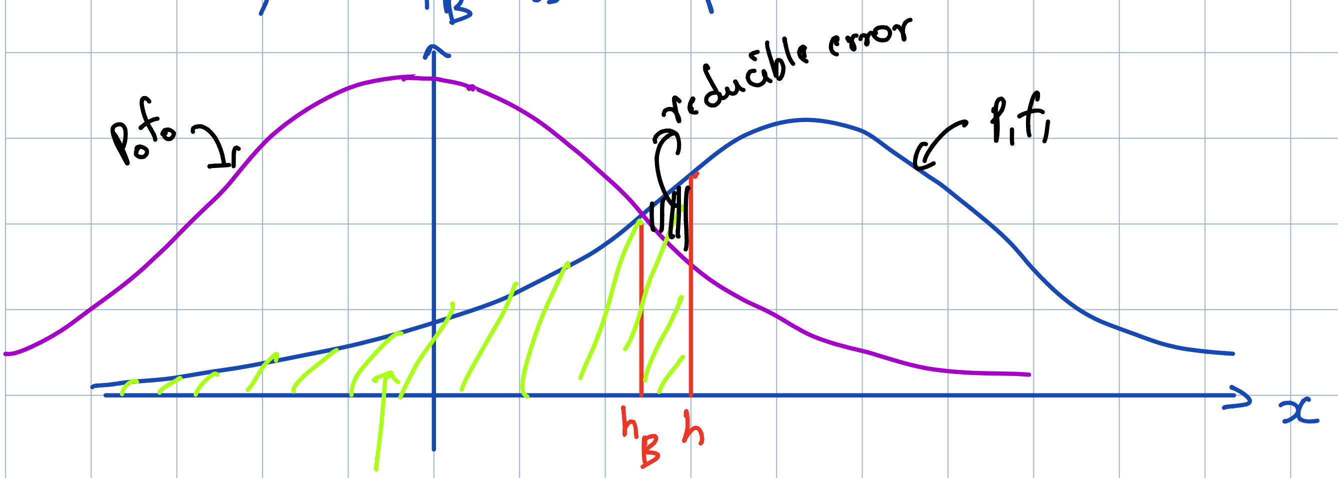
$$R_0(h_B) = \{ x \mid p_0 f_0(x) \geq p_1 f_1(x) \}$$

$$R_1(h_B) = \{ x \mid p_1 f_1(x) > p_0 f_0(x) \}$$

So, $$F(h_B) = \int_{R_1(h)} p_0 f_0(x)\, dx + \int_{R_0(h)} p_1 f_1(x)\, dx$$

$$= \int_{\mathcal{X}} \min\left(p_0 f_0(x),\ p_1 f_1(x)\right) dx$$

So, h_B is optimal. ■



$$\int_{R_0(h)} p_1 f_1(x)\,dx$$

Generalizing Bayes classifier to handle loss functions:-

Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$

$L(h(x), y(x))$ is the loss suffered by h on pattern x.

Performance metric $F(h) = E[L(h(x), y(x))]$ → (*)

Special case:- (0-1) loss function → $L(a,b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}$

Then, $F(h) = P(h(x) \neq y(x))$ for above L.

In general, $L(0,1) \neq L(1,0)$ $[L(0,0) = L(1,1) = 0]$

h_B for optimizing (*) is

$$h_B(x) = \begin{cases} 0 & \text{if } \dfrac{q_0(x)}{q_1(x)} > \dfrac{L(0,1)}{L(1,0)} \\ 1 & \text{else} \end{cases}$$

Note: If $L(1,0) = L(0,1)$, we recover the Bayes classifier for 0-1 loss function

Extending Bayes classifier to multi-class classification problems:-

Let $\{0, 1, --- M-1\}$ be the M-class labels.

