16/2/22 ML Assignment :1

Name Mahendra Kumar

Roll no 1203209

Q.1 Given, There are only classes (1 & 2 s)

$P_1 = P_2 = 1/2$

also, class conditional density's f_i's are

$$f_i(x) = \frac{1}{\pi b} \cdot \frac{1}{1+\left(\frac{x-a_i}{b}\right)^2} \qquad i=1,2.$$

(a) To show: f_i's are probability density functions.

Sol: If f_i's are probability density function then

$$\int_{-\infty}^{\infty} f_i(x)\,dx = 1.$$

$$LHS = \int_{-\infty}^{\infty} f_i(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\pi b} \cdot \frac{1}{1+\left(\frac{x_i-a}{b}\right)^2}\,dx$$

$$= \frac{1}{\pi b} \cdot \left[\tan^{-1}\left(\frac{x_i-a}{b}\right) \cdot b\right]_{-\infty}^{\infty}$$

$$= \frac{1}{\pi}\left[\tan^{-1}(\infty) - \tan^{-1}(-\infty)\right]$$

$$= \frac{1}{\pi}\left(\pi/2 - (-\pi/2)\right) = \frac{\pi}{\pi} = 1. = RHS.$$

Hence Proved. □.

So $\boxed{\int_{-\infty}^{\infty} f_i(x)\,dx = 1}$ Ans.

(b) Using Bayes rule,

$$q_i(x) = \frac{b_i f_i(x)}{b \sum_{i=1}^{2} b_i f_i(x)}$$

posterior

$\because b_1 = b_2 = \frac{1}{2}$

$$\Rightarrow q_i(x) = \frac{\frac{1}{2} f_i(x)}{\frac{1}{2}(f_1(x) + f_2(x))} = \frac{f_i(x)}{f_1(x) + f_2(x)}$$

Decision boundary:

if $q_1(x) \geq q_2(x) \Rightarrow x$ belongs to 1st class.

$$q_1(x) \geq q_2(x) \Leftrightarrow \frac{f_1(x)}{f_1(x)+f_2(x)} \geq \frac{f_2(x)}{f_1(x)+f_2(x)}$$

$f_1(x)+f_2(x) \neq 0$

$$\Rightarrow f_1(x) \geq f_2(x)$$

$$\Rightarrow \left(\frac{1}{\pi b}\right) \cdot \frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} \geq \left(\frac{1}{\pi b}\right) \cdot \frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}$$

$$\Rightarrow (x-a_2)^2 \geq (x-a_1)^2$$

$$\Rightarrow x^2 - 2a_2x + a_2^2 \geq x^2 - 2a_1x + a_1^2$$

$$\Rightarrow 2x(a_2-a_1) \leq a_2^2 - a_1^2$$

(b) $2x(a_2 - a_1) \leq (a_2 - a_1)(a_2 + a_1)$.

case-1 if $a_1 = a_2$ this is always true. i.e. all point belongs to class 1.

case-2 if $a_1 \neq a_2$ then.

$$2x \leq a_2 + a_1$$

$$\boxed{x \leq \left(\frac{a_2 + a_1}{2}\right)} \rightarrow \text{decision boundary.}$$

(c) probability of error for Bayes classifier with 0-1 loss function.

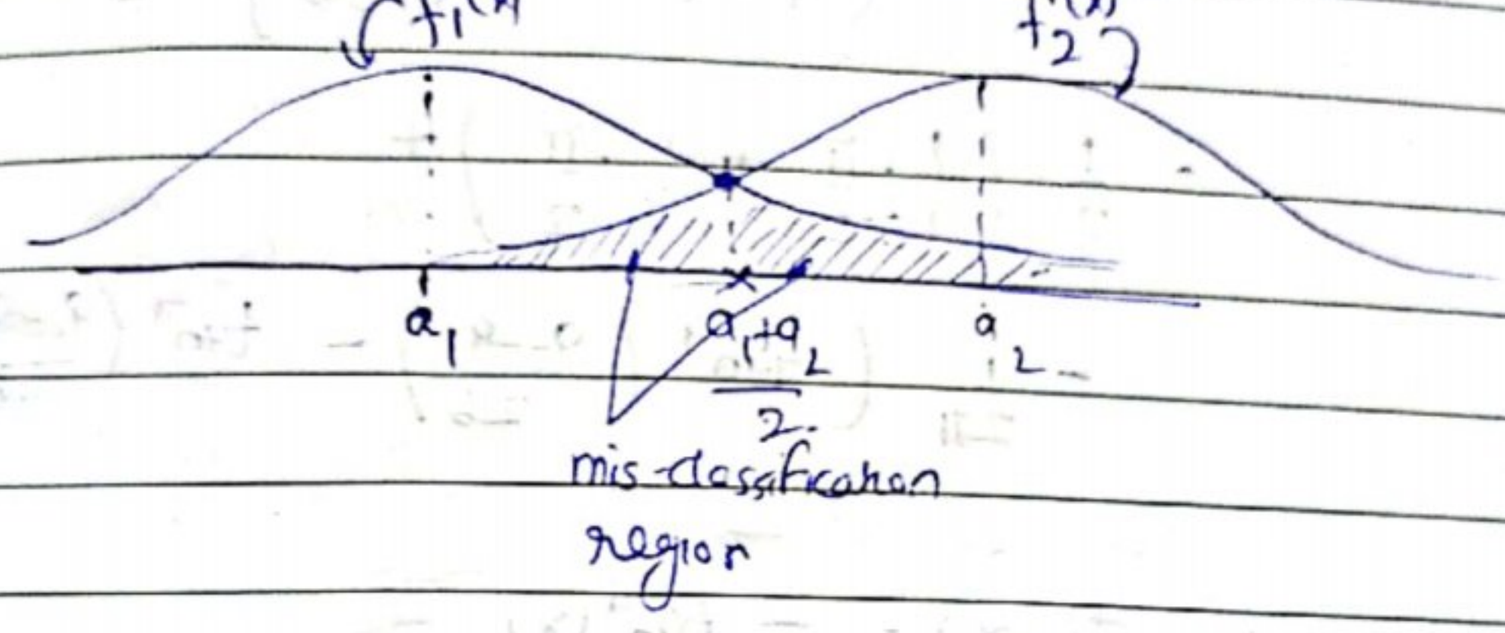
from previous part it is clear that our bayes classifier is $h_B(x)$ given by

$$h_B(x) = \begin{cases} 1 & \text{if } x \leq \frac{a_1 + a_2}{2} \\ 2 & \text{if } x > \frac{a_1 + a_2}{2} \end{cases}$$

classmate
Date
Page

continue...

(c)



$P_{error} = \int_{\frac{a_1+a_2}{2}}^{a_2} P(\text{error})$

So probability of error is

$P_{error} = P(h_B(x) \neq \text{true class})$

$= P(\text{true class} = 1) \cdot P(h_B(x) \neq \text{true class} \mid \text{true class} = 1)$

$+ P(\text{true class} = 2) \cdot P(h_B(x) \neq \text{true class} \mid \text{true class} = 2)$

$= p_1 \cdot P(h_B(x) \neq 1) + p_2 \cdot P(h_B(x) \neq 2)$

$P(h_B(x) \neq 1)$: Case when $x > \frac{a_1+a_2}{2}$ but x belongs to class 1.

