Introduction to Machine Learning

Homework 5 - Solution

Feb 15th, 2017 *Ori Ram*

Question 1

(a) The likelihood is

$$\begin{aligned}\mathcal{L}(\lambda; x_1, .., x_m) &= Pr\left[x_1, .., x_m|\lambda\right] \\ &= \prod_{i=1}^{m} \lambda e^{-\lambda x_i} \\ &= \lambda^m e^{-\lambda \sum_{i=1}^{m} x_i}\end{aligned} \tag{1}$$

So the log-likelihood is

$$\begin{aligned}\ell(\lambda; x_1, .., x_m) &= \log \mathcal{L}(\lambda; x_1, .., x_m) \\ &= \log\left(\lambda^m e^{-\lambda \sum_{i=1}^{m} x_i}\right) \\ &= m \log \lambda - \lambda \sum_{i=1}^{m} x_i\end{aligned} \tag{2}$$

We differentiate w.r.t λ to get

$$\frac{d\ell}{d\lambda} = \frac{m}{\lambda} - \sum_{i=1}^{m} x_i \tag{3}$$

and therefore we have

$$\lambda_{ML} = \frac{m}{\sum_{i=1}^{m} x_i} \tag{4}$$

(The second derivative is $-m/\lambda^2 < 0$, so we indeed have a maximum).

(b)

$$\begin{aligned}\lambda_{MAP} &= \arg\max_{\lambda} \mathcal{L}(\lambda; \mathbf{x}_1, .., \mathbf{x}_m) p(\lambda) \\ &= \arg\max_{\lambda} \lambda^m e^{-\lambda \sum_{i=1}^{m} x_i} e^{-\lambda} \\ &= \arg\max_{\lambda} \lambda^m e^{-\lambda\left(1+\sum_{i=1}^{m} x_i\right)} \\ &= \arg\max_{\lambda} \left[m \log \lambda - \lambda\left(1 + \sum_{i=1}^{m} x_i\right)\right]\end{aligned} \tag{5}$$

We again differentiate w.r.t λ and equating to zero:

$$\frac{m}{\lambda} - \sum_{i=1}^{m} x_i - 1 = 0 \Longrightarrow \lambda_{MAP} = \frac{m}{1 + \sum_{i=1}^{m} x_i} \tag{6}$$

(The second derivative is the same like in **(a)** so it's indeed a maximum).

Question 2

$$\begin{aligned}
\arg\min_{p\in\mathcal{F}} D_{KL}[\hat{p}; p] &= \arg\min_{p\in\mathcal{F}} \sum_x \hat{p}(x) \log \frac{\hat{p}(x)}{p(x)} \\
&= \arg\min_{p\in\mathcal{F}} \sum_x \left(\hat{p}(x) \log \hat{p}(x) - \hat{p}(x) \log p(x)\right) \\
&= \arg\min_{p\in\mathcal{F}} \sum_x -\hat{p}(x) \log p(x) \\
&= \arg\max_{p\in\mathcal{F}} \sum_x \hat{p}(x) \log p(x) \\
&= \arg\max_{p\in\mathcal{F}} \sum_x \frac{|\{i|x_i = x\}|}{n} \log p(x) \\
&= \arg\max_{p\in\mathcal{F}} \sum_{i=1}^{n} \log p(x_i) \\
&= \arg\max_{p\in\mathcal{F}} \log \prod_{i=1}^{n} p(x_i) \\
&= \arg\max_{p\in\mathcal{F}} \prod_{i=1}^{n} p(x_i) \\
&= \arg\max_{p\in\mathcal{F}} Pr_p(x_1, .., x_n)
\end{aligned}$$

Question 3

(a) The parameters of the model are $\{p_r\}_{r=0}^{31}$.

