University of Illinois Spring 2018

CS 446: Machine Learning
Homework 9

Due on Tuesday, April 3, 2018, 11:59 a.m. Central Time

1. **[16 points]** Gaussian Mixture Models & EM
   Consider a Gaussian mixture model with K components ($k \in \{1, \ldots, K\}$), each having mean μ_k, variance σ_k^2, and mixture weight π_k. All these are parameters to be learned, and we subsume them in the set θ. Further, we are given a dataset $X = \{x_i\}$, where $x_i \in \mathbb{R}$. We also use $Z = \{z_i\}$ to denote the latent variables, such that $z_i = k$ implies that x_i is generated from the k^{th} Gaussian.

   (a) What is the log-likelihood of the data $\log p(X; \theta)$ according to the Gaussian Mixture Model? (use μ_k, σ_k, π_k, K, x_i, and X). Don't use any abbreviations.

   Your answer:

$$
\begin{aligned}
\mathrm{LL} &= \ln \prod_{x_i \in X} p\left(x_i; \theta\right) \\
&= \sum_{x_i \in X} \ln p\left(x_i; \theta\right) \\
&= \sum_{x_i \in X} \ln \sum_{k=1}^{K} p\left(x_i; \mu_k, \sigma_k\right) \\
&= \sum_{x_i \in X} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x_i | \mu_k, \sigma_k\right) \\
&= \sum_{x_i \in X} \ln \sum_{k=1}^{K} \pi_k \left(2\pi\sigma_k^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_k^2}\left(x_i - \mu_k\right)^2\right)
\end{aligned}
$$

   Where the last equality follows because we are dealing 1-d data so there is no need for determinant or matrix inverses.

   (b) For learning θ using the EM algorithm, we need the conditional distribution of the latent variables Z given the current estimate of the parameters $\theta^{(t)}$ (we will use the superscript (t) for parameter estimates at step t). What is the posterior probability $p(z_i = k | x_i; \theta^{(t)})$? To simplify, wherever possible, use $\mathcal{N}(x_i | \mu_k, \sigma_k)$ to denote a Gaussian distribution over $x_i \in \mathbb{R}$ having mean μ_k and variance σ_k^2.

Your answer:

$$
\begin{aligned}
p\left(z_i = k | x_i; \theta^{(t)}\right) &= \frac{p\left(z_i = k, x_i; \theta^{(t)}\right)}{p\left(x_i; \theta^{(t)}\right)} \\
&= \frac{p\left(z_i = k; \theta^{(t)}\right) p\left(x_i | z_i = k; \theta^{(t)}\right)}{p\left(x_i; \theta^{(t)}\right)} \\
&= \frac{p\left(z_i = k; \theta^{(t)}\right) p\left(x_i | z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^{K} p\left(x_i, z_i = j; \theta_k^{(t)}\right)} \\
&= \frac{p\left(z_i = k; \theta^{(t)}\right) p\left(x_i | z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^{K} p\left(z_i = j; \theta^{(t)}\right) p\left(x_i | z_i = j; \theta^{(t)}\right)} \\
&= \frac{\pi_k \mathcal{N}\left(x_i | \mu_k^{(t)}, \sigma_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_i | \mu_j^{(t)}, \sigma_j^{(t)}\right)}
\end{aligned}
$$

(c) Find $\mathbb{E}_{z_i|x_i;\theta^{(t)}}[\log p(x_i, z_i; \theta)]$. Denote $p(z_i = k|x_i; \theta^{(t)})$ as z_{ik}, and use all previous notation simplifications.