$L(i,j) \rightarrow$ loss function

classifier predicts (i) true class label (j)

$$R(h) = E(L(h(x), y(x))) \leftarrow (**)$$

Goal: minimize $R(h)$

Deriving the Bayes classifier:

Let $R(i \mid x) = E(L(h(x), y(x)) \mid h(x)=i, x)$

$$= E(L(i, y(x)) \mid -\text{"}-)$$

$$= \sum_{j=0}^{M-1} L(i,j)\, P(y(x)=j \mid x)$$

$$= \sum_{j=0}^{M-1} L(i,j)\, q_j(x)$$

$$R(h) = E[E(L(h(x), y(x)) \mid x]]$$

$$= \int R(h(x) \mid x)\, f(x)\, dx$$

need to minimize this $\forall x \in \mathcal{X}$.

<u>Bayes Classifier:</u>

$$h_B(x) = i \quad \text{if} \quad \sum_{j=0}^{M-1} L(i,j)\, q_j(x) \leq \sum_{j=0}^{M-1} L(k,j)\, q_j(x), \quad \forall k$$

Claim: h_B is optimal for $(**)$

Special Case for $M=2$:

$h_B(x) = 0$ if

$$L(0,0)\, q_0(x) + L(0,1)\, q_1(x) \leq L(1,0)\, q_0(x) + L(1,1)\, q_1(x) \quad \text{—— } (***)$$

If $L(0,0) = L(1,1) = 0$, then $(***)$ is equivalent to

$$\frac{q_0(x)}{q_1(x)} \geq \frac{L(0,1)}{L(1,0)}$$

<u>H.W.:</u> For 0-1 loss function, with a M-class classification problem, derive h_B.

Special Case : $X \in \mathbb{R}$

$$f_i \sim N(\mu_i, \sigma_i^2), \quad i = 0, 1$$

$$f_i(x) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(- \frac{(x-\mu_i)^2}{2\sigma_i^2} \right), \quad i = 0, 1$$

$$h_B(x) = 0 \quad \text{if}$$

$$p_0 f_0(x) L(1,0) > p_1 f_1(x) L(0,1)$$

$$\log(p_0 L(1,0)) + \log(f_0(x)) > \log(p_1 L(0,1)) + \log(f_1(x))$$

(***)

$$\log(p_0 L(1,0)) - \log(\sigma_0) - \frac{1}{2}\log(2\pi) - \frac{(x-\mu_0)^2}{2\sigma_0^2}$$

$$> \log(p_1 L(0,1)) - \log(\sigma_1) - \frac{1}{2}\log(2\pi) - \frac{(x-\mu_1)^2}{2\sigma_1^2}$$

$$\frac{1}{2} x^2 \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) + x \left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2} \right)$$

$$+ \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2} \right) + \log\left(\frac{\sigma_1}{\sigma_0}\right) + \log\left(\frac{p_0 L(1,0)}{p_1 L(0,1)} \right)$$

$$> 0$$

Special Cases:

