CSCI 567 Homework 4

Shamim Samadi - USC ID: 5188327056

October 29, 2017

Problem 1. Generative models

- **Question 1.1**

$$L(\theta; x_1, x_2, ..., x_n) = logP(X = x_1, ..., x_n|\theta) = log(P(x_1|\theta)P(x_2|\theta)...P(x_n|\theta)) \tag{1}$$

Thus

$$L(\theta; x_1, x_2, ..., x_n) = log(\frac{1}{\theta})^n 1[0 < x_1 \leq \theta, ..., 0 < x_n \leq \theta] = nlog(\frac{1}{\theta})1[0 < x_1 \leq \theta, ..., 0 < x_n \leq \theta] \tag{2}$$

From $0 < x_1 \leq \theta, 0 < x_2 \leq \theta, ..., 0 < x_n \leq \theta$, we get: $\theta \geq max(x_1, x_2, ..., x_n)$. Thus:

$$\hat{\theta}_{ML} = argmax_\theta\{nlog(\frac{1}{\theta})\} = max(x_1, x_2, ..., x_n) \tag{3}$$

- **Question 1.2**

• first part:

$$P(k|x_n, \theta_1, \theta_2, w_1, w_2) = \frac{P(x_n, \theta_1, \theta_2, w_1, w_2|k)P(k)}{P(x_n, \theta_1, \theta_2, w_1, w_2)} \tag{4}$$

$$P(k|x_n, \theta_1, \theta_2, w_1, w_2) = \frac{w_k \frac{1}{\theta_k} 1[0 < x_n \leq \theta_k]}{\sum_{j=1}^{2} w_j \frac{1}{\theta_j} 1[0 < x_n \leq \theta_j]} \tag{5}$$

• second part - assuming $\theta = w_1, w_2, \theta_1, \theta_2$,

$$Q(\theta, \theta^{OLD}) = \sum_{n=1}^{N} \sum_{k=1}^{2} q(z_n = k) logP(x_n, z_n = k|\theta) \tag{6}$$

$$q(z_n = k) = P(z_n = k|\theta^{OLD}) = P(z_n = k|{w_1}^{OLD}, {w_2}^{OLD}, {\theta_1}^{OLD}, {\theta_2}^{OLD}) \tag{7}$$

We also have:

$$logP(x_n, z_n = k|\theta) = log(P(x_n|z_n = k, \theta_k)P(z_n = k|\theta_k)) = logP(x_n|z_n = k, \theta_k) + logP(z_n = k|\theta_k) \tag{8}$$

$$logP(x_n, z_n = k|\theta) = logU(x_n|z_n = k, \theta_k) + logw_k = logw_k + log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k] \quad (9)$$

From 6, and using 5, 7 and 9, we get:

$$Q(\theta, \theta^{OLD}) = \sum_{n=1}^{N}\sum_{k=1}^{2}(\frac{{w_k}^{OLD}\frac{1}{{\theta_k}^{OLD}}1[0 < x_n \leq {\theta_k}^{OLD}]}{\sum_{j=1}^{2}{w_j}^{OLD}\frac{1}{{\theta_j}^{OLD}}1[0 < x_n \leq \theta_j]})(logw_k + log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]) \quad (10)$$

- third part - defining $P_{OLD}(k|x_n) = \frac{{w_k}^{OLD}\frac{1}{{\theta_k}^{OLD}}1[0<x_n\leq{\theta_k}^{OLD}]}{\sum_{j=1}^{2}{w_j}^{OLD}\frac{1}{{\theta_j}^{OLD}}1[0<x_n\leq\theta_j]}$, we have:

$$Q(\theta, \theta^{OLD}) = \sum_{n=1}^{N}\sum_{k=1}^{2}P_{OLD}(k|x_n)(logw_k + log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]) \quad (11)$$

In the M-step, we follow the following update rule: $\theta^{new} \leftarrow argmax_\theta\{Q(\theta, \theta_{OLD})\}$

$$Q(\theta, \theta^{OLD}) = \sum_{n=1}^{N}\sum_{k=1}^{2}P_{OLD}(k|x_n)logw_k + \sum_{n=1}^{N}\sum_{k=1}^{2}P_{OLD}(k|x_n)log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]) \quad (12)$$

which means w_k and θ_k can be optimized separately. Also, term $P_{OLD}(k|x_n)$ is independent of optimization parameters and can be treated as a constant. Thus

- ${\theta_k}^*$:

$argmax_{\theta_k}\{\sum_{n=1}^{N}\sum_{k=1}^{2}P_{OLD}(k|x_n)log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]\} =$
$argmax_{\theta_k}\{\sum_{n=1}^{N}\sum_{k=1}^{2}log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]\} =$
$argmax_{\theta_k}\{\sum_{n=1}^{N}\sum_{k=1}^{2}\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]\} =$
$argmin_{\theta_k}\{\sum_{n=1}^{N}\sum_{k=1}^{2}\theta_k 1[0 < x_n \leq \theta_k]\}$

