ML Homework 4

Mehul Yesminkumar

1 Generative Models

1.1

$$P(X = x|\theta) = \begin{cases} 1/\theta & \text{if } x\epsilon(0,\theta) \\ \\ 0 & \text{otherwise} \end{cases} \implies \frac{1}{\theta}\mathbf{1}[0 \le x \le \theta]$$

$$Likelihood\ func\ is:\ \ L(x_1, ..., x_n|\theta) = \prod_{r=1}^{N} \frac{1}{\theta}\mathbf{1}[0 \le x_r \le \theta]$$

Likelihood function is inversely proportional to θ

To maximize likelihood we need smallest value of θ

$\theta^{MLE} = max(x_1, x_2 ... x_n)$

1.2.1

$$P(k|x_n, \theta_1, \theta_2, w_1, w_2) = \frac{w_k U(X = x|\theta_k)}{w_1 U(X = x|\theta_1) + w_2 U(X = x|\theta_2)}$$

1.2.2

$$Q(\theta, \theta_{old}) = \sum_n \sum_k P(k|x_n, \theta_1^{old}, \theta_2^{old}, w_1^{old}, w_2^{old}) log(P(x_n, k|\theta_1, \theta_2, w_1, w_2)$$

$$\implies \sum_n \sum_{k=1,2} \frac{w_k^{old} U(X = x/\theta_k^{old})}{\sum_{d=1}^k w_d^{old} \frac{1}{\theta_d^{old}} \mathbf{1}[0 \leq x \leq \theta_d^{old}]]} log(w_k U[x = x_n|\theta_k])$$

$$\implies \sum_n \sum_k P_{old}(k/x_n).log(w_k U[x = x_n|\theta_k])$$

1.2.3

$$\theta^{new} = argmax_\theta A(\theta, \theta^{old})$$

$$\theta_1 = argmax_\theta \sum_n P_{old}(k = 1|x_n) * log[\frac{1}{\theta_1} I[0 < x_n \leq \theta_1] * w_1]$$

$$\theta_2 = argmax_\theta \sum_n P_{old}(k = 2|x_n) * log[\frac{1}{\theta_2} I[0 < x_n \leq \theta_2] * w_2]$$

We can remove the points not in distribution for θ_1^{old}

$$\implies P_{old}(k/x_n) \ would \ turn \ out \ to \ be \ 0$$

$$\theta_1 = max(x_n) \qquad \forall x \epsilon [0 < x_n \leq \theta_1^{old}]$$

$$\theta_1^{new} = max(x_1, x_2 ... x_N) \quad \forall n = 1, 2 ... N \ such \ that \ 0 \leq x_n \leq \theta_1^{old}$$

$$\theta_2^{new} = max(x_1, x_2 ... x_N) \quad \forall n = 1, 2, ..N$$

2 Mixture density models

$$
\begin{aligned}
P(x_b|x_a) &= \frac{P(x_a, x_b)}{P(x_a)} \quad (conditional\, probability) \\
&= \frac{\sum_{k=1}^{K} \pi_k P(x_b, x_a|k)}{\sum_{k=1}^{K} \pi_k P(x_a|k)} \qquad as\ P(x) = \sum_{k=1}^{K} \pi_k P(x|k)
\end{aligned}
$$

$$
\begin{aligned}
P(x_b|x_a) &= \frac{\sum_{k=1}^{K} \pi_k P(x_b|x_a, k) P(x_a|k)}{\sum_{k=1}^{K} \pi_k P(x_a|k)} \\
P(x_b|x_a) &= \frac{\sum_{k=1}^{K} \pi_k P(x_a|k) P(x_b|x_a, k)}{\sum_{k=1}^{K} \pi_k P(x_a|k)} \\
P(x_b|x_a) &= \sum_{k=1}^{K} \frac{\pi_k P(x_a|k)}{\sum_{k=1}^{K} \pi_k P(x_a|k)} P(x_b|x_a, k)
\end{aligned}
$$

$$
\begin{aligned}
\lambda_k = \frac{\pi_k P(x_a|k)}{\sum_{k=1}^{K} \pi_k P(x_a|k)} &\implies \frac{\pi_k P(x_a|k)}{P(x_a)} \\
Verifying \sum_{k=1}^{K} \lambda_k = 1 &\implies \sum_{k=1}^{K} \lambda_k = \sum_{k=1}^{K} \frac{\pi_k P(x_a|k)}{P(x_a)} \\
&\implies \frac{P(x_a)}{P(x_a)} \sum_{k=1}^{K} \pi_k = 1
\end{aligned}
$$

3 GMM and K-means

$$
\gamma(z_{nk}) = \frac{\pi_k e^{-\dfrac{||x_n - \mu_k||_2^2}{2\sigma^2}}}{\sum_j \pi_j e^{-\dfrac{||x_n - \mu_j||_2^2}{2\sigma^2}}}
$$