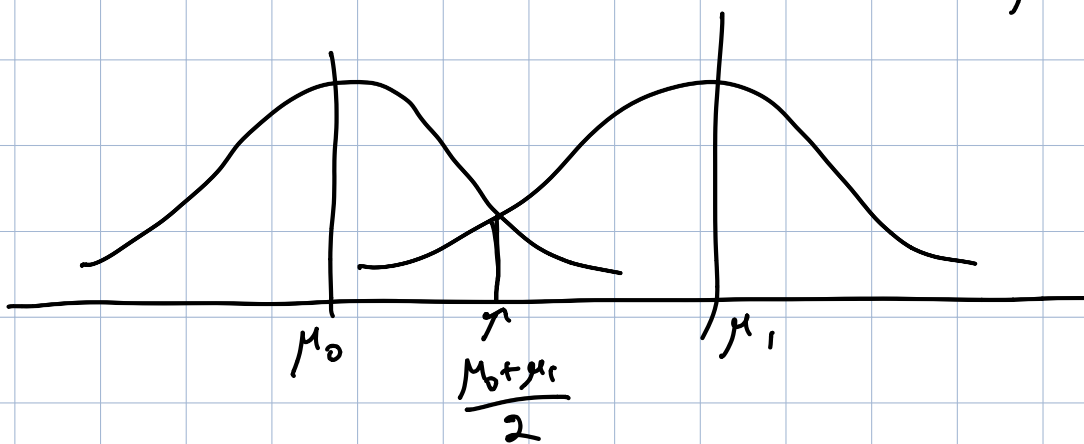
equi-probable classes

① $\sigma_0 = \sigma_1 = \sigma, \quad P_0 = P_1, \quad L(1,0) = L(0,1)$

Then, $h_B(x) = 0$ if

$$\frac{x}{\sigma^2}(\mu_0 - \mu_1) - \frac{1}{2\sigma^2}(\mu_0^2 - \mu_1^2) > 0$$

i.e., $x > \frac{\mu_0 + \mu_1}{2}$ (assuming $\mu_0 > \mu_1$)



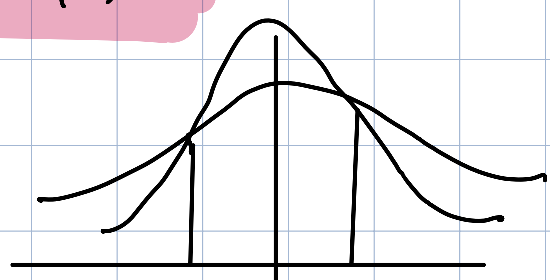
② $\mu_0 = \mu_1 = 0, \quad P_0 = P_1, \quad L(1,0) = L(0,1)$

$h_B(x) = 0$ if

$$\frac{x^2}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) - \log\left(\frac{\sigma_0}{\sigma_1}\right) > 0$$

$\sigma_0 > \sigma_1 \Rightarrow$

$$\frac{x^2}{2} > \frac{\sigma_1^2 \sigma_0^2 \log(\sigma_0/\sigma_1)}{(\sigma_0^2 - \sigma_1^2)}$$



③ Generalization to multivariate class-conditional densities:

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right), \quad i=0,1$$

$h_B(x) = 0$ if

CHECK!

$$\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_0^{-1}) x + x^T (\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1)$$

$$+ \frac{1}{2} (\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0)$$

$$+ \log\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) + \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$

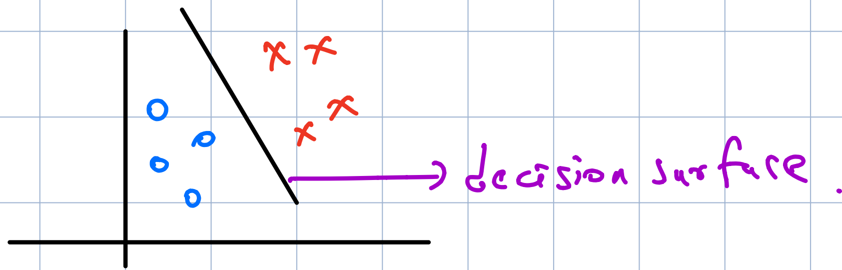
H.W: Workout the special cases as in the univariate setting.

DISCRIMINANT FUNCTIONS:

$$h(x) = \begin{cases} 0 & \text{if } g(x) > 0 \\ 1 & \text{else} \end{cases}$$

Example: 0-1 loss function. Bayes classifier is based on the discriminant function $g(x) = q_0(x) - q_1(x)$

$g(x) = 0$ is the decision surface.



Maximum likelihood estimation

① To implement Bayes classifier, we need class conditional densities & prior probabilities

② Suppose we are given i.i.d. samples from a class conditional distribution

$$\{x_1, \ldots x_n\}$$

③ We adopt a "parametric" approach to estimation

$\mathcal{D} = \{x_1, \ldots x_n\}$ iid from the distribution of r.v. X parameterized by θ.