$P(h_B(x) \neq 2)$: Case when $x \leq \frac{a_1+a_2}{2}$ but x belongs to class 2.

Decision boundary line

$= p_1 \cdot 1 \cdot \int_{\frac{a_1+a_2}{2}}^{\infty} f_1(x)\,dx + p_2 \cdot 1 \cdot \int_{-\infty}^{\frac{a_1+a_2}{2}} f_2(x)\,dx + 0 + 0$

$= \frac{1}{2} \cdot \left(\frac{b}{\pi b}\left[\tan^{-1}(\infty) - \tan^{-1}\left(\frac{\frac{a_1+a_2}{2} - a_1}{b}\right)\right]\right)$

$+ \frac{1}{2} \cdot \left(\frac{b}{\pi b}\left(\tan^{-1}\left(\frac{\frac{a_1+a_2}{2} - a_2}{b}\right) - \tan^{-1}(-\infty)\right)\right]$

$$\Rightarrow P_{error} = \frac{1}{2}\cdot\left(\frac{1}{\pi}\left(\frac{\pi}{2} - \tan^{-1}\left(\frac{a_2-a_1}{2b}\right)\right)\right)$$

$$+\frac{1}{2}\cdot\left(\frac{1}{\pi}\cdot\left(\pi\tan^{-1}\left(\frac{a_1-a_2}{2b}\right) - \frac{\pi}{2}\left(\frac{-\pi}{2}\right)\right)\right)$$

$$= \frac{1}{2}\cdot\left(\frac{1}{\pi}\cdot\frac{\pi}{2} + \frac{1}{\pi}\cdot\frac{\pi}{2}\right) +$$

$$-\frac{1}{2\pi}\left(\tan^{-1}\left(\frac{a_2-a_1}{2b}\right) - \tan^{-1}\left(\frac{a_1-a_2}{2b}\right)\right)$$

$\because \tan^{-1}(-x) = -\tan^{-1}(x)$

$$\Rightarrow P_{error} = \frac{1}{2}\cdot\frac{\pi}{\pi}\left(\frac{\pi}{\pi}\right) - \frac{1}{2\pi}\left(2\tan^{-1}\left(\frac{a_2-a_1}{2b}\right)\right)$$

$$\boxed{P_{error} = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left(\frac{a_2-a_1}{2b}\right)} \quad \underline{\text{Ans}}$$

Q-2. Given

$P(h(x)=i|x)$ → probability of choosing class i out of a class.

(a) Risk of the classifier on 0-1 loss function

$R(h(x)=i|x)$ → risk in choosing class i out of a classes defined as

$$R(h(x)=i|x) = \sum_{j=1}^{a} L(h(x)=i, Y=j)\, P(h(x)=j|x)$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{a} 1 \cdot P(h(x)=j|x) + 0 \times P(h(x)=i|x)$$

$$R(h(x)=i|x) = 1 - P(Y=i|x)$$

$$\left(\because \sum_{j=1}^{a} P(h(x)=j|x) = 1\right)$$

(b) Current our decision rule is to for given x choose i s.t.

$$i = \underset{j}{\operatorname{argmax}}\ P(h(x)=j|x) \quad \cdots$$

$$\Leftrightarrow \quad i = \underset{j}{\operatorname{argmin}}\ (-P(h(x)=j|x)$$

$$i = \underset{j}{\operatorname{argmin}}\ (1 - P(h(x)=j|x)) \quad \text{---} (2)$$

to get best classifier we want.
risk $R(h(x)=i|x)$ to be minimum
for chosen i i.e.

$$i = \underset{j}{\operatorname{argmin}}\ R(h(x)=j|x)$$

$$i = \underset{j}{\operatorname{argmin}}\ \left(1 - P(Y=j|x)\right) \quad -(2)$$

Comparing (1) & (2) we will get improved classifier if:

$$i = \underset{j}{\operatorname{argmin}}\ \left(1 - P(Y=j|x)\right) = \underset{j}{\operatorname{argmin}}\ \left(1 - P(h(x)=j|x)\right)$$

i.e. $1 - P(Y=i|x) = 1 - P(h(x)=i|x)$

$\Leftrightarrow P(h(x)=1|x) = P(Y=1|x)$

$\Rightarrow$ i.e. choose class i out of a class if based on decision rule then $P(h(x)=i|x)$ is equal to posterior probability of $Y=i$ given x.

Page

Q-3. Given, loss function L

$$L(h(x)=i, Y=j) = \begin{cases} 0 & i=j \\ \lambda_r & i=K+1 \rightarrow \text{rejection action} \\ \lambda_m & \text{otherwise} \end{cases}$$

$\lambda_r \rightarrow$ loss for choosing rejection class (K+1)

$\lambda_m \rightarrow$ loss for mis-classification

$K \rightarrow$ total no. of classes.

Let $R(h(x)=i \mid x) \rightarrow$ expected loss / risk associated with taking i^{th} action for given x then.

$i = 1, 2, \ldots K$ when # no of classes $= K$

$$R(h(x)=i \mid x) = \sum_{j=1}^{K} L(h(x)=i, Y=j)\, P(Y=j \mid x)$$

$$= \sum_{j=1}^{i-1} L(h(x)=i, Y=j) P(Y=j \mid x) +$$

$$L(\underbrace{h(x)=i, Y=i}_{\downarrow\downarrow\ 0}) P(Y=i \mid x) +$$

$$\sum_{j=i+1}^{K} L(h(x)=i, Y=j)\, P(Y=j \mid x)$$

$$= \phi \cdot \sum_{\substack{j=1 \\ j \neq i}}^{K} L(\underbrace{h(x)=i, Y=j}_{\downarrow\downarrow\ \lambda_m \ \forall\ j \ \&\ i=1,2\ldots K}) P(Y=j \mid x)$$

$$= \lambda_m \sum_{\substack{j=1 \\ j \neq i}}^{K} P(Y=j \mid x)$$

$$R(h(x)=i \mid x) = \lambda_m \left(1 - P(Y=i \mid x)\right)$$

$$\left(\because \sum_{j=1}^{K} P(Y=j \mid x) = 1\right)$$

$$\Rightarrow R(h(x)=i \mid x) = \lambda_m \left(1 - q_i(x)\right)$$

↳ posterior probability for choosing i given x.

case-2 when $i = K+1$ i.e. rejection

$$R(h(x)=K+1 \mid x) = \sum_{j=1}^{K} L\underbrace{(h(x)=K+1, Y=j)}_{\Downarrow} P(Y=j \mid x)$$

$$\lambda_r \ \forall j$$

$$= \lambda_r \sum_{J=1}^{K} P(Y=j \mid x)$$

$$R(h(x)=K+1 \mid x) = \lambda_r$$

Now, using bayesian decision theory. we decide a class i if

(i) $R(h(x)=i \mid x) \leq R(h(x)=j \mid x) \quad \forall j = 1, 2, \ldots K$

⇒ of i.e. and.

(ii) & $R(h(x)=i \mid x) \leq R(h(x)=K+1 \mid x)$.