As $\sigma \to 0$, $\gamma(z_{nk})$ will go to zero except for term j.For term j, $\gamma(z_{nj})$ will go to 1. So $\gamma(z_{nk})$

$$
\gamma(z_{nk}) = \begin{cases} 1, & k = arg\min_k ||x_n - \mu_k||^2 \\ 0, & \text{otherwise.} \end{cases} \qquad \text{As } \sigma \to 0 \implies \gamma(z_{nk}) = r_{nk}
$$

$$Given\ G = \sum_n^N \sum_k^K \gamma(z_{nk})[log\pi_k + log(N(x_n|\mu_k, \sigma_2 I))]$$

$$logN(x_n|\mu_k, \sigma^2 I) = log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{-||x_n - \mu_k||^2}{2\sigma^2})\Big)$$

$$\implies G = \sum_n^N \sum_k^K \gamma(z_{nk})[log\pi_k + (-\frac{1}{2}log(2\pi) - log(\sigma) - \frac{1}{2}\frac{||x_n - \mu_k||^2}{\sigma^2})]$$

As $\sigma \to 0$, multiplying equation with σ_2 we get $G = -\frac{1}{2}\sum_n^N \sum_k^K r_{nk}||x_n - \mu_k||_2^2 + C$

Hence proved that maximizing the log likelihood for GMM is equivalent

to minimizing the distortion measure J for the K-means algorithm.

4 Naive Bayes

4.1

$$P(X = x, Y = c) = P(Y = c)P(X = x|Y = c)$$

$$log(LL) = log\prod_{n=1}^{N} P(X = x, Y = c) \qquad (log - likelihood\, of\, above)$$

$$= log\prod_{n=1}^{N}(\pi_c \prod_{d=1}^{D} P(X_d = x_d|Y = c))$$

$$Given\ P(X_d = x_d|Y = c; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma_{cd}^2}} exp\Big(\frac{-(x_d - \mu_{cd})^2}{2\sigma_{cd}^2}\Big)$$

$$\implies \sum_n \Big(log\pi_{y_n} + \sum_{d=1}^{D} log\Big(\frac{1}{\sqrt{2\pi\sigma_{y_n d}^2}} exp\Big(\frac{-(x_d - \mu_{y_n d})^2}{2\sigma_{y_n d}^2}\Big)\Big)\Big)$$

$$\implies \sum_n log\pi_{y_n} + \sum_n \sum_{d=1}^{D} log\Big(\frac{1}{\sqrt{2\pi\sigma_{y_n d}^2}} exp\Big(\frac{-(x_d - \mu_{y_n d})^2}{2\sigma_{y_n d}^2}\Big)\Big)$$

$$Log\, Likelihood = \sum_n log\pi_{y_n} + \sum_{n,d} \frac{-1}{2}log(2\pi\sigma_{y_n d}^2) - \frac{(x_d - \mu_{y_n d})^2}{2\sigma_{y_n d^2}}$$

4.2

$$L = \sum_{n} log\pi_{y_n} + \sum_{n,d} \frac{-1}{2} log(2\pi\sigma_{y_n d}^2) - \frac{(x_d - \mu_{y_n d})^2}{2\sigma_{y_n d^2}}$$

derivative w.r.t σ_{cd}^2 $$\frac{\partial(L)}{\partial\sigma_{cd}^2} = -\sum_{n:y_n=c}^{N} \frac{2\pi}{4\pi\sigma_{cd}^2} + \sum_{n:y_n=c}^{N} \frac{(x_d - \mu_{cd})^2}{2\sigma_{cd}^4} = 0$$

$$\implies \sigma_{cd}^2 = \frac{\sum_{n:y_n=c}^{N}(x_{nd} - \mu_{cd})^2}{N_c} \qquad N_c = no\,of\,datapoints\,labeled\,as\,c$$

derivative w.r.t μ_{cd} $$\frac{\partial(L)}{\partial\mu_{cd}} = \sum_{n:y_n=c}^{N} \frac{2(x_{nd} - \mu_{cd})}{2\sigma_{cd}^2} = 0$$

$$\implies \mu_{cd} = \frac{\sum_{n:y_n=c}^{N} x_{nd}}{N_c} \qquad N_c = no\,of\,datapoints\,labeled\,as\,c$$

derivative w.r.t π_c $$\frac{\partial(L)}{\partial\pi_c} = \sum_{n:y_n=c} \frac{1}{\pi_c} + \lambda = 0 \quad Using\ Lagrange\ Theorem$$

$$\implies \sum_{c}^{C} \pi_c = \sum_{c}^{C} \frac{-N_c}{\lambda} \qquad \lambda = -N$$

$$\implies \pi_c = \frac{N_c}{N} \qquad N_c = no\,of\,datapoints\,labeled\,as\,c$$

Hence we get : $\pi_c = \frac{N_c}{N}; \quad \sigma_{cd}^2 = \frac{\sum_{n:y_n=c}^{N}(x_{nd} - \mu_{cd})^2}{N_c}; \quad \mu_{cd} = \frac{\sum_{n:y_n=c}^{N} x_{nd}}{N_c}$