Aim: Use $\mathcal{D}$ to estimate θ

Example:- $f(x|\theta) \sim N(\theta_1, \theta_2)$

(θ_1, θ_2) unknown

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\,\theta_2} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right)$$

$\hat{\theta}_n \leftarrow$ estimate of θ.

<u>Example:-</u> Suppose X is drawn uniformly at random from the set $\{1, ---, n\}$

Suppose "n" is unknown

Suppose we observe a sample "k".

Want an estimator of n given $X=k$ is observed.

<u>Max-likelihood (ML) idea:</u>

The pmf of X $\quad P_X(k) = \frac{1}{n} I\{1 \le k \le n\}$

Think of $P_X(k)$ as a function of n.

& find the maximizer of this function

$P_X(k)$ is 0 for $n \le k-1$, jumps to $\frac{1}{k}$ at $n=k$ & decreases beyond k. So, it is maximized at $n=k$.

$$\hat{n}_{MC}(k) = k.$$

Suppose we are given $\{x_1, \ldots x_n\}$ iid $\sim f(\cdot|\theta)$

E.g. $x_i \sim N(\underset{\substack{\uparrow \\ \text{unknown}}}{\theta}, 1) \leftarrow$ one-parameter distribution, $i = 1 \sim n$

$\hat{\theta}_n \leftarrow$ estimate of θ

Sample mean $\rightarrow \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\hat{\theta}_2 = \frac{x_1 + x_2}{2}$, $\hat{\theta}_1 = x_1$

Why choose $\hat{\theta}_n$ over $\hat{\theta}_2$ & $\hat{\theta}_1$?

Use a "mean-square error" objective

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta\right)^2$$

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

$$= E\left[(\hat{\theta} - E\hat{\theta})^2\right] + \left(E(\hat{\theta}) - \theta\right)^2 + 2\underbrace{E\left[(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta)\right]}_{=(E\hat{\theta} - \theta)E(\hat{\theta} - E\hat{\theta}) = 0}$$

$$= \underbrace{E\left[(\hat{\theta} - E\hat{\theta})^2\right]}_{\text{Variance}} + \underbrace{(E\hat{\theta} - \theta)^2}_{\text{bias}}$$

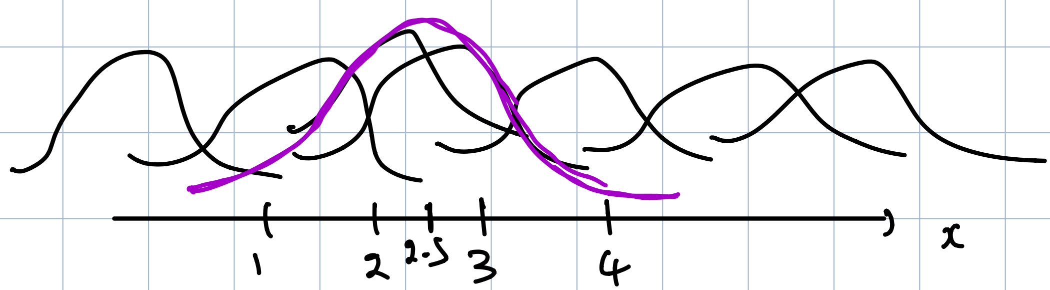
Estimator $\hat{\theta}$ unbiased if $E\hat{\theta} = \theta$

$Var(\hat{\theta}_n) = \frac{1}{n}$, $\quad Var(\hat{\theta}_2) = \frac{1}{2}$, $\quad Var(\hat{\theta}_1) = 1$

Max-likelihood estimation:

$\mathcal{D} = \{x_1, --- x_n\}$

E.g. $\quad x_i \sim \mathcal{N}(\theta, 1)$



Likelihood function $L(\theta) = \prod_{j=1}^{n} f(x_j | \theta)$

ML estimate: $\quad \theta^* \in \underset{\theta}{\arg\max} \{L(\theta) = L(x_1, \dots x_n, \theta)\}$

$$L(\theta^*) \geq L(\theta) \quad \forall \theta$$

Log-likelihood $\quad \ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} \log f(x_j | \theta)$