Your answer:

$$
\begin{aligned}
\mathbb{E}_{z_i|x_i;\theta^{(t)}}[\log p(x_i, z_i; \theta)] &= \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\log \prod_{k=1}^{K} p\left(x_i, z_i; \theta_k\right)^{\delta(z_i=k)}\right] \\
&= \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\sum_{k=1}^{K} \log p\left(x_i, z_i = k; \theta_k\right)^{\delta(z_i=k)}\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\log p\left(x_i, z_i = k; \theta_k\right)^{\delta(z_i=k)}\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\delta\left(z_i = k\right) \log p\left(x_i, z_i = k; \theta_k\right)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\delta\left(z_i = k\right) \log p\left(z_i = k|x_i; \theta_k\right) p\left(x_i; \theta_k\right)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\delta\left(z_i = k\right) \log \pi_k p\left(x_i; \theta_k\right)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{z_i|x_i;\theta^{(t)}}\left[\delta\left(z_i = k\right)\right] \log \pi_k p\left(x_i; \theta_k\right) \\
&= \sum_{k=1}^{K} p\left(z_i = k|x_i; \theta^{(t)}\right) \log \pi_k p\left(x_i; \theta_k\right) \\
&= \sum_{k=1}^{K} z_{ik} \log \pi_k p\left(x_i; \theta_k\right) \\
&= \sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log p\left(x_i; \theta_k\right)\right) \\
&= \sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log \mathcal{N}\left(x_i|\mu_k, \sigma_k\right)\right)
\end{aligned}
$$

The idea of using a product with an indicator function as an exponent was taken from the textbook. I really liked that trick so decided to use it. Apparently there are easier ways to derive this result by just using the definition of an expectation.

(d) $\theta^{(t+1)}$ is obtained as the maximizer of $\sum_{i=1}^{N} \mathbb{E}_{z_i|x_i;\theta^{(t)}}[\log p(x_i, z_i; \theta)]$. Find $\mu_k^{(t+1)}$, $\pi_k^{(t+1)}$, and $\sigma_k^{(t+1)}$, by using your answer to the previous question.

Your answer: Summarizing:

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{ik}$$
$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} z_{ik}x_i}{N\pi_k^{(t+1)}}$$
$$\sigma_k^{2,(t+1)} = \frac{\sum_{i=1}^{N} z_{ik}\left(x_i - \mu_k^{(t+1)}\right)^2}{N\pi_k^{(t+1)}}$$

The whole derivation can be found at the end of this report. For some reason this template does not like boxes that span multiple pages.

(e) How are kMeans and Gaussian Mixture Model related? (There are three conditions)

Your answer:

$$\pi_k^{(t+1)} = \frac{1}{K} \;\; \forall k$$
$$\sigma_k^{(t+1)} = c \;\; \forall k$$
$$c \downarrow 0$$

Other relations include:
1) distance measure is different. k-Means uses Euclidean distance whereas GMM uses a Gaussian probability.
2) kMeans assumes the data is spherically clustered, as consequence of using Euclidean distance.

Lets take the objective function from c)

$$\sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log \mathcal{N} \left(x_i | \mu_k, \sigma_k \right) \right)$$

and lets also remember the explicit constraint that $\sum_k \pi_k = 1$. Note that we need to sum over the entire data set so we then can write the dual of the objective function:

$$\begin{aligned} F &= \sum_{i=1}^{N} \left(\sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log \mathcal{N} \left(x_i | \mu_k, \sigma_k \right) \right) \right) + \lambda \left(\sum_k \pi_k - 1 \right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \mathcal{N} \left(x_i | \mu_k, \sigma_k \right) + \lambda \left(\sum_k \pi_k - 1 \right) \end{aligned}$$

Now lets take the derivative with respect to the variables we want to update on each step and set it to zero:

a) For π_k (refers to $\pi_k^{(t+1)}$):

$$\frac{\partial F}{\partial \pi_k} = \sum_{i=1}^{N} z_{ik} \frac{1}{\pi_k} + 0 + \lambda = 0 \iff \pi_k = \frac{\sum_{i=1}^{N} z_{ik}}{-\lambda}$$

By using the constraint:

$$\begin{aligned} 1 = \sum_{k=1}^{K} \pi_k &= \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} z_{ik}}{-\lambda} \\ &= \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik}}{-\lambda} \\ &= \frac{\sum_{i=1}^{N} 1}{-\lambda} \\ &= \frac{N}{-\lambda} \\ \rightarrow \lambda &= -N \end{aligned}$$