Simplifying (i) gives

Simplification - (i)

Case-1. if $i < K+1$ then.

$\lambda_m (1 - q_i(x)) \leq \lambda_m (1 - q_j(x))$

$\Rightarrow q_i(x) \geq q_j(x)$ provided $\lambda_m \neq 0$

Case-2. if $i = K+1$.

$\lambda_r \leq \lambda_m (1 - q_j(x))$

$\Rightarrow q_j(x) \leq 1 - \frac{\lambda_r}{\lambda_m}$

Since $q_j(x) \geq 0 \Rightarrow$ we choose rejection class over any other class if $1 - \frac{\lambda_r}{\lambda_m} \geq q_j(x) \geq 0 \Rightarrow \boxed{\lambda_r \leq \lambda_m}$

Simplification - (ii)

$$R(h(x) = i \mid x) \leq R(h(x) = K+1 \mid x)$$

Case-I Clearly for $i = K+1$ equality holds so no need to check.

Case-II. for $i < K+1$

LHS $\Rightarrow \lambda_m (1 - q_i(x)) \leq \lambda_r \cdot 1 \cdot \Leftarrow$ RHS

$\Rightarrow q_i(x) \geq 1 - \frac{\lambda_r}{\lambda_m}$.

Since $q_i(x) \leq 1$ this value make sense if

$$1 - \frac{\lambda_r}{\lambda_m} \leq q_i(x) \leq 1.$$

$$\Rightarrow \boxed{\frac{\lambda_r}{\lambda_m} \geq 0}$$ which is always true.

as

Part (b) of question.

- What happen if $\lambda_r = 0$.

See Simplification -(i) case -2 : if $\lambda_r = 0$ then this case always hold. i.e. we will always reject. and expected loss will in turn be 0.

- What happen if $\lambda_r > \lambda_m$

See Simplification -(i) case - II, if $\frac{\lambda_r}{\lambda_m} > 1$.

$\Rightarrow q_i(x) \geq 1 - \frac{\lambda_r}{\lambda_m}$ but $q_i(x) \geq 0 \ \forall i$ (non-negative)

$\hookrightarrow$ negative.

$\Rightarrow$ this condition holds for every i

So we will definitely not select rejection class.

$\Rightarrow$ So by Simplification-(i) Case-1 we will select class j having maximum posterior probability i.e. $q_j(x)$.

Q-4 Maximum Likelihood Estimation

① exponential distribution

the pdf of exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Let take $D = \{X_i\}_{i=1}^{N}$ be a iid sample from given distribution with unknown parameter λ. then likelihood will be

$$L(D|\lambda) = f(X_1 = x_1, \ldots X_N = x_n)$$

$$L(D|\lambda) = \prod_{i=1}^{N} f(X_i = x_i) \quad (\because \lambda i's \text{ are iid})$$

take log on both sides.

$$\log(L(D|\lambda)) = l(\lambda) = \sum_{i=1}^{N} \log\left(f(X_i = x_i)\right)$$

$l(\lambda)$ (say)

$$= \sum_{i=1}^{N} \log\left(\lambda e^{-\lambda x_i}\right)$$

$$= \sum_{i=1}^{N} \log\lambda - \sum_{i=1}^{N} \lambda x_i$$

$$\boxed{l(\lambda) = N\log\lambda - \lambda \sum_{i=1}^{N} x_i} \quad \text{— (1)}$$

$$\lambda_{MLE} = \underset{\lambda}{\operatorname{argmax}}\ \ell(\lambda)$$

taking differentiation on both sides of (1) to get w.r.t. λ we get.

$$\frac{d(\ell(\lambda))}{d\lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i \quad \text{— (2)}$$

equating (2) to zero gives maximum likelihood estimation of λ.

$$\lambda_{MLE} = \frac{N}{\sum_{i=1}^{N} x_i}$$

(3) Multivariate Gaussian Distribution

Assume we have N random vectors each of size p :- $X^{(1)}, X^{(2)} \ldots X^{(N)}$ where each random vector can be interpreted as observation across p variables. If each $X^{(i)}$ are i.i.d. as multivariate Gaussian vectors.

$$X^{(i)} \sim \mathcal{N}(\vec{\mu}, \Sigma)$$

↳ covariance matrix of shape $p \times p$

where $\vec{\mu}$, Σ are unknown parameters.

Note: Note

Note that by independence of random vectors the joint density of data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ is product of individual densities.

i.e. $\prod_{i=1}^{N} f_{X^{(i)}}(x^{(i)} | \vec{\mu}, \Sigma)$

take log on both sides gives log-likelihood function

$$l(\vec{\mu}, \Sigma | \mathcal{D}) = \log \left(\prod_{i=1}^{N} f_{X^{(i)}}(x^{(i)} | \vec{\mu}, \Sigma) \right)$$

$$= \log \left(\prod_{i=1}^{N} \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{\left(-\frac{1}{2} (x^{(i)} - \vec{\mu})^T \Sigma^{-1} (x^{(i)} - \vec{\mu}) \right)} \right)$$

$$= \sum_{i=1}^{N} -\frac{p}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x^{(i)} - \vec{\mu})^T \Sigma^{-1} (x^{(i)} - \vec{\mu})$$

$$\boxed{l(\vec{\mu}, \Sigma | \mathcal{D}) = -\frac{Np}{2} \log 2\pi - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x^{(i)} - \vec{\mu})^T \Sigma^{-1} (x^{(i)} - \vec{\mu})} \quad \text{--- (1)}$$

Deriving μ_{MLE}

take derivative w.r.t. $\vec{\mu}$ given

$$\frac{\partial}{\partial \mu} \left(l(\vec{\mu}, \Sigma | \mathcal{D}) \right) = \frac{-1}{2} \left(\sum_{i=1}^{N} 2\Sigma^{-1} (x^{(i)} - \vec{\mu}) \right)$$

{ ∵ $\frac{\partial}{\partial w}(w^T A w) = 2Aw$ if A is symmetric }

equating it to 0 given :

$$\Rightarrow \sum_{i=1}^{N} \Sigma^{-1} \cdot (x^{(i)} - \vec{u}) = 0 \quad \ldots$$

$$\Rightarrow \sum_{i=1}^{N} (x^{(i)} - \vec{u}) = 0 \quad \left(\because \text{by multiply } \Sigma \text{ on both sides} \right.$$

$$\left. \& \; \Sigma \Sigma^{-1} = I \rightarrow \text{identity matrix.} \right.$$

$$\left. \& \; \Sigma \cdot 0 = 0 \rightarrow \text{null matrix} \right)$$

$$\Rightarrow \sum_{i=1}^{N} x^{(i)} - N\vec{u} = 0 \Rightarrow \boxed{\vec{u}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}}$$

Ans

Deriving Σ_{MLE}.

Deriving MLE for covariance matrix Σ we will use following results from linear algebra & calculus.

