Machine Learning - Quiz 3 ①
Parametric Methods & Clustering

— Kinshuk Vasisht, 19
M.Sc., Semester 2

30/30

1. To find the maximum likelihood estimate of a parameter set θ for the given dataset $\mathcal{X}$, the steps are:

1) Define the likelihood for the ~~instance x^t~~ parameters θ given a single instance x^t, using the probability of the assumed distribution of x^t:

$$l(\theta | x^t) = p(x^t | \theta).$$

2) Define the likelihood for the parameters θ of the distribution of $\mathcal{X} = \{x^t\}_{t=1}^N$ given the entire dataset $\mathcal{X}$:

$$l(\theta | \mathcal{X}) = l(\theta | x^1, x^2 \cdots x^n) \quad — ①$$

If the instances are i.i.d., ① simplifies to:

$$l(\theta | \mathcal{X}) = \prod_{t=1}^{N} l(\theta | x^t)$$

3) Next, maximize the likelihood function wrt. different parameters in θ to find the MLE for θ.
Since computation using l is complex, this is equivalent to maximizing the log-likelihood of θ given $\mathcal{X}$, defined as:

$$\begin{aligned}\mathcal{L}(\theta | \mathcal{X}) &= \log_e l(\theta | \mathcal{X}) \\ &= \log_e \left(\prod_{t=1}^{N} l(\theta | x^t)\right) \\ &= \Sigma_{t=1}^{N} \log l(\theta | x^t) \\ &= \Sigma_{t=1}^{N} \log (p(x^t | \theta))\end{aligned}$$

4) Next, to maximize $\mathcal{L}$, differentiate partially wrt $\theta_i \ \forall \ \theta_i \in \theta$ & equate to 0 to find MLE for θ_i.
Repeat 4) for all $\theta_i \in \theta$.

Maximize $\mathcal{L} \Rightarrow \dfrac{\partial \mathcal{L}}{\partial \theta_i} = 0$

(2)

Date :
Page No.

Kinshuk Vasisht , 19

Parameter to estimate = λ.

2: Given distribution = $\lambda e^{-\lambda x} = p(x^t|\lambda)$

Now, likelihood of λ given an instance x^t

$$= \quad l(\lambda|x^t) = p(x^t|\lambda)$$
$$= \lambda e^{-\lambda x^t}$$

Now, likelihood for the entire dataset $\mathcal{X}$ (assuming x^t are i.i.d):

$$\Rightarrow \quad l(\lambda|\mathcal{X}) = \prod_{t=1}^{N} l(\lambda|x^t)$$
$$= \prod_{t=1}^{N} \lambda e^{-\lambda x^t}$$

Now, log-likelihood of $\theta(\lambda)$ given $\mathcal{X}$ is :

$$\mathcal{L}(\lambda|\mathcal{X}) = \log l(\lambda|\mathcal{X})$$
$$= \sum_{t=1}^{N} \log(l(\lambda|x^t))$$
$$\left(\because l(\lambda|\mathcal{X}) = \prod_{t=1}^{N} l(\lambda|x^t)\right)$$

$$= \sum_{t=1}^{N} \log(\lambda e^{-\lambda x^t})$$

$$\underline{\mathcal{L}(\lambda|\mathcal{X}) = \sum_{t=1}^{N} [\log\lambda + (-\lambda x^t)]}$$

Now, maximizing $\mathcal{L}$ implies differentiating wrt λ & equating the expression to 0.

Now, $$\frac{\partial \mathcal{L}(\lambda|\mathcal{X})}{\partial\lambda} = \frac{\partial}{\partial\lambda} \sum_{t=1}^{N} [\log\lambda - \lambda x^t]$$
$$= \sum_{t=1}^{N} \left[\frac{\partial \log\lambda}{\partial\lambda} - \frac{\partial}{\partial\lambda}\lambda x^t\right]$$
$$= \sum_{t=1}^{N} \left[\frac{1}{\lambda} - x^t \cancel{\frac{\partial\lambda}{\partial\lambda}}\right]$$
$$= \sum_{t=1}^{N} \left[\frac{1}{\lambda} - x^t\right]$$

Date :
Page No.

Kinshuk Vasisht, 19

Now, $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$

$$\Rightarrow \sum_{t=1}^{N} \left[\frac{1}{\lambda} - x^t \right] = 0$$

$$\Rightarrow \frac{\sum_{t=1}^{N} 1}{\lambda} - \sum_{t=1}^{N} x^t = 0 \quad (\because \lambda \text{ is independent of } t)$$

$$\Rightarrow \frac{N}{\lambda} - \sum_{t=1}^{N} x^t = 0 \quad \left(\because \sum_{t=1}^{N} 1 = N \text{ (N instances in } \mathcal{X}) \right)$$

$$\Rightarrow \frac{N}{\lambda} = \sum_{t=1}^{N} x^t$$

$$\Rightarrow \hat{\lambda}_{MLE} = \frac{N}{\sum_{t=1}^{N} x^t}$$

Also, we have, $E[x]$, where $x \sim \text{Exponential}(\lambda)$
$= 1/\lambda$

$$\therefore \frac{1}{\hat{\lambda}_{MLE}} = \frac{\sum_{t=1}^{N} x^t}{N}$$

$\therefore$ The maximum likelihood estimate for λ for a dataset $\mathcal{X}$ with i.i.d $x^t \in \mathcal{X}$ s.t. $x^t \sim \text{Exponential}(\lambda)$ is $\hat{\lambda}_{MLE} = \frac{N}{\sum_{t=1}^{N} x^t}$ (inverse of the average value of the given sample, consistent with the estimator for sample mean)