Example: Univariate normal

$$X_i \sim N(\theta_1, \theta_2^2), \quad i = 1 \dots n$$

$$l(\theta) = \log L(\theta) = \sum_{j=1}^{n} \left(-\log \theta_2 - \frac{1}{2}\log 2\pi - \frac{(x_j - \theta_1)^2}{2\theta_2^2} \right)$$

$$l(\theta) = -n \log \theta_2 - \frac{n}{2}\log 2\pi - \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{2\theta_2^2}$$

To find ML estimates $\hat{\theta}_1, \hat{\theta}_2$ of θ_1 & θ_2, do

(i) $\frac{\partial l}{\partial \theta_1} = 0$ & (ii) $\frac{\partial l}{\partial \theta_2} = 0$

$$\hat{\theta}_1 = \frac{1}{n} \sum_{j=1}^{n} x_j \qquad \hat{\theta}_2^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \hat{\theta}_1)^2$$

not unbiased

H.W.: ① Assume $\mathrm{Exp}(\lambda)$ likelihood & find the ML estimate

② Generalize to multivariate case

$$x_i \sim N(\mu, \Sigma), \quad i = 1 \dots n, \quad x_i \in \mathbb{R}^d$$

Calculate ML estimates for μ, Σ.

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \,; \quad \hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^T$$

(Recall $\Sigma = E(x-\mu)(x-\mu)^T$

<u>Example:-</u> Suppose a coin with bias θ is tossed n times & k heads occur.

$$P_{\hat{\theta}}(X=k) = \binom{n}{k}(\hat{\theta})^k (1-\hat{\theta})^{n-k} = L(\hat{\theta})$$

$$\hat{\theta}_{ML} = \arg\max_{\hat{\theta}} (\hat{\theta})^k ((1-\hat{\theta})^{n-k}, \quad n, k \text{ known}$$

Let $1 \le k \le n-1$

$$\frac{d}{d\hat{\theta}}\left(\hat{\theta}^k (1-\hat{\theta})^{n-k}\right) = \frac{k\hat{\theta}^k}{\hat{\theta}}(1-\hat{\theta})^{n-k} - \frac{(n-k)}{(1-\hat{\theta})}\hat{\theta}^k(1-\hat{\theta})^{n-k}$$

$$= \left(\frac{k}{\hat{\theta}} - \frac{(n-k)}{(1-\hat{\theta})}\right)\hat{\theta}^k (1-\hat{\theta})^{n-k}$$

$$= (k - n\hat{\theta})\,\hat{\theta}^{k-1}(1-\hat{\theta})^{n-k-1} \quad \text{———} (*)$$

If $\hat{\theta} \le \frac{k}{n}$, then $(*) \ge 0$

if $\hat{\theta} > \frac{k}{n}$, then $(*) < 0$

So, $\hat{\theta}_{ML} = \frac{k}{n}$

For $k=0$, $L(\hat{\theta}) = (1-\hat{\theta})^n$ & $\hat{\theta}_{ML} = 0$

For $k=n$, $L(\hat{\theta}) = \hat{\theta}^n$ & $\hat{\theta}_{ML} = 1$

H.W. ① Work out MLE estimate for (i) Poisson (λ) likelihood. & (ii) Geometric (θ) likelihood

---

Bayesian estimation

Frequentists	Bayesian
Probability = long run frequency	Probability = degree of belief (subjective)
Goal = algorithms with long-run freq. guarantees	Goal:- State & analyze beliefs