① trace is invariant on cyclic permutation of matrix products.

$$tr[ABC] = tr[CAB] = tr(BCA)$$

② ∵ $x^T A x$ is a scalar we take take trace & it remain same.

i.e. $tr(x^T A x) = x^T A x$

③ $\frac{\partial}{\partial A}(tr(AB)) = B^T$

(1) $\frac{\partial}{\partial A}(\log|A|) = (A^{-1})^T$

Combining these properties we get:

$$\frac{\partial}{\partial A}(x^T A x) = \frac{\partial}{\partial A}(x^T x A) = (x x^T)^T = x x^T$$

Now back to problem —

We can write log-likelihood function and compute derivative w.r.t Σ^{-1} & using $|\Sigma^{-1}| = |\Sigma|^{-1}$ we get:

$$l(\mu, \Sigma | D) = -\frac{Np}{2}\log 2\pi + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{N}(x^{(i)} - \vec{\mu})^T \Sigma^{-1} (x^{(i)} - \vec{\mu})$$

$$\frac{\partial}{\partial \Sigma^{-1}} l(\mu, \Sigma | D) = 0$$

$$\Rightarrow 0 + \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}(x^{(i)} - \vec{\mu})(x^{(i)} - \vec{\mu})^T = 0$$

$$\left(\because (\Sigma^{-1})^T = \Sigma^T = \Sigma\right)$$

$$\Rightarrow N\Sigma = \sum_{i=1}^{N}(x^{(i)} - \mu)(x^{(i)} - \vec{\mu})^T$$

$$\Rightarrow \boxed{\Sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \vec{\mu})(x^{(i)} - \vec{\mu})^T}$$

Q.5

unknown parameter $\theta := \mu$

Data $D = \{x_i\}_{i=1}^{N}$ as iid sampled from normal distribution. $f_X(x|\theta) \sim N(x; \mu, \sigma^2)$ ← known.

i.e. $f_{X_i}(x_i|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ — (I)

also given $\theta = \mu \sim N(\mu_0, \sigma_0^2)$ ← known.

i.e. $b(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$ — (II).

this is pdf not probability.

posterior distribution $b(\mu|D)$ ← this is p.d.f. not probability.

(a) Baysian estimation: with parameter corresponding to mode of posterior distribution by bayes rule.

$$b(\mu|D) = \frac{f_X(D|\mu) \cdot b(\mu)}{f_X(D)} \quad \text{— (A)}$$

$\because f_X(D)$ is independent of unknown parameter μ

$\therefore \theta_{MAP} = \underset{\theta=\mu}{\text{argmax}}\ b(\mu|D)$

MAP = means maximum aposteriori

mode → μ at which $b(\mu|D)$ attain its maximum value.

$$\mu_{MAP} = \underset{\theta=\mu}{\text{argmax}} \frac{f_X(D|\mu) \cdot b(\mu)}{f_X(D)}$$

$$= \underset{\theta=\mu}{\text{argmax}}\ f_X(D|\mu) \cdot b(\mu)$$

Assuming samples $\mathcal{D} = \{x_i\}_{i=1}^{N}$ are iid

$$\mu_{MAP} = \underset{\mu}{\operatorname{argmax}} \left(\prod_{i=1}^{N} f_{X_i}(x_i|\mu) \right) \cdot p(\mu)$$

↑ log likelihood ↑ prior

taking log on both sides gives.

$$l(\mu) = \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{N} \log \left(f_{X_i}(x_i|\mu) + \log p(\mu) \right)$$

↳ lets call this $l(\mu)$

$$l(\mu) = \sum_{i=1}^{N} \log \left(f_{X_i}(x_i|\mu) + \log (p(\mu)) \right)$$

$$= \sum_{i=1}^{N} \left(\frac{-1}{2} \log 2\pi - \log(\sigma) - \frac{1}{2\sigma^2} (x_i - \mu)^2 \right)$$

$$+ \log \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) \cdot - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2$$

taking differentiation on both side w.r.t. μ and set it to zero gives. critical point

$$\frac{\partial l(\mu)}{\partial \mu} = 0 \Rightarrow \left[\sum_{i=1}^{N} \frac{1}{\sigma^2} (x_i - \mu) \right] - \frac{1}{\sigma_0^2} (\mu - \mu_0) = 0$$

$$\Rightarrow \left(\sum_{i=1}^{N} \frac{1}{\sigma^2} x_i + \frac{\mu_0}{\sigma_0^2} \right) = \mu \left(\sum_{i=1}^{N} \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right)$$

$$\mu = \frac{\left(\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i + \frac{\mu_0}{\sigma_0^2}\right)}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\mu_{MAP} = \frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{N} x_i + \frac{\mu_0}{\sigma_0^2}}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}\right) \quad \text{--- (II)}$$

So density function using parameter corresponding to mode of posterior distribution is

$$f_X(x \mid \mu_{MAP}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_{MAP})^2}{2\sigma^2}} \quad \text{- Ans}$$

where μ_{MAP} is as defined in (III)

② Baysian estimator corresponding to mean of posterior distribution

~~$\hat{\mu}$ ...~~

~~again by changing~~ Using (*) from previous part

$$p(\mu \mid D) = \frac{f_X(D \mid \mu) \cdot p(\mu)}{f_X(D)} \rightsquigarrow \text{independent of } \mu$$

$$\propto f_X(D \mid \mu) \cdot p(\mu)$$

$$= \text{Constant} \times f_X(D \mid \mu)\, p(\mu)$$

$$p(\mu \mid D) = K \cdot f_X(D \mid \mu)\, p(\mu)$$

↳ say

Assuming samples are i.i.d. we get...

$$p(\mu|D) = K \times \left(\prod_{i=1}^{N} p_{x_i}(x_i|\mu) \right) \cdot p(\mu)$$

$$= K \times \left(\prod_{i=1}^{N} \frac{e^{-(x_i-\mu)^2}}{\sqrt{2\pi}\sigma} \right) \cdot e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

$$\propto e^{-\left[\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} + \frac{(\mu-\mu_0)^2}{2\sigma_0^2} \right]}$$

$\propto$ proportional symbol

$$-\sum_{i=1}^{N} (x_i-\mu)^2\sigma_0^2 - \sigma^2(\mu-\mu_0)^2$$

$$\propto e \qquad 2\sigma_0^2\sigma^2$$

On simplification we get-

$$\propto \exp\left\{ \left[-\mu^2\left(\sum_{i=1}^{N} \sigma_0^2 + \sigma^2 \right) + 2\mu\left[\left(\sum_{i=1}^{N} x_i\sigma_0^2 \right) + \mu_0\sigma^2 \right] - \left(\left[\sum_{i=1}^{N} x_i^2\sigma_0^2 \right] + \mu_0^2\sigma^2 \right) \right] \right\}$$

$$\propto \exp\left\{ -\mu^2 + 2\mu\left[\frac{\sum x_i\sigma_0^2 + \mu_0\sigma^2}{\sum \sigma_0^2 + \sigma^2} \right] - \left[\frac{\sum x_i^2\sigma_0^2 + \mu_0^2\sigma^2}{\sum \sigma_0^2 + \sigma^2} \right] \right\}$$

Let $\sum_{i=1}^{N} x_i$

Let $\mu_1 = \dfrac{\sum_{i=1}^{N} x_i \sigma_0^2 + \mu_0 \sigma^2}{\sum_{i=1}^{N} \sigma_0^2 + \sigma_1^2}$ & $\sigma_1^2 = \dfrac{\sigma_0^2 \sigma^2}{\sum_{i=1}^{N} \sigma_0^2 + \sigma^2}$

