Ahmad Fallahpour

CSCI567 2017 Homework Assignment 4 10/29/2017

1 Generative models

1.1

Likelihood of the observations:

$$L(D) = \prod_{i=1}^{N} P(x_i;\theta) \overset{iid}{=} \frac{1}{\theta^N}\mathbf{1}[0 < x_i \leq \theta] \tag{1}$$

Maximum likelihood:

$$\theta^* = \max_i x_i \tag{2}$$

If θ is smaller than even one of x_i, eq.(1) would be zero. So, θ should be greater than the greatest x_i. On the other hand, eq.(1) is a decreasing function. So, it is maximized when θ has its smallest value which is indicated in eq.(2).

1.2

$$P(k|x_n,\theta_1,\theta_2,\omega_1,\omega_2) = \frac{P(x_n|k,\theta_1,\theta_2,\omega_1,\omega_2)P(k|\theta_1,\theta_2,\omega_1,\omega_2)}{P(x_n|\theta_1,\theta_2,\omega_1,\omega_2)} = \frac{\omega_k U(x_n|\theta_k)}{\sum_{k'=1}^{2}\omega_{k'}U(x_n|\theta_{k'})} \tag{3}$$

expected complete-data log-likelihood:

$$Q(\theta,\theta^{OLD}) = \sum_{n=1}^{N}\sum_{k=1}^{2} P(k|x_n,\theta_1^{OLD},\theta_2^{OLD},\omega_1^{OLD},\omega_2^{OLD}) log P(x_n,k|\theta_1,\theta_2,\omega_1,\omega_2) \tag{4}$$

$$Q(\theta,\theta^{OLD}) = \sum_{n=1}^{N}\sum_{k=1}^{2} \frac{\omega_k^{OLD}U(x_n|\theta_k^{OLD})}{\omega_1^{OLD}U(x_n|\theta_1^{OLD}) + \omega_2^{OLD}U(x_n|\theta_2^{OLD})} log(\omega_k U(x_n|\theta_k)) \tag{5}$$

M-step:

$$\theta^{NEW} \leftarrow argmax Q(\theta,\theta^{OLD}) \tag{6}$$

$$\theta_1^{NEW},\theta_2^{NEW},\omega_1^{NEW},\omega_2^{NEW} \leftarrow \underset{\theta_1,\theta_2,\omega_1,\omega_2}{argmax} \sum_{n=1}^{N}\sum_{k=1}^{2} P_{OLD}(k|x_n) log(\omega_k U(x_n|\theta_k)) \tag{7}$$

Similar to what is explained in part 1.1, the function is decreasing and both θ_1^{NEW} and θ_1^{NEW} we should be greater than the greatest x_i. So:

$$\theta_1^{NEW} = \theta_2^{NEW} = \underset{i}{max} x_i \tag{8}$$

2 Mixture density models

2.1

$$P(x) = \sum_{k=1}^{K} \pi_k P(x|k) \tag{9}$$

$$P(x_a, x_b) = \sum_{k=1}^{K} \pi_k P(x_a, x_b|k) \tag{10}$$

$$P(x_a)P(x_b|x_a) = \sum_{k=1}^{K} \pi_k P(x_a|k)P(x_b|x_a, k) \tag{11}$$

$$P(x_b|x_a) = \sum_{k=1}^{K} \frac{\pi_k P(x_a|k)}{P(x_a)} P(x_b|x_a, k) \tag{12}$$

$$P(x_b|x_a) = \sum_{k=1}^{K} \lambda_k P(x_b|x_a, k) \tag{13}$$

$$\lambda_k = \frac{\pi_k P(x_a|k)}{P(x_a)} = \frac{\pi_k P(x_a|k)}{\sum_{k'=1}^{K} \pi_{k'} P(x_a|k')} \tag{14}$$

Form eq. (14) we can easily verify that:

$$\lambda_k \geq 0, \sum_{k=1}^{K} \lambda_k = 1 \tag{15}$$

3 The connection between GMM and K-means

3.1

$$\gamma(z_{nk}) = \frac{\pi_k exp(-\|x_n - \mu_k\|^2/2\sigma^2)}{\sum_j \pi_j exp(-\|x_n - \mu_j\|^2/2\sigma^2)} \tag{16}$$

We can rewite eq.(16) as follow:

$$\gamma(z_{nk}) = \frac{\pi_k}{\pi_k + \sum_{j\neq k} \pi_j exp((\|x_n - \mu_k\|^2 - \|x_n - \mu_j\|^2)/2\sigma^2)} \tag{17}$$

When $\sigma \to 0$, the denominator of eq.(17) can goes to π_k or ∞ which means $\gamma(z_{nk})$ can be 1 or 0.

$$if k = arg\min_{k'} \|x_n - \mu_{k'}\|^2 \implies (\|x_n - \mu_k\|^2 - \|x_n - \mu_j\|^2) < 0, \forall j \neq k \tag{18}$$

$$So.if \sigma \to 0, then \sum_{j\neq k} \pi_j exp((\|x_n - \mu_k\|^2 - \|x_n - \mu_j\|^2)/2\sigma^2) \to 0 \implies \gamma(z_{nk}) = 1 \tag{19}$$

$$if k \neq arg\min_{k'} \|x_n - \mu_{k'}\|^2 \implies \exists j, (\|x_n - \mu_k\|^2 - \|x_n - \mu_j\|^2) > 0 \tag{20}$$

$$So.if \sigma \to 0, then \sum_{j\neq k} \pi_j exp((\|x_n - \mu_k\|^2 - \|x_n - \mu_j\|^2)/2\sigma^2) \to \infty \implies \gamma(z_{nk}) = 0 \tag{21}$$