For any sample $\mathbf{x}$ and $\mathbf{z} \in \{\mathbf{s}_0, .., \mathbf{s}_{31}\} \equiv S$ we define $\delta_{\mathbf{xz}}$ to be 1 if $\mathbf{x}$ can be generated from $\mathbf{z}$ and 0 if not. We have

$$Pr[\mathbf{x}, \mathbf{z}|\mathbf{p}] = Pr[\mathbf{z}|\mathbf{p}] \cdot Pr[\mathbf{x}|\mathbf{z}, \mathbf{p}] = p_{\mathbf{z}} \cdot \delta_{\mathbf{xz}} \frac{1}{10} \tag{7}$$

The likelihood is therefore

$$\begin{aligned}\mathcal{L}(\mathbf{p};\mathbf{x}_1..\mathbf{x}_n) &= \Pr\left[\mathbf{x}_1..\mathbf{x}_n|\mathbf{p}\right] \\ &= \prod_{i=1}^{n} Pr\left[\mathbf{x}_i|\mathbf{p}\right] \\ &= \prod_{i=1}^{n}\sum_{\mathbf{z}\in S} Pr\left[\mathbf{x}_i,\mathbf{z}|\mathbf{p}\right] \\ &= \prod_{i=1}^{n}\sum_{\mathbf{z}\in S}\frac{1}{10}p_{\mathbf{z}}\delta_{\mathbf{x}_i\mathbf{z}}\end{aligned} \tag{8}$$

So the log-likelihood is

$$\ell(\mathbf{p};\mathbf{x}_1,..,\mathbf{x}_n) = \sum_{i=1}^{n}\log\left(\sum_{\mathbf{z}\in S}\frac{1}{10}p_{\mathbf{z}}\delta_{\mathbf{x}_i\mathbf{z}}\right) \tag{9}$$

(b) We will use EM in order to estimate the parameters $\mathbf{p}$.

E-step

$$\begin{aligned}Q(\mathbf{p}|\mathbf{p}^{(t)}) &= \sum_{i=1}^{n}\sum_{\mathbf{z}\in S} Pr\left[\mathbf{z}|\mathbf{x}_i,\mathbf{p}^{(t)}\right]\log Pr\left[\mathbf{x}_i,\mathbf{z}|\mathbf{p}\right] \\ &= \sum_{i=1}^{n}\sum_{\mathbf{z}\in S}\left(\frac{Pr\left[\mathbf{z},\mathbf{x}_i|\mathbf{p}^{(t)}\right]}{\sum_{\mathbf{z}'\in S} Pr\left[\mathbf{z}',\mathbf{x}_i|\mathbf{p}^{(t)}\right]}\right)\log\left(\frac{1}{10}p_{\mathbf{z}}\delta_{\mathbf{x}_i\mathbf{z}}\right) \\ &\equiv \sum_{i=1}^{n}\sum_{\mathbf{z}\in S} a_{i\mathbf{z}}^{(t)}\log\left(\frac{1}{10}p_{\mathbf{z}}\delta_{\mathbf{x}_i\mathbf{z}}\right)\end{aligned} \tag{10}$$

for

$$\begin{aligned}a_{i\mathbf{z}}^{(t)} &= \frac{Pr\left[\mathbf{z},\mathbf{x}_i|\mathbf{p}^{(t)}\right]}{\sum_{\mathbf{z}'\in S} Pr\left[\mathbf{z}',\mathbf{x}_i|\mathbf{p}^{(t)}\right]} \\ &= \frac{p_{\mathbf{z}}^{(t)}\delta_{\mathbf{x}_i\mathbf{z}}}{\sum_{\mathbf{z}'\in S}\left(p_{\mathbf{z}'}^{(t)}\delta_{\mathbf{x}_i\mathbf{z}'}\right)}\end{aligned} \tag{11}$$

So $a_{i\mathbf{z}}^{(t)}$ is a constant which we know how to calculate at time t.

M-step We consider the optimization problem

$$\begin{aligned}\max_{\mathbf{p}} \quad & Q\left(\mathbf{p}|\mathbf{p}^{(t)}\right) \\ \text{s.t} \quad & \sum_{\mathbf{z}\in S} p_{\mathbf{z}} = 1\end{aligned} \tag{12}$$

We don't write the non-negativity constraints, because they will be satisfied anyway. We will write the Lagrangian:

$$L(\mathbf{p},\lambda)=\sum_{i=1}^{n}\sum_{\mathbf{z}\in S}a_{i\mathbf{z}}^{(t)}\log\left(\frac{1}{10}p_{\mathbf{z}}\delta_{\mathbf{x}_i\mathbf{z}}\right)-\lambda\left(\sum_{\mathbf{z}\in S}p_{\mathbf{z}}-1\right) \tag{13}$$

Differentiating it w.r.t $\mathbf{p}$ and equating to zero leads to

$$\frac{\partial L}{\partial p_{\mathbf{z}}}=\sum_{i=1}^{n}\frac{a_{i\mathbf{z}}^{(t)}}{p_{\mathbf{z}}}-\lambda=0\Longrightarrow p_{\mathbf{z}}=\frac{\sum_{i=1}^{n}a_{i\mathbf{z}}^{(t)}}{\lambda}$$

And in order to get a valid distribution, we have:

$$p_{\mathbf{z}}^{(t+1)}=\frac{\sum_{i=1}^{n}a_{i\mathbf{z}}^{(t)}}{\sum_{\mathbf{z}'\in S}\sum_{i=1}^{n}a_{i\mathbf{z}'}^{(t)}}=\frac{\sum_{i=1}^{n}a_{i\mathbf{z}}^{(t)}}{n} \tag{14}$$

We can differentiate L again to get a negative second derivative $\left(-\sum_{i=1}^{n}a_{i\mathbf{z}}^{(t)}\right)/p_{\mathbf{z}}^2<0$, so we indeed have a maximum.

Question 4

(a) The parameters of the model are $\theta=\left\{p_1,..,p_m,\boldsymbol{\mu}_1,..,\boldsymbol{\mu}_m,\sigma_1^2,..,\sigma_m^2\right\}$ where m is the number of clusters. First, we see that

$$\begin{aligned}Pr[\mathbf{x},\mathbf{z}|\theta]&=Pr[\mathbf{z}|\theta]\cdot Pr[\mathbf{x}|\mathbf{z},\theta]\\&=p_{\mathbf{z}}\cdot\frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{z}})^T\left(\sigma_{\mathbf{z}}^2\cdot I\right)^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{z}})\right)}{\sqrt{|2\pi\sigma_{\mathbf{z}}^2\cdot I|}}\\&=p_{\mathbf{z}}\cdot\left(2\pi\sigma_{\mathbf{z}}^2\right)^{-k/2}\cdot\exp\left(-\frac{1}{2\sigma_{\mathbf{z}}^2}||\mathbf{x}-\boldsymbol{\mu}_{\mathbf{z}}||^2\right)\end{aligned} \tag{15}$$

Next, we will derive an EM procedure for a local maximization of the parameters:

E-step:

$$\begin{aligned}Q(\theta|\theta^{(t)})&=\sum_{i=1}^{n}\sum_{j=1}^{m}Pr\left[\mathbf{z}_i=j|\mathbf{x}_i,\theta^{(t)}\right]\log Pr\left[\mathbf{x}_i,\mathbf{z}_i=j|\theta\right]\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{Pr\left[\mathbf{z}_i=j,\mathbf{x}_i|\theta^{(t)}\right]}{\sum_{j'=1}^{m}Pr\left[\mathbf{z}_i=j',\mathbf{x}_i|\theta^{(t)}\right]}\right)\log Pr[\mathbf{x}_i,\mathbf{z}_i=j|\theta]\\&\equiv\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}^{(t)}\cdot\left(\log p_j-\frac{k}{2}\log\left(2\pi\sigma_j^2\right)-\frac{1}{2\sigma_j^2}||\mathbf{x}_i-\boldsymbol{\mu}_j||^2\right)\end{aligned} \tag{16}$$

where again

$$\begin{aligned} a_{ij}^{(t)} &= \frac{Pr\left[\mathbf{z}_i = j, \mathbf{x}_i | \theta^{(t)}\right]}{\sum_{j'=1}^{m} Pr\left[\mathbf{z}_i = j', \mathbf{x}_i | \theta^{(t)}\right]} \\ &= \frac{p_j^{(t)} \cdot \left(2\pi(\sigma_j^2)^{(t)}\right)^{-k/2} \cdot \exp\left(-\frac{1}{2(\sigma_j^2)^{(t)}} ||\mathbf{x}_i - \boldsymbol{\mu}_j^{(t)}||^2\right)}{\sum_{j'=1}^{m} p_{j'}^{(t)} \cdot \left(2\pi(\sigma_{j'}^2)^{(t)}\right)^{-k/2} \cdot \exp\left(-\frac{1}{2(\sigma_{j'}^2)^{(t)}} ||\mathbf{x}_i - \boldsymbol{\mu}_{j'}^{(t)}||^2\right)} \end{aligned} \tag{17}$$

and it is a constant which we can calculate at time t (it depends only on $\mathbf{x}_1, .., \mathbf{x}_n$ and $\theta^{(t)}$).

M-step We consider the optimization problem

$$\begin{aligned} \max_{\theta} \quad & Q\left(\theta | \theta^{(t)}\right) \\ \text{s.t} \quad & \sum_{j=1}^{m} p_j = 1 \end{aligned} \tag{18}$$

We don't write the non-negativity constraints, because they will be satisfied anyway. We will write the Lagrangian:

$$\mathcal{L}\left(p_1, .., p_m, \boldsymbol{\mu}_1, .., \boldsymbol{\mu}_m, \sigma_1^2, .., \sigma_m^2, \lambda\right) = Q(\theta | \theta^{(t)}) - \lambda\left(\sum_{i=1}^{m} p_i - 1\right)$$

Assigning $Q(\theta | \theta^{(T)})$ we get

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^{(t)} \cdot \left(\log p_j - \frac{k}{2}\log\left(2\pi\sigma_j^2\right) - \frac{1}{2\sigma_j^2}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2\right) - \lambda\left(\sum_{i=1}^{m} p_i - 1\right)$$

Now we will differentiate:

$$\frac{\partial \mathcal{L}}{\partial p_j} = \sum_{i=1}^{n} a_{ij}^{(t)} \cdot \frac{1}{p_j} - \lambda = 0 \Longrightarrow p_j^{(t+1)} = \frac{\sum_{i=1}^{n} a_{ij}^{(t)}}{\lambda}$$

and again to get a valid distribution we have

$$p_j^{(t+1)} = \frac{\sum_{i=1}^{n} a_{ij}^{(t)}}{\sum_{j'=1}^{m}\sum_{i=1}^{n} a_{ij'}^{(t)}} = \frac{\sum_{i=1}^{n} a_{ij}^{(t)}}{n} \tag{19}$$

We can differentiate again and this is indeed a maximum.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_j} &= \frac{1}{\sigma_j^2}\sum_{i=1}^{n} a_{ij}^{(t)}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right) = 0 \\ &\Longrightarrow \boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} a_{ij}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} a_{ij}^{(t)}} \end{aligned} \tag{20}$$

The Hessian matrix is

$$\left(-\frac{1}{\sigma_j^2}\sum_{i=1}^{n} a_{ij}^{(t)}\right) I$$

which is negative definite, therefore this is indeed a maximum.

$$\frac{\partial \mathcal{L}}{\partial \sigma_j^2} = \sum_{i=1}^{n} a_{ij}^{(t)} \left(-\frac{k}{2}\cdot\frac{1}{\sigma_j^2} + \frac{1}{2(\sigma_j^2)^2}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2\right) = 0$$

$$\frac{k}{2\cancel{\sigma_j^2}}\sum_{i=1}^{n} a_{ij}^{(t)} = \frac{1}{2(\sigma_j^2)\cancel{^2}}\sum_{i=1}^{n} a_{ij}^{(t)}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$$

We divided because $\sigma_j^2 > 0$, therefore

$$(\sigma_j^2)^{(t+1)} = \frac{1}{k}\cdot\frac{\sum_{i=1}^{n} a_{ij}^{(t)}||\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)}||^2}{\sum_{i=1}^{n} a_{ij}^{(t)}} \tag{21}$$