& $C_1 = \text{constant}_2 = -\left[\dfrac{\sum_{i=1}^{N} x_i^2 \sigma_0^2 + \mu_0^2 \sigma^2}{\sum_{i=1}^{N} \sigma_0^2 + \sigma^2}\right]$

then

$$p(\mu|D) \propto \exp\left(\frac{-\mu^2 + 2\mu\mu_1 + C_1}{2\sigma_1^2}\right)$$

$$\propto \exp\left(\frac{-(\mu-\mu_1)^2}{2\sigma_1^2}\right) \cdot \exp\left(\frac{C_1 + \mu_1^2}{2\sigma_1^2}\right)$$ (also constant)

(by adding & subtracting μ_1^2 in numerator)

$$\Rightarrow p(\mu|D) \propto \exp\left(-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}\right)$$

$$\Rightarrow p(\mu|D) = \text{constant} \times \exp\left(\frac{-(\mu-\mu_1)^2}{2\sigma_1^2}\right)$$

Clearly $p(\mu|D)$ is normally distributed s.t

$$p(\mu|D) \sim N(\mu_1, \sigma_1)$$

Let's Simplify μ_1 & σ_1 and a little

$$\mu_1 = \frac{\sum_{i=1}^{N} x_i \sigma_0^2 + \mu_0 \sigma^2}{\sum_{i=1}^{N} \sigma_0^2 + \sigma^2} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{\mu_0 \sigma^2}{\sigma_0^2 N}}{1 + \frac{\sigma^2}{N\sigma_0^2}}$$

$\because$ Sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\Rightarrow \mu_1 = \frac{\bar{x} + \frac{\mu_0 \sigma^2/\sigma_0^2}{N}}{1 + \frac{(\sigma^2/\sigma_0^2)}{N}}$$

and Variance σ_1^2 is

$$\sigma_1^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2} = \frac{\sigma_0^2 (\sigma/\sigma_0)^2}{N + (\sigma/\sigma_0)^2}$$

$$\sigma_1^2 = \frac{\sigma^2}{N + (\sigma/\sigma_0)^2}$$

I guess you noticed we have found our answer i.e.

μ mean of posterior distribution is

$$\mu_{mean} = \mu_1 = \frac{\frac{\bar{x} + \mu_0 \sigma^2/\sigma_0^2}{N}}{\frac{1 + \sigma^2/\sigma_0^2}{N}}$$ Ans =.

Key observation:

a) $N \to \infty$.

$\mu_{mean} \to \bar{x} = \mu_{MLE}$

μ_{mean} → mean posterior estimation

μ_{MLE} → maximum likelihood estimator.

So density function using mean of posterior distribution is

$$f_X(x|\mu_{mean}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_{mean})^2}{2\sigma^2}}$$

where μ_{mean} is as defined above.

Q6. Derive E & M steps for

(a) Mixture of gaussians

Let $X_i \sim N(\mu, \sigma^2)$. In the scenario our X_i comes from mixture model with (say) K mixture components. So we have,

Conditional distribution

$$(X_i | Z_i = k) \sim N(\mu_k, \sigma_k^2)$$

where $Z_i \in \{1, 2, \ldots, K\}$

↳ Identifiable.

So, marginal distribution of X_i is

$$P(X_i = x_i) = \sum_{j=1}^{K} P(X_i = x_i | Z_i = j)\, P(Z_i = j)$$

$$P(X_i = x_i) = \sum_{j=1}^{K} \pi_j P(X_i = x_i | Z_i = j)$$

Similarly, Joint probability of observation $X_1, X_2, \ldots, X_N$ assuming iid sample is

$$P(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} \left(\sum_{j=1}^{K} \pi_j P(X_i = x_i | Z_i = j) \right)$$

$$\Rightarrow f_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} \left(\sum_{j=1}^{K} \pi_j f_{X_i}(x_i | \mu_j, \sigma_j^2) \right) \quad \hookrightarrow (1)$$

where $f_{X_i}(x_i | \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\,\sigma_j} e^{-\frac{(x_i - \mu_i)^2}{2\sigma_j^2}}$

Aim :- find maximum likelihood estimates π_j, μ_j & σ_j's given a dataset $\{X_i = x_i\}_{i=1}^{N}$

Optimization Problem

likelihood function $L(\theta)$ is

$$L(\theta) = \prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j f_{X_i}(x_i | \mu_j, \sigma_j^2) = \prod_{i=1}^{N} \sum_{j=1}^{k} \pi_j f_j$$

↓ for notational convenience.

log likelihood will be

$$l(\theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{k} \pi_j f_j \right)$$

MLE $\hat{\theta}_{MLE} = \text{argmax } l(\theta)$ s.t. $\sum_{j=1}^{K} \pi_j = 1$

$$L(\theta, \lambda) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \pi_j f_j \right) + \lambda \left(1 - \sum_{j=1}^{K} \pi_j \right)$$

$$\frac{\partial L(\theta, \lambda)}{\partial \pi_j} = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \frac{f_j}{\sum_{j=1}^{K} \pi_j f_j} - \lambda = 0} \quad \text{— (1)}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \mu_j} = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \frac{\pi_j f_j}{\sum_{j=1}^{K} \pi_j f_j} \left(\frac{-1}{2} \cdot \frac{2(x_i - \mu_j)}{\sigma_j^2} \cdot (-1) \right) = 0}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \sigma_j} = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \frac{\pi_j}{\sum_{j=1}^{K} \pi_j f_j} \left(\frac{-1}{\sigma_j^2} f_j + f_j \cdot \left(\frac{-2}{\sigma_j^3} \right) \left(\frac{-1}{2} (x_i - \mu_j)^2 \right) \right) = 0} \quad \text{— (2)}$$

↳ (3)

Let's denote:

$$\frac{\pi_j f_j}{\sum_{k=1}^{K} \pi_k f_k} = \gamma_{ij}$$

then by ①, ② & ③ we have.

② $\Rightarrow \sum_{i=1}^{N} \gamma_{ij} (x_i - \mu_j) = 0 \Rightarrow$

$$\boxed{\hat{\mu}_j = \frac{\sum_{i=1}^{N} \gamma_{ij} x_i}{\sum_{i=1}^{N} \gamma_{ij}}}$$

③ $\Rightarrow \sum_{i=1}^{N} \gamma_{ij} \left(\frac{(x_i - \mu_j)^2}{\sigma_j^2} - 1 \right) = 0 \Rightarrow$

$$\boxed{\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N} \gamma_{ij} (x_i - \mu_j)^2}{\sum_{i=1}^{N} \gamma_{ij}}}$$

① $\Rightarrow \sum_{i=1}^{N} \frac{\gamma_{ij}}{\pi_j} - \lambda = 0 \Rightarrow \sum_{i=1}^{N} \gamma_{ij} - \lambda \pi_j = 0$

taking Sum over j on both sides given.