Bayesian approach:-

Parameter θ is given a prior density $\pi(\theta)$

Observe data $\{x_1, \ldots x_n\}$ sampled from $f(x|\theta)$

Compute posterior density (use Bayes rule)

$$f(\theta | x_1, \ldots x_n) \propto L(\theta)\, \pi(\theta)$$

Example: $D_n = \{x_1, \ldots x_n\}$

$x_i \sim Ber(\theta)$

Uniform prior $\pi(\theta) = 1$

Posterior probability

$$f(\theta | \mathcal{D}_n) \propto \underbrace{\pi(\theta)}_{\text{prior}} \underbrace{\mathcal{L}_n(\theta)}_{\text{likelihood}}$$

$$\propto 1 \times \prod_{i=1}^{n} f(x_i | \theta)$$

$$\propto \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

Denote $S_n = \sum_{i=1}^{n} x_i$

$$\propto \theta^{S_n} (1-\theta)^{n-S_n}$$

$$\propto \theta^{(S_n+1)-1} (1-\theta)^{(n-S_n+1)-1}$$

$$\boxed{\text{Beta density: } \pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}$$

So, $f(\theta | \mathcal{D}_n) \sim \text{Beta}(S_n+1, n-S_n+1)$

$$\text{Posterior mean} = \frac{S_n+1}{n+2}$$

H.W. Redo the calculation above for $\pi(\theta) \sim \text{Beta}(\alpha, \beta)$

Bayes estimation

Working procedure:

1. Choose a prior density $\pi(\theta)$
2. Observe data $\mathcal{D}_n = \{x_1, \dots x_n\}$ sampled iid from $f(x|\theta) \leftarrow$ likelihood
3. Update beliefs:

Calculate posterior density

Bayes rule

$$f(\theta \mid \underbrace{x_1, \dots, x_n}_{\mathcal{D}_n}) = \frac{f(x_1 \dots x_n|\theta)\,\pi(\theta)}{f(x_1 \dots x_n)}$$

$$= \frac{L_n(\theta)\,\pi(\theta)}{C_n}$$

Normalization constant $\longrightarrow C_n$

$$\underbrace{f(\theta| x_1 \dots x_n)}_{\text{posterior}} \propto \underbrace{L_n(\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}}$$

Point estimates

1. Posterior mean $\bar{\theta}_n = \int \theta\, f(\theta|\mathcal{D}_n)\, d\theta$

2. Maximum aposteriori probability (MAP) = mode of the posterior

Example: $\mathcal{D}_n = \{x_1, \dots x_n\}$

$x_i \sim Ber(\theta)$

Beta prior i.e., $\pi(\theta) \sim Beta(\alpha, \beta)$

Then,

$$f(\theta \mid \mathcal{D}_n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\theta^{S_n}(1-\theta)^{n-S_n}$$

$$\propto \theta^{\alpha+S_n-1}(1-\theta)^{\beta+n-S_n-1}$$

$$\Rightarrow \quad \theta \mid \mathcal{D}_n \sim \text{Beta}(\alpha+S_n,\ \beta+n-S_n)$$

Special case: $\alpha=\beta=1$ leads to uniform (flat) prior

Posterior mean $\quad \bar{\theta}_n = \dfrac{\alpha+S_n}{\alpha+\beta+n}$

Sum to 1

$$= \left(\frac{n}{\alpha+\beta+n}\right)\underset{\text{ML-estimate}}{\left(\frac{S_n}{n}\right)} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\underset{\text{prior mean}}{\left(\frac{\alpha}{\alpha+\beta}\right)}$$

Note: Conjugacy: Prior & posterior belong to the same family of parameterized distributions.

E.g. Bernoulli case: Beta prior $\Rightarrow$ conjugacy.