Using ${\theta_2}^{OLD} \geq max(x_1, ..., x_n)$ and ${\theta_2}^{OLD} \geq min(x_1, ..., x_n)$, we get:

$argmax_{\theta_1}\{\sum_{n=1}^{N}\sum_{k=1}^{2}P_{OLD}(k|x_n)log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]\} = min(x_1, ..., x_n)$
which leads to: ${\theta_1}^{t+1} \leftarrow min(x_1, ..., x_n)$

$argmax_{\theta_2}\{\sum_{n=1}^{N}\sum_{k=1}^{2}P_{OLD}(k|x_n)log\frac{1}{\theta_k}1[0 < x_n \leq \theta_k]\} = max(x_1, ..., x_n)$
which leads to: ${\theta_2}^{t+1} \leftarrow max(x_1, ..., x_n)$

- ${w_k}^*$: $argmax\sum_n\sum_k log(w_k)$

we know that: $0 < w_k \leq 1$, thus: ${w_1}^{t+1} \leftarrow 1$ and ${w_2}^{t+1} \leftarrow 1$.

Problem 2. Mixture density models

- **Question 2.1**

$$P(x) = \sum_{k=1}^{K}\pi_k P(x|k) \quad (13)$$

$$P(x_a, x_b) = \sum_{k=1}^{K} \pi_k P(x_a, x_b|k) \tag{14}$$

$$P(x_b|x_a)P(x_a) = \sum_{k=1}^{K} \pi_k P(x_a, x_b|k) \tag{15}$$

$$P(x_b|x_a) = \frac{1}{P(x_a)} \sum_{k=1}^{K} \pi_k P(x_a, x_b|k) \tag{16}$$

$$P(x_b|x_a) = \frac{1}{P(x_a)} \sum_{k=1}^{K} \pi_k P(x_b|x_a, k)P(x_a|k) \tag{17}$$

$$P(x_b|x_a) = \sum_{k=1}^{K} (\frac{\pi_k}{P(x_a)} P(x_a|k))P(x_b|x_a, k) \tag{18}$$

thus:

$$\lambda_k = \frac{\pi_k}{P(x_a)} P(x_a|k) \tag{19}$$

Problem 3. The connection between GMM and K-means

- **Question 3.1**

$$\gamma(z_{nk}) = \frac{\pi_k exp(\frac{-||x_n - \mu_k||^2}{2\sigma^2})}{\sum_j \pi_j exp(\frac{-||x_n - \mu_j||^2}{2\sigma^2})} \tag{20}$$

when $\sigma \to 0$: $exp(\frac{-||x_n-\mu_k||^2}{2\sigma^2}) \to \frac{-||x_n-\mu_k||^2}{2\sigma^2}$, and:

$$\gamma(z_n k) \to \frac{\pi_k \frac{-||x_n-\mu_k||^2}{2\sigma^2}}{\sum_j \pi_j \frac{-||x_n-\mu_j||^2}{2\sigma^2}} = \frac{\pi_k(-||x_n - \mu_k||^2)}{\sum_j \pi_j(-||x_n - \mu_j||^2)} \tag{21}$$

the dominant term in the denominator is: $argmax(-||x_n - \mu_j||^2) = argmin(||x_n - \mu_j||^2)$. Thus:

$$\gamma(z_{nk}) = \begin{cases} \frac{\pi_k(-||x_n-\mu_k||^2)}{\pi_k(-||x_n-\mu_k||^2)} = 1, & k = argmin(||x_n - \mu_j||^2) \\ 0, & O.W. \end{cases} = r_{nk}$$

now we only need to show: $log\pi_k + logN(x_n|\mu_k, \sigma^2 I) \to c||x_n - \mu_k||^2$ as $\sigma \to 0$ (c is a constant).

$$log\pi_k + logN(x_n|\mu_k, \sigma^2 I) = log\pi_k + log(\frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-||x_n - \mu_k||^2}{2\sigma^2})) \tag{22}$$

$$log\pi_k + logN(x_n|\mu_k, \sigma^2 I) = log\pi_k + log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{||x_n - \mu_k||^2}{2\sigma^2} = -\frac{||x_n - \mu_k||^2}{2\sigma^2} + const. \tag{23}$$

Ignoring the constant terms, maximizing the expected complete-data log-likelihood (in GMM) comes down to maximizing the term $-\frac{||x_n-\mu_k||^2}{2\sigma^2}$, and since $2\sigma^2$ is the same for all classes, this maximization is equivalent to maximizing the term $-||x_n - \mu_k||^2$, which is equivalent to minimizing $||x_n - \mu_k||^2$, as in the K-means algorithm.

Problem 4. Naive Bayes

- **Question 4.1**

$$logP(D|\theta) = logP(X = x_1, x_2, ..., x_N; Y = y_1, y_2, ..., y_N|\pi_c, \mu_{cd}, \sigma_{cd}) \tag{24}$$

since observations are i.i.d, we get:

$$L = \sum_{n=1}^{N}[logP(y_n = c|\pi_c) + \sum_{d=1}^{D} logP(x_{nd}|y_n = c, \mu_{cd}, \sigma_{cd})] \tag{25}$$

$$L = \sum_{n=1}^{N}[log\pi_{y_n} + \sum_{d=1}^{D} log(\frac{1}{\sqrt{2\pi\sigma_{nd}{}^2}} exp(\frac{-(x_{nd} - \mu_{nd})^2}{2\sigma_{nd}{}^2}))] \tag{26}$$

$$L = \sum_{n=1}^{N} log\pi_{y_n} + \sum_{n=1}^{N}\sum_{d=1}^{D} log(\frac{1}{\sqrt{2\pi\sigma_{nd}{}^2}} exp(\frac{-(x_{nd} - \mu_{nd})^2}{2\sigma_{nd}{}^2})) \tag{27}$$

$$L = \sum_{n:y_n=c}^{N} log\pi_c + \sum_{c=1}^{C}\sum_{n:y_n=c,d} log(\frac{1}{\sqrt{2\pi\sigma_{cd}{}^2}} exp(\frac{-(x_{nd} - \mu_{cd})^2}{2\sigma_{cd}{}^2})) \tag{28}$$

$$L = \sum_{n:y_n=c}^{N} log\pi_c - \sum_{c=1}^{C}\sum_{n:y_n=c,d} [\frac{1}{2} log(2\pi\sigma_{cd}{}^2) + \frac{(x_{nd} - \mu_{cd})^2}{2\sigma_{cd}{}^2}] \tag{29}$$

- **Question 4.2**

$$L(\pi_c) = \sum_{n:y_n=c} log\pi_c \quad s.t. \sum_{c=1}^{C} \pi_c = 1 \tag{30}$$

Lagrange multipliers:

$$L(\pi_c, \lambda) = \sum_{n:y_n=c} log\pi_c - \lambda(\sum_{c=1}^{C} \pi_c - 1) \tag{31}$$

$$\frac{\partial L}{\partial \pi_c} = \sum_{n:y_n=c} \frac{1}{\pi_c} - \lambda = 0 \rightarrow \pi_c = \frac{1}{\lambda} \sum_{n:y_n=c} 1 \tag{32}$$

$$\frac{\partial L}{\partial \lambda} = 0 \rightarrow \sum_{c} \pi_c = 1 \tag{33}$$

Using 32 and 33, we get:

$$\sum_{c} \frac{1}{\lambda} \sum_{n:y_n=c} 1 = 1 \rightarrow \frac{1}{\lambda} \sum_{c} \sum_{n:y_n=c} 1 = 1 \rightarrow \lambda = \sum_{c} \sum_{n:y_n=c} 1 \tag{34}$$

now again using 32, we get:

$$\pi_c{}^* = \frac{\sum_{n:y_n=c} 1}{\sum_c \sum_{n:y_n=c} 1} = \frac{N_c}{N} \tag{35}$$

where N_c is the number of data points labeled as c.

Now,

$$L(\mu_{cd}, \sigma_{cd}) = -\sum_{c=1}^{C} \sum_{n:y_n=c,d} [\frac{1}{2} log(2\pi {\sigma_{cd}}^2) + \frac{(x_{nd} - \mu_{cd})^2}{2{\sigma_{cd}}^2}] \tag{36}$$

μ_{cd} and σ_{cd} can be estimated separately for each class c:

$$L_c(\mu_{cd}, \sigma_{cd}) = -\sum_{n:y_n=c,d} [\frac{1}{2} log(2\pi {\sigma_{cd}}^2) + \frac{(x_{nd} - \mu_{cd})^2}{2{\sigma_{cd}}^2}] \tag{37}$$

$$\frac{\partial L_c}{\partial \mu_{cd}} = \sum_{n:y_n=c,d} \frac{2(x_{nd} - \mu_{cd})}{2{\sigma_{cd}}^2} = 0 \rightarrow \sum_{n:y_n=c,d} x_{nd} - \mu_{cd} = 0 \tag{38}$$

defining the number of data points with label c as N_c $(N_c = \#(n : y_n = c))$:

$${\mu_{cd}}^* = \frac{\sum_{n:y_n=c,d} x_{nd}}{N_c} \tag{39}$$

Also

$$\frac{\partial L_c}{\partial \sigma_{cd}} = 0 \rightarrow \sum_{n:y_n=c,d} \frac{1}{\sigma_{cd}} (\frac{(x_{nd} - \mu_{cd})^2}{{\sigma_{cd}}^3}) = 0 \rightarrow \sum_{n:y_n=c,d} [{\sigma_{cd}}^2 - (x_{nd} - \mu_{cd})^2] = 0 \tag{40}$$

$$\rightarrow N_c {\sigma_{cd}}^2 - \sum_{n:y_n=c,d} (x_{nd} - \mu_{cd})^2 = 0 \tag{41}$$

$${\sigma_{cd}}^{2*} = \frac{\sum_{n:y_n=c,d} (x_{nd} - \mu_{cd})^2}{N_c} \tag{42}$$