$$\sum_{j=1}^{K} \sum_{i=1}^{N} = \lambda \sum_{j=1}^{K} \pi_j = 0 \Rightarrow \sum_{i=1}^{N} \left(\sum_{j=1}^{K} \frac{\pi_j f_j}{\pi_1 f_1 + \dots + \pi_k f_k} \right)$$

$$= \lambda \cdot 1 = 0$$

$$\left(\because \sum_{j=1}^{K} \pi_j = 1 \right)$$

$$\Rightarrow \sum_{i=1}^{N} 1 - \lambda = 0 \Rightarrow \boxed{\lambda = N}$$

$$\Rightarrow \pi_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{N} \Rightarrow \boxed{\hat{\pi}_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{N}}$$

Init: Randomly choose $\theta^0 = (\pi_j^{(0)}, \mu_j^{(0)}, \sigma_j^{2(0)}) \quad \forall j = 1, 2, \ldots K.$

E-step:- $\gamma_{ij}^{(t+1)} = \frac{\pi_j f_j(x_i | \theta^{(t)})}{\pi_1 f_1 + \ldots + \pi_k f_k}.$

M-step $\pi_j^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ij}^{(t+1)}}{N}$ ||

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ij}^{(t+1)} x_i}{\sum_{i=1}^{N} \gamma_{ij}^{(t+1)}}$$

$$\sigma_j^{2(t+1)} = \sum_{i=1}^{N} \frac{\gamma_{ij}^{(t+1)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^{N} \gamma_{ij}^{(t+1)}}.$$

loop till : convergence.

(b) mixture of bernoulli

Let $X_i \sim \text{Bernoulli}(p)$. In this scenario our x_i can come from mixture model with (say) K mixture components so we have; conditional distribution

$$(X_i \mid Z_i = k) \sim \text{Bernoulli}(p_k)$$

where $Z_i \in \{1, 2, \ldots, K\}$

↳ latent variable

So marginal distribution of X_i is

$$P(X_i = x_i) = \sum_{j=1}^{K} P(X_i = x_i \mid Z_i = j) \cdot P(Z_i = j)$$

$$= \sum_{j=1}^{K} \pi_j \, P(X_i = x_i \mid Z_i = j)$$

Similarly joint probability ~~distribution~~ of observation $X_1, X_2, \ldots, X_N$ assuming iid sample is

$$P(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} \left(\sum_{j=1}^{K} \pi_j \, P(X_i = x_i \mid Z_i = j) \right)$$

for bernoulli distribution.

$$P(X_i = x_i \mid Z_i = j) = p_j^{x_i} (1 - p_j)^{(1 - x_i)}$$

Aim: find maximum likelihood estimators for π_j's, b_j's

given datadet $\sum_{i=1}^{N} x_i = x_i^2{}_{i=1}$

likelihood function $L(\theta)$ is

$$L(\theta) = \prod_{i=1}^{N} \left(\sum_{j=1}^{K} \pi_j b_j^{x_i} (1-b_j)^{R\ 1-x_i} \right)$$

log. likelihood will be:

$$\ell(\theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \pi_j b_j^{x_i} (1-b_j)^{1-x_i} \right)$$

$$\theta_{MLE} = \underset{\pi_j, b_j}{\operatorname{argmax}}\ \ell(\theta) \quad \text{s.t.} \quad \sum_{j=1}^{K} \pi_j = 1.$$

$$L(\theta, \lambda) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \pi_j b_j^{x_i} (1-b_j)^{1-x_i} \right) + \lambda \left(1 - \sum_{j=1}^{K} \pi_j \right) = 0$$

$$\frac{\partial L(\theta, \lambda)}{\partial \pi_j} = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \frac{f_j}{\pi_1 f_1 + \pi_2 f_2 + \ldots + \pi_K f_K} - \lambda = 0} \quad \text{--- (1)}$$

where $f_j := b_j^{x_i} (1-b_j)^{1-x_i}$

$$\frac{\partial L(\theta, \lambda)}{\partial b_j} = 0 \Rightarrow \sum \frac{\pi_j}{\pi_1 f_1 + \ldots + \pi_K f_K} \cdot \left(x_i b_j^{(x_i - 1)} (1-b_j)^{1-x_i} + (1-x_i) b_j^{x_i} (1-b_j)^{-x_i} \right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\pi_j \overbrace{b_j^{x_i}(1-b_j)^{1-x_i}}^{f_j}}{\pi_1 f_1 + \ldots + \pi_k f_k} \left\{ x_i b_j^{-1} - (1-x_i)(1-b_j)^{-1} \right\} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\pi_j f_j}{\pi_1 f_1 + \ldots + \pi_k f_k} \left\{ x_i \left(\frac{1-b_j}{b_j} \right) - (1-x_i) \right\} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \gamma_{ij} \left\{ x_i - b_j \right\} = 0$$

$$\boxed{\gamma_{ij} = \frac{\pi_j f_j}{\pi_1 f_1 + \pi_2 f_2 + \ldots + \pi_k f_k}}$$ then

$$\frac{\partial L(\theta, \lambda)}{\partial b_j} = 0 \Rightarrow \sum_{i=1}^{N} \gamma_{ij} \left\{ \frac{x_i}{b_j} - \cancel{x_i} - 1 + \cancel{x_i} \right\} = 0$$

$$\Rightarrow \boxed{\hat{b}_j = \frac{\sum_{i=1}^{N} \gamma_{ij} x_i}{\sum_{i=1}^{N} \gamma_{ij}}} \quad \text{--- (2)}$$

from (1) we also have

$$\sum_{i=1}^{N} \frac{f_j}{\pi_1 f_1 + \ldots + \pi_k f_k} - \lambda = 0 \Rightarrow \sum_{i=1}^{N} \frac{\gamma_{ij}}{\pi_j} - \lambda = 0$$

$$\Rightarrow \sum_{i=1}^{N} \gamma_{ij} - \lambda \pi_j = 0$$ taking sum over j on both sides gives

$$\sum_{j=1}^{K} \sum_{i=1}^{N} \gamma_{ij} - \lambda \sum_{j=1}^{K} \pi_j = 0 \Rightarrow \lambda = \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{\pi_j f_j}{\pi_1 f_1 + \ldots + \pi_k f_k}$$

$$\Rightarrow \boxed{\lambda = N}$$

So $\hat{\pi}_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{\lambda} \Rightarrow \boxed{\hat{\pi}_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{N}}$

EM Algorithm

Init: Randomly choose $\theta^0 = (\pi_j^{(0)}, b_j^{(0)}) \; \forall \; j = 1, 2, \ldots, k$

E-Step $\gamma_{ij}^{(t+1)} = \frac{\pi_j \, b_j^{x_i} \, (1-b_j)^{1-x_i}}{\sum_{l=1}^{k} \pi_l \, b_l^{x_i} \, (1-b_l)^{1-x_i}}$

M-Step $\pi_j^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ij}^{(t+1)}}{N}$

$b_j^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ij}^{(t+1)} x_i}{\sum_{i=1}^{N} \gamma_{ij}}$

loop till convergence

Q-7. Bernoulli distribution of $D = \{x_i\}_{i=1}^{N}$

$$f_X(x_i|b) = b^{x_i}(1-b)^{1-x_i}$$

given conjugate prior to be Beta distribution

i.e.

$$g(b) \sim \text{Beta}(\alpha, \beta)$$

$$\Leftrightarrow g(b) = \text{constant} \times b^{\alpha-1}(1-b)^{\beta-1}$$

by bayes rule

$$f(b|D) = \frac{f_X(D|b) \cdot g(b)}{f_X(D)}$$

$f_X(D)$ is independent of b. and is just a scaling constant.