<u>Example:</u> $\quad \mathcal{D}_n = \{x_1, \ldots, x_n\}$

$$f(x \mid \theta) \sim N(\theta, \underset{\text{known}}{\sigma^2})$$

Prior $\pi(\theta) \sim N(\theta_0, \sigma_0^2)$

$$f(\theta \mid \mathcal{D}_n) \propto \left[\prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2\right)\right] \exp\left(-\frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right)$$

$$\propto \exp\Bigg(-\frac{1}{2}\underbrace{\left[\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma}\right)^2 + \left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right]}_{\text{quadratic in } \theta}\Bigg) \quad (*)$$

So, $f(\theta \mid \mathcal{D}_n) \sim N(\theta_n, \sigma_n^2)$

$$f(\theta \mid \mathcal{D}_n) \propto \exp\left(-\frac{1}{2}\left(\frac{\theta-\theta_n}{\sigma_n}\right)^2\right) \quad (**)$$

Equating the co-efficients of θ & θ^2 in $(*)$ & $(**)$

$$\frac{1}{\sigma_n^2} = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \Leftarrow \text{Coeff. of } \theta^2$$

$$\frac{\theta_n}{\sigma_n^2} = \frac{n}{\sigma^2}\frac{S_n}{n} + \frac{\theta_0}{\sigma_0^2} \Leftarrow \text{Co-eff of } \theta$$

$$\Rightarrow \theta_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\left(\frac{S_n}{n}\right) + \left(\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\right)\theta_0$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

Some observations:-

(1) θ_n lies between $\frac{S_n}{n}$ & θ_0

(2) $\sigma_0 \neq 0 \Rightarrow \theta_n \to \frac{S_n}{n}$ as $n \to \infty$

(3) $\sigma_0 \to 0 \Rightarrow$ Prior dominates

(4) $\sigma_0 >> \sigma \Rightarrow \theta_n \approx \frac{S_n}{n}$

Ref: Duda Hart Stork's book for Bayes Classifier & ML/Bayes estimation

---

Minimum mean square error estimation

Ref: Sec 4.9 of Bruce Hajek's notes on probability (ECE 313 UIUC)

I Constant estimators:

Y is a r.v. & we wish to estimate Y by a constant δ.

$$MSE(\delta) = E(Y-\delta)^2$$

$$= \int_{-\infty}^{\infty} (y-\delta)^2 f_Y(y)\,dy$$

It is easy to see that $\delta^* = EY$ minimizes the MSE.

So, EY is the MMSE estimate.

II Unconstrained estimators

Y is a r.v. & we observe X.

$$MSE = E(Y - g(X))^2$$

want to find the g that minimizes the MSE

Suppose you observe $X=10$.

$$E[Y \mid X=10] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\,dy$$

Claim: $E[Y \mid X=x]$ is the MMSE estimate

$$MSE = E\left[(Y - g(X))^2\right]$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} (y - g(x))^2 f_{Y|X}(y|x)\,dy\right] f_X(x)\,dx$$

minimized by $g^*(x) = E[Y \mid X=x]$

$$MMSE = E\left[\left(Y - E[Y|X]\right)^2\right]$$

show this $\overset{?}{=}$ $E[Y^2] - E\left[\left(E[Y|X]\right)^2\right]$

Some problems

① Bayes estimation, Gaussian: unknown mean
known variance σ^2

Aim: Sequential estimation of posterior mean & variance

From class notes (see above),

① $$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$= \frac{1}{\sigma^2} + \frac{n-1}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$= \frac{1}{\sigma^2} + \frac{1}{\sigma_{n-1}^2}$$

② $$\frac{\theta_n}{\sigma_n^2} = \frac{S_n}{\sigma^2} + \frac{\theta_0}{\sigma_0^2}$$

$$= \frac{S_{n-1}}{\sigma^2} + \frac{x_n}{\sigma^2} + \frac{\theta_0}{\sigma_0^2}$$

$$= \frac{\theta_{n-1}}{\sigma_{n-1}^2} + \frac{x_n}{\sigma^2}$$

Pb 2) $\quad x_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \quad x_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad x_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$$\bar{x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$C = \frac{1}{3} \sum_{i=1}^{3} x_i x_i^T = \frac{2}{3} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

u_1

Top eigenvalue $= \frac{4}{3}$ & corresponding eigenvector $= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$\tilde{x}_i = (x_i^T u_1) u_1 + \cancel{(\bar{x}^T u) u_2} \; 0$$