Therefore, we proved $\gamma(z_{nk}) = r_{nk}$ if $\sigma \to 0$

Now, we want to maximize following:

$$\underset{\mu_k}{maximize} \sum_n^N \sum_k^K \gamma(z_{nk})[log\pi_k + log\aleph(x_n|\mu_k, \sigma^2\mathbf{I})] \tag{22}$$

$$\underset{\mu_k}{maximize} \sum_n^N \sum_k^K \gamma(z_{nk})log(\frac{exp(-\|x_n - \mu_k\|^2)}{(2\pi\sigma^2)^{N/2}}) \tag{23}$$

$$\underset{\mu_k}{maximize} \sum_n^N \sum_k^K \gamma(z_{nk})[log(exp(-\|x_n - \mu_k\|^2) - log(2\pi\sigma^2)^{N/2}] \tag{24}$$

$$\underset{\mu_k}{maximize} \sum_n^N \sum_k^K \gamma(z_{nk})[-\|x_n - \mu_k\|^2] \tag{25}$$

$$\underset{\mu_k}{maximize} - \sum_n^N \sum_k^K \gamma(z_{nk})(-\|x_n - \mu_k\|^2) \Leftrightarrow \underset{\mu_k}{minimize} \sum_n^N \sum_k^K \gamma(z_{nk})\|x_n - \mu_k\|^2 = J \tag{26}$$

4 Naive Bayes

4.1

$$\ell = logP(D) = log \prod_{n=1}^N P(X = x_n, Y = y_n) = log \prod_{n=1}^N [P(Y = y_n)P(X = x_n|Y = y_n)] \tag{27}$$

$$\ell = \sum_{n=1}^N [log\big(P(Y = y_n) \prod_{d=1}^D P(X = x_{nd}|Y = y_n)\big)] \tag{28}$$

$$\ell = \sum_{n=1}^N [logP(Y = y_n) + log\big(\prod_{d=1}^D P(X = x_{nd}|Y = y_n)\big)] \tag{29}$$

$$\ell = \sum_{n=1}^N logP(Y = y_n) + \sum_{n=1}^N log\Big(\prod_{d=1}^D \frac{1}{\sqrt{2\pi\sigma^2_{y_n d}}} exp(-\frac{(x_{nd} - \mu_{y_n d})^2}{2\sigma^2_{y_n d}})\Big) \tag{30}$$

$$\ell = \sum_{n=1}^N \pi_{y_n} + \sum_{n=1}^N \sum_{d=1}^D log\big(\frac{1}{\sqrt{2\pi\sigma^2_{y_n d}}} exp(-\frac{(x_{nd} - \mu_{y_n d})^2}{2\sigma^2_{y_n d}})\big) \tag{31}$$

$$\ell = \sum_{n=1}^N \pi_{y_n} - \sum_{n=1}^N \sum_{d=1}^D \frac{1}{2} log(2\pi\sigma^2_{y_n d}) - \sum_{n=1}^N \sum_{d=1}^D \frac{(x_{nd} - \mu_{y_n d})^2}{2\sigma^2_{y_n d}} \tag{32}$$

$$(\pi_c^*, \mu_{cd}^*, {\sigma^2_{cd}}^*) = argmax \sum_{n=1}^N \pi_{y_n} - \sum_{n=1}^N \sum_{d=1}^D \frac{1}{2} log(2\pi\sigma^2_{y_n d}) - \sum_{n=1}^N \sum_{d=1}^D \frac{(x_{nd} - \mu_{y_n d})^2}{2\sigma^2_{y_n d}} \tag{33}$$

$$\ell = \sum_{c=1}^C \sum_{n:y_n=c} \pi_c - \sum_{c=1}^C \sum_{n:y_n=c} \sum_{d=1}^D \frac{1}{2} log(2\pi\sigma^2_{cd}) - \sum_{c=1}^C \sum_{n:y_n=c} \sum_{d=1}^D \frac{(x_{nd} - \mu_{cd})^2}{2\sigma^2_{cd}} \tag{34}$$

4.2

$$\frac{\partial \ell}{\partial \mu_{cd}} = \sum_{n:y_n=c} \frac{-2(x_{nd} - \mu_{cd})}{2\sigma_{cd}^2} = 0 \rightarrow \mu_{cd} = \frac{\sum_{n:y_n=c} x_{nd}}{\text{\# of data points labeled as c}} \tag{35}$$

$$\frac{\partial \ell}{\partial \sigma_{cd}} = -\sum_{n:y_n=c} \frac{1}{\sigma_{cd}} + \sum_{n:y_n=c} \frac{(x_{nd} - \mu_{cd})^2}{\sigma_{cd}^3} = \sum_{n:y_n=c} \frac{(x_{nd} - \mu_{cd})^2 - \sigma_{cd}^2}{\sigma_{cd}^3} = 0 \tag{36}$$

$$\sum_{n:y_n=c} \frac{(x_{nd} - \mu_{cd})^2 - \sigma_{cd}^2}{\sigma_{cd}^3} = 0 \rightarrow \sigma_{cd}^2 = \frac{\sum_{n:y_n=c} (x_{nd} - \mu_{cd})^2}{\text{\# of data points labeled as c}} \tag{37}$$

To find π_c, we just consider the first term of eq.(34) because the other two don't effect in derivation. Also, we should consider the constraint of summing up to one and the Lagrangian multiplier:

$$\frac{\partial}{\partial \pi_c} \left(\sum_{c=1}^{C} \sum_{n:y_n=c} \pi_c + \lambda \Big(\sum_{c}^{C} \pi_c - 1 \Big) \right) \rightarrow \pi_c = \frac{\text{\# of data points labeled as c}}{N} \tag{38}$$