MAP estimate of bernoulli distribution i.e Bernoulli (p ← probability of success)

$$b_{MAP} = \underset{b}{\text{argmax}}\, f(b|D)$$

$$= \underset{b}{\text{argmax}} \left(\prod_{i=1}^{N} f(x_i|b)\right) \cdot g(b)$$

taking log won't change argmax So,

$$b_{MAP} = \underset{b}{\text{argmax}} \sum_{i=1}^{N} \log(f(x_i|b)) + \log g(b)$$

$$\log \ell(p) = \sum_{i=1}^{N} \log\left(f(x_i|p)\right) + \log\left(g(p)\right)$$

$$= \sum_{i=1}^{N} \left\{ x_i \log p + (1-x_i)\log(1-p) \right\} +$$

$$\log(\text{constant}) + (\alpha-1)\log p + (\beta-1)\log(1-p)$$

Taking log ratio.

Take differentiation w.r.t. p on both side

$$\frac{\partial \ell(p)}{\partial p} = 0 \Rightarrow \sum_{i=1}^{N}\left(\frac{x_i}{p} - \frac{(1-x_i)}{1-p}\right) + \frac{\alpha-1}{p} - \frac{(\beta-1)}{1-p} = 0$$

On simplification we get:

$$\hat{p}_{MAP} = \frac{N + (\alpha-1) + (\beta-1)}{\sum_{i=1}^{N} x_i + (\alpha-1)} \quad \text{or} \quad p_{MAP} = \frac{\sum_{i=1}^{N} x_i + (\alpha-1)}{N + (\alpha-1) + (\beta-1)}$$

Ans

Become estimate of

Q-8 Multinomial distribution

multinomial distribution.

$D = \{X_i\}_{i=1}^{N}$ s.t $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$

$\sum_{j=1}^{K} x_{ij} = n$

$D = \{Y_i\}_{i=1}^{N}$ where $Y_i = \{X_j\}_{j=1}^{K}$ s.t $\sum_{i=1}^{K} x_i = n$ (say)

then probability mass function of multinomial distribution is

$f(x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)} \mid p_1, p_2, \ldots, p_k) = P(X_1^{(i)} = x_1^{(i)}, X_2^{(i)} = x_2^{(i)} \ldots X_k^{(n)} = x_k^{(i)})$

(this sample)

$= \frac{n!}{x_1! x_2! \ldots x_k!} \cdot p_1^{x_1^{(i)}} p_2^{x_2^{(i)}} \ldots p_k^{x_k^{(i)}}$

$= \text{Constant} \times p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$

(say c_i)

s.t $\sum x_i = n$, $x_i \geq 0$

Given conjugate prior is dirichlet distribution i.e

$g(p_1, p_2, \ldots p_k) \sim \text{dirichlet}(\alpha_1, \alpha_2 \ldots \alpha_k)$

$\Leftrightarrow g(p_1, p_2, \ldots, p_k) = (\text{Constant}) \times \prod_{i=1}^{K} p_i^{\alpha_i - 1}$

s.t $\sum_{i=1}^{K} p_i = 1$ & $p_i \geq 0$

$\forall i = 1, 2, \ldots k$

MAP estimate of multinomial distribution

Pb Let $\vec{p} = (p_1, p_2, \ldots, p_k)$ s.t. $\sum_{i=1}^{k} p_i = 1$

$$\vec{p}_{MAP} = \underset{\vec{p}}{\operatorname{argmax}}\ f(\vec{p}|D) \quad \text{s.t.} \sum_{i=1}^{k} p_i = 1,$$

by now we know that this is equivalent to

$$\vec{p}_{map} = \underset{p_1, p_2 \ldots p_k}{\operatorname{argmax}} \left(\sum_{i=1}^{N} \log \left(f(x_1^{(i)}, \ldots, x_k^{(i)} | p_1, \ldots p_k) + \log g(p_1, \ldots, p_k) \right) \right)$$

Using lagrange's multiplier

$$\Rightarrow \text{let } l(p_1, p_2 \ldots p_k) = \sum_{i=1}^{N} \log \left(f(x_1^{(i)}, \ldots x_k^{(i)} | p_1 \ldots p_k) \right) + \log \left(g(p_1 \ldots p_k) \right] + \lambda \left(1 - \sum_{j=1}^{k} p_j \right)$$

$$= \sum_{i=1}^{N} \log (C_i) + \sum_{i=1}^{N} \sum_{j=1}^{k} x_j^{(i)} \log p_j + \log(\text{constant}) + \sum_{j=1}^{k} (\alpha_j - 1) \log p_j + \lambda \left(1 - \sum_{j=1}^{k} p_j\right)$$

Now

$$\frac{\partial l}{\partial p_j} = 0 \Rightarrow \sum_{i=1}^{N} \frac{x_j^{(i)}}{p_j} + \frac{(\alpha_j - 1)}{p_j} - \lambda = 0$$

$$\Rightarrow \boxed{p_j = \frac{\sum_{i=1}^{N} x_j^{(i)} + (\alpha_j - 1)}{\lambda}} \quad \text{---(1)}$$

$$\therefore \lambda b_j = \sum_{i=1}^{m} x_j^{(i)} + (\alpha_j - 1)$$

take summation on both sides, w.r.t. j

$$\therefore \sum_{j=1}^{K} \lambda b_j = \sum_{j=1}^{K} \sum_{i=1}^{N} x_j^{(i)} + \sum_{j=1}^{K} (\alpha_j - 1)$$

$$\Rightarrow \lambda \left(\sum_{j=1}^{K} b_j \right) = \sum_{i=1}^{N} \left(\sum_{j=1}^{K} x_j^{(i)} \right) + \sum_{j=1}^{K} (\alpha_j - 1)$$

$$\Rightarrow \boxed{\lambda = mN + \sum_{j=1}^{K} (\alpha_j - 1)} \quad \text{--- (2)}$$

Substitute (2) in (1) we get

$$\hat{b}_{j_{MAP}} = \frac{\sum_{i=1}^{N} x_j^{(i)} + (\alpha_j - 1)}{mN + \sum_{k=1}^{K} (\alpha_k - 1)}$$

$\forall j = 1, 2, \ldots K$

$$\text{So } \vec{b}_{map} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_K) \quad \text{Ans}$$

is required map estimation of multinomial distribution.

Q-1 Generalized EM for MAP estimation.

Sol: The expectation-maximization algorithm, or EM algorithm is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables.

Let X → set of all observed variables.

Z → set of all hidden/latent variables.

θ → set of all unknown parameters

$D = \{x_i\}$, q ~ distribution of Z

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|D)$$ } using Bayes rule.

$$= \underset{\theta}{\operatorname{argmax}}\ \frac{p(\theta, D)}{p(D)} \cdot \log(\theta)$$ → prior distribution

$$= \underset{\theta}{\operatorname{argmax}} \log\left(\frac{p(\theta, D)\ \log(\theta)}{p(D)}\right)$$

($\because \log$ is increasing function).

$$= \underset{\theta}{\operatorname{argmax}} \log p(\theta, D) - \underset{\theta}{\operatorname{argmax}} \log(p(D))$$ → independent of θ

$$= \underset{\theta}{\operatorname{argmax}} \log(p(\theta, D))$$

$$= \underset{\theta}{\operatorname{argmax}} \log\left[p(D|\theta) \cdot p(\theta)\right]$$ → likelihood, → prior

$$= \underset{\theta}{\operatorname{argmax}} \left(\log(p(D|\theta) + \log(p(\theta))\right)$$

assuming N iid samples.

$$p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

$$\therefore \theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\theta) + \log(p(\theta))$$

$$\text{where } l(\theta) = \sum_{i=1}^{N} \log(p(x_i|\theta)) + \log(p(\theta))$$

$$= \sum_{i=1}^{N} \log\left(\sum_{c} P(x_i, z_i=c|\theta)\right) + \log(p(\theta))$$

(assume node z_i is hidden variable.
& $\sum_c P(x_i, z_i=c|\theta) = P(x_i|\theta)$)

$$= \sum_{i=1}^{N} \log\left(\sum_{c} q(z_i=c)\left\{\frac{P(x_i, z_i=c|\theta)}{q(z_i=c)}\right\}\right) + \log(p(\theta))$$

(where $\sum q(z_i=c)$ is a probability

s.t. $\sum_c q(z_i=c) = 1, q(z_i=c) \geq 0$)

$$\geq \sum_{i=1}^{N} \sum_{c} q(z_i=c) \log\left(\frac{P(x_i, z_i=c|\theta)}{q(z_i=c)}\right) + \log(p(\theta))$$

(Using Jensen's inequality. i.e $\log(E(\cdot)) \geq E(\log(\cdot))$]

$\Rightarrow l(\theta) \geq l(\theta, q)$ ↳ just a notation $\forall\, q$ of $q(z_i=c) \geq 0$ & $\sum_c q(z_i=c) = 1$

E-step. find q^* s.t.

$$l(\theta) = l(\theta, q^*)$$

* for any q, since

$$\because l(\theta) - (l(\theta, q)) = \left(\sum_{i=1}^{N} \log\left(p(x_i|\theta)\right) + \log(p(\theta)) \right)$$

$$- \left(\sum_{i=1}^{N} \sum_{c} q(z_i=c) \log \left(\frac{p(x_i, z_i=c|\theta)}{q(z_i=c)} \right) \right)$$

$$- \log(p(\theta))$$

$$= \sum_{i=1}^{N} \left(\sum_{c} q(z_i=c) \right) \log\left(p(x_i|\theta)\right) - \sum_{i=1}^{N} \sum_{c} q(z_i=c) \times \log\left(\frac{p(x_i, z_i=c)}{q(z_i=c)} \right)$$

($\sum_c q(z_i=c) = 1$)

$$= \sum_{i=1}^{N} \sum_{c} q(z_i=c) \log \left(\frac{p(x_i|\theta)\ q(z_i=c)}{p(x_i, z_i=c|\theta)} \right)$$

$$= \sum_{i=1}^{N} \left[\sum_{c} q(z_i=c) \log \left(\frac{p(x_i)\ q(z_i=c)}{p(z_i=c|x_i, \theta)} \right) \right]$$

↳ KL-divergence b/w $q(z_i=c)$ & $p(z_i|x_i, \theta)$

$$= \sum_{i=1}^{N} KL\left(q(z_i) \,||\, p(z_i|x_i, \theta) \right)$$

for for E-step find q^* s.t.

s.t. $l(\theta) - l(\theta, q^*) = 0$; where

$$\Rightarrow \sum_{i=1}^{N} KL(q^*(z_i) \| p(z_i | x_i, \theta)) = 0$$

but we know $KL(p \| q) \geq 0 \quad \forall p, q$

& $KL(p \| q) = 0$ iff $p = q$

$$\Rightarrow \boxed{q^*(z_i) := p(z_i | x_i, \theta)} \rightarrow \text{independent of prior } p(\theta)$$

So E-step is $q^*(z_i) = p(z_i | x_i, \theta)$

M-step find $\theta^* = \arg\max_{\theta} l(\theta, q^*)$

$$\Rightarrow \theta^* = \arg\max_{\theta} \left\{ \sum_{i=1}^{N} \sum_{c} q(z_i = c) \cdot \ln \left(\frac{p(x_i, z_i = c | \theta)}{q(z_i = c)} \right) + \log(p(\theta)) \right\}$$

$$\Rightarrow \theta^* = \arg\max_{\theta} \left\{ \sum_{i=1}^{N} \sum_{c} q(z_i = c) \ln(p(x_i, z_i = c) | \theta) + \ln(p(\theta)) \right\}$$

$(\because q(z_i = c)$ is independent of $\theta)$

NOTE: θ^* depends on prior As $p(\theta)$

* proof of correctness -

$$\ell(\theta^{(old)}) \leq \sum_{i=1}^{N} \sum_{c} q^*(z_i=c) \ln \frac{p(x_i, z_i=c)}{q^*(z_i=c)}$$

$$\therefore \ell(\theta^{(old)}) = \ell(\theta^{(old)}, q^*) + \sum_{i=1}^{N} KL(q^*(z_i) \| \underbrace{p(z_i|x_i, \theta^{(old)})}_{\downarrow 0})$$

$$\left(\because q^*(z_i) = p(z_i|x_i, \theta^{(old)})\right)$$

$$\Rightarrow \ell(\theta^{(old)}) = \ell(\theta^{(old)}, q^*)$$

$$= \sum_{i=1}^{N} \sum_{c} q^*(z_i=c) \log\left(\frac{p(x_i, z_i=c)}{q^*(z_i=c)}\right) + \log(p(\theta))$$

$$\ell(\theta^{(old)}) = \sum_{i=1}^{N} \sum_{c} q^*(z_i=c) \log(p(x_i, z_i=c)) + \log(p(\theta)) - \sum_{i=1}^{N} \sum_{c} \log(q^*(z_i=c))$$

$$\ell(\theta^{(new)}) = \ell(\theta^{(new)}, q^*) + \sum_{i=1}^{N} KL(q^*(z_i) \| \underbrace{p(z_i|x_i, \theta^{(new)})}_{\geq 0})$$

$$= \sum_{i=1}^{N} \sum_{c} q^*(z_i=c) \ln(p(x_i, z_i=c)) - \sum_{i=1}^{N} \sum_{c} \log(q^*(z_i=c)) + \log(p(\theta)) + KL(q^*(z_i) \| p(z_i|x_i, \theta^{(new)}))$$

$$\Rightarrow \ell(\theta^{(new)}) - \ell(\theta^{(old)}) = KL(q^*(z_i) \| p(z_i|x_i, \theta^{(new)})) \geq 0$$

$$\Rightarrow \ell(\theta^{(new)}) \geq \ell(\theta^{(old)}) \Rightarrow \text{H.P.}$$

CLASSMATE

Date

Page

END OF ASSIGNMENT

X X X