$$\tilde{x}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \quad \tilde{x}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \tilde{x}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Pb 3) Linear estimator

$$MSE = E\left[(Y - (aX + b))^2 \right]$$
$$= E\left[((Y - aX) - b)^2 \right]$$

So, for a given a, $\quad b = E(Y - aX)$
$$= \mu_Y - a\mu_X$$

The estimator has the form $aX + \mu_Y - a\mu_X$

$$MSE = E\left[\left(Y - \mu_Y - aX + a\mu_X\right)^2\right]$$

$$= E\left[\left[(Y - aX) - (\mu_Y - a\mu_X)\right]^2\right]$$

$$= Var(Y - aX)$$

$$= Cov(Y - aX, Y - aX)$$

$$= Var(Y) - 2a\, Cov(Y, X) + a^2 Var(X) \quad (*)$$

Differentiate $(*)$ wrt a to obtain

$$a^* = \frac{Cov(Y, X)}{Var(X)}$$

Best linear estimator $= L^*(X) = a^* X + b^*$

$$= a^* X + \mu_Y - a^* \mu_X$$

$$= \left(\frac{Cov(Y, X)}{Var\, X}\right)(X - \mu_X) + \mu_Y$$

$$= \mu_Y + \sigma_Y \rho_{X,Y} \left(\frac{X - \mu_X}{\sigma_X}\right)$$

$$MMSE = Var(Y - a^* X)$$

$$= \sigma_Y^2 - \left(\frac{\sigma_Y^2 \rho_{X,Y}^2}{\sigma_X^2}\right) \cdot \sigma_X^2$$

$$= \sigma_Y^2 (1 - \rho_{X,Y}^2)$$

Pb 4) ML estimation for Geometric (θ)

$\{x_1 - - x_n\}$

$S_n = \sum_{i=1}^{n} x_i$

$$L(\theta) = \prod_{i=1}^{n} (1-\theta)^{x_i - 1} \theta$$

$$= (1-\theta)^{S_n - n} \theta^n$$

$$L'(\theta) = -(1-\theta)^{S_n - n - 1} (S_n - n) \theta^n + (1-\theta)^{S_n - n} n \theta^{n-1}$$

$$= (1-\theta)^{S_n-n-1} \theta^{n-1} \left[-(S_n - n)\theta + n(1-\theta) \right]$$

$$= (1-\theta)^{S_n-n-1} \theta^{n-1} \left[n - S_n \theta \right]$$

$L' > 0$ if $n > S_n \theta$ i.e., $\theta < \frac{1}{\left(\frac{S_n}{n}\right)}$

$= 0$ if $\theta = \frac{1}{\left(\frac{S_n}{n}\right)}$

< 0 if $\theta > \frac{1}{\left(\frac{S_n}{n}\right)}$

So, $\hat{\theta}_{ML} = \frac{1}{\left(\frac{S_n}{n}\right)}$

Pb 5) Intuitive justification for ML estimate of variance being biased

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right) = \sigma^2$$

Recall $\hat{\sigma}^2_{MC} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_n)^2$

Observe that

$$\sum_{i=1}^{n}(x_i-\hat{\mu}_n)^2 \le \sum_{i=1}^{n}(x_i-\mu)^2 \quad \text{— } (*)$$

Why?

$$\text{RHS} = \sum_{i=1}^{n}\left((x_i-\hat{\mu}_n)+(\hat{\mu}_n-\mu)\right)^2$$

Cross term vanishes

$$= \sum_{i=1}^{n}(x_i-\hat{\mu}_n)^2 + \sum_{i=1}^{n}(\hat{\mu}_n-\mu)^2 + 0$$

$$\ge \sum_{i=1}^{n}(x_i-\hat{\mu}_n)^2$$

From $(*)$,
$$E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_n)^2\right) \le \sigma^2$$