Then

$$\begin{aligned} 0 = \frac{\partial F}{\partial \pi_k} &= \sum_{i=1}^{N} z_{ik} \frac{1}{\pi_k} + \lambda = \sum_{i=1}^{N} z_{ik} \frac{1}{\pi_k} - N \\ \rightarrow \pi_k^{(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} z_{ik} \end{aligned}$$

b) For μ_k:

$$
\begin{aligned}
0 &= \frac{\partial F}{\partial \mu_k} \\
&= \frac{\partial}{\partial \mu_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\log \mathcal{N}\left(x_i|\mu_k,\sigma_k\right) + \lambda\left(\sum_k \pi_k - 1\right)\right) \\
&= \frac{\partial}{\partial \mu_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\log \mathcal{N}\left(x_i|\mu_k,\sigma_k\right) + \lambda\left(\sum_k \pi_k - 1\right)\right) \\
&= \frac{\partial}{\partial \mu_k}\sum_{i=1}^{N} z_{ik}\log \mathcal{N}\left(x_i|\mu_k,\sigma_k\right) \\
&= \frac{\partial}{\partial \mu_k}\sum_{i=1}^{N} z_{ik}\log \left(2\pi\sigma_k^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)^2\right) \\
&= \frac{\partial}{\partial \mu_k}\left(\sum_{i=1}^{N} z_{ik}\log \left(2\pi\sigma_k^2\right)^{-\frac{1}{2}} + \sum_{i=1}^{N} z_{ik}\log\exp\left(-\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)^2\right)\right) \\
&= \frac{\partial}{\partial \mu_k}\left(-\sum_{i=1}^{N} z_{ik}\log\left(\left(2\pi\right)^{\frac{1}{2}}\sigma_k\right) + \sum_{i=1}^{N} z_{ik}\left(-\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)^2\right)\right) \\
&= \frac{\partial}{\partial \mu_k}\left(\sum_{i=1}^{N} z_{ik}\left(-\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)^2\right)\right) \\
&= \sum_{i=1}^{N} z_{ik}\left(\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)\right) \\
&\rightarrow \sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k\right) = 0 \\
&\rightarrow \sum_{i=1}^{N} z_{ik}x_i = \sum_{i=1}^{N} z_{ik}\mu_k \\
&\rightarrow \mu_k = \frac{\sum_{i=1}^{N} z_{ik}x_i}{\sum_{i=1}^{N} z_{ik}} \\
&\rightarrow \mu_k = \frac{\sum_{i=1}^{N} z_{ik}x_i}{N\pi_k^{(t+1)}} = \mu_k^{(t+1)}
\end{aligned}
$$

c) For σ_k:
Reusing some derivations from b)

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \sigma_k}\left(-\sum_{i=1}^{N} z_{ik}\log\left((2\pi)^{\frac{1}{2}}\sigma_k\right)+\sum_{i=1}^{N} z_{ik}\left(-\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)^2\right)\right)\\
&= \frac{\partial}{\partial \sigma_k}\left(-\sum_{i=1}^{N} z_{ik}\log\sigma_k+\sum_{i=1}^{N} z_{ik}\left(-\frac{1}{2\sigma_k^2}\left(x_i-\mu_k\right)^2\right)\right)\\
&= -\frac{1}{\sigma_k}\sum_{i=1}^{N} z_{ik}+\frac{1}{\sigma_k^3}\sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k\right)^2
\end{aligned}
$$

Multiplying by σ_k^3 on both sides

$$
\begin{aligned}
0 &= -\sigma_k^2\sum_{i=1}^{N} z_{ik}+\sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k\right)^2\\
\sigma_k^2\sum_{i=1}^{N} z_{ik} &= \sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k\right)^2\\
\sigma_k^2 &= \frac{\sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k\right)^2}{\sum_{i=1}^{N} z_{ik}}\\
\sigma_k^2 &= \frac{\sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k\right)^2}{N\pi_k^{(t+1)}} = \frac{\sum_{i=1}^{N} z_{ik}\left(x_i-\mu_k^{(t+1)}\right)^2}{N\pi_k^{(t+1)}}
\end{aligned}
$$