3\. a) $d(x) - \theta$ denotes the deviation between the value of the estimator d for the dataset x & the value of the parameter θ.

b) $E[d(x) - \theta]$ denotes the expected value of the deviation between the estimator d ~~over~~ all sample datasets x_i & the parameter θ. This is the bias associated with the estimator θ, defined as $b_\theta(d)$.

c) $E[d(x) - \theta]^2$ denoted the squared value of the bias associated with the estimator d, and is a

Kinshuk Vasisht, 19

component of the Mean squared error between the parameters θ & the estimator d, $\therefore$ MSE $= E[(d(x) - \theta)^2]$.

d) $E[(d(x) - E[d(x)])^2]$ is the expected value of the squared deviation of an estimator d from the expected value of the estimator over all sample datasets x_i. This represents the variance associated with the estimator & is a component in the mean square error between d & θ.

4

4. In the E step, we have :
$$Q(\phi/\phi^{l}) = E[\mathcal{L}_c(\phi/x,Z) \mid x, \phi^{l}]$$

- Here, $Q(\phi/\phi^{l})$ denotes the expectation of the complete likelihood function $\mathcal{L}_c$ conditioned over the dataset x & the current set of parameters (parameters for the current iteration, iteration l) ϕ^{l}.
- ϕ denotes the parameters to be estimated, which for a general mixture model include the priors of the hidden variables denoting cluster associations $P(G_i)$ & the distributional parameters of the assumed distribution of x^t in the cluster G_i, i.e. θ_i s.t $p(x^t|G^t) \sim \Delta(\theta_i)$, where Δ is any distribution assumed for $x^t|G^t$.
- ϕ^{l} is the value of the parameters from the parameter space for the current iteration l, which is used in the M-step to improve (maximize) Q & derive the next set of parameters ϕ^{l+1}.
- $\mathcal{L}_c(\phi/x,Z)$ is the complete likelihood function for the parameters ϕ given the dataset x & the set of hidden variables Z ($|Z| = |x|$)

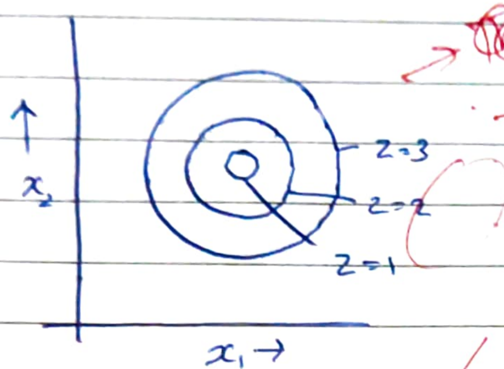
Kinshuk Vasisht, 19

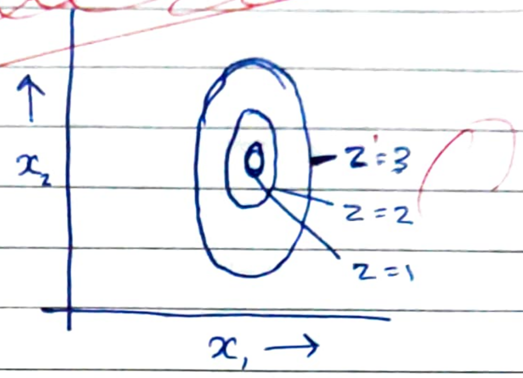
Date:

Page No.

assumed to be present but unknown for the dataset. Using the complete likelihood the ~~best~~ set of parameters ϕ can be estimated which maximizes the incomplete likelihood $L(\phi|x)$.

5.





a) $\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma^2$ (say)

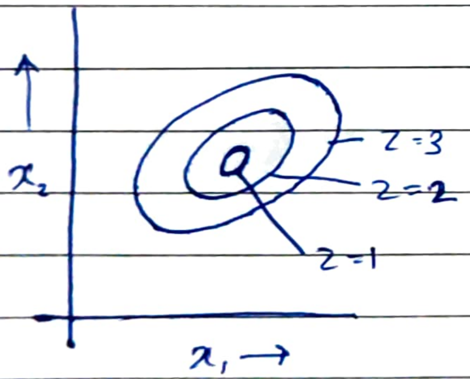
$\sigma_{x_1 x_2} = 0.$

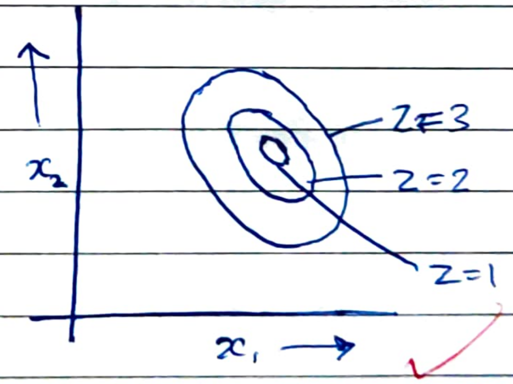
$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

b) $\sigma_{x_1}^2 \neq \sigma_{x_2}^2$ $(\sigma_{x_2}^2 > \sigma_{x_1}^2)$

$\sigma_{x_1 x_2} = 0.$

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & 0 \\ 0 & \sigma_{x_2}^2 \end{bmatrix}$$





$\sigma_{x_1}^2 \neq \sigma_{x_2}^2$

$\sigma_{x_1 x_2} > 0$

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix}$$

$\sigma_{x_1}^2 \neq \sigma_{x_2}^2$

$\sigma_{x_1 x_2} < 0$

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix}$$

10