Machine Learning 1 - Homework 5

Pascal Mattia Esser

October 18th, 2017

1

1.1

$$
\begin{aligned}
p(x) &= \prod_{n}^{N} p(x_n) && (1)\\
&= \prod^{N} \sum^{K} \pi_k p(x_n|\lambda_n) && (2)\\
&= \prod^{N} \sum^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} exp(-\lambda_k) && (3)
\end{aligned}
$$

1.2

$$
\ln(p(x)) = \sum_{n}^{N} ln \sum_{k}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} exp(-\lambda_k) \tag{4}
$$

1.3

$$
\begin{aligned}
r_{nk} &= \frac{\pi_k p(x_n|\lambda_n)}{\sum_j^K \pi_j p(x_n|\lambda_j)} && (5)\\
&= \frac{\pi_k \frac{1}{x_n!} \lambda_k^{x_n} exp(-\lambda_k)}{\sum_j^K \pi_j \frac{1}{x_n!} \lambda_j^{x_n} exp(-\lambda_j)} && (6)\\
&= \frac{\pi_k \lambda_k^{x_n} exp(-\lambda_k)}{\sum_j^K \pi_j \lambda_j^{x_n} exp(-\lambda_j)} && (7)
\end{aligned}
$$

1.4

$$\frac{\partial \ln p(x)}{\partial \lambda_k} = \sum^N \frac{\pi_k}{\sum_j^K \pi_j p(x_n|\lambda_j)} \frac{\partial p(x_n|\lambda_k)}{\partial \lambda_k} \tag{8}$$

$$\frac{\partial p(x_n|\lambda_k)}{\partial \lambda_k} = \frac{1}{x_n!}(x_n \lambda_k^{x_n - 1} exp(-\lambda_k) - \lambda_k^{x_n} exp(-\lambda_k) \tag{9}$$

$$\frac{\partial \ln p(x)}{\partial \lambda_k} = \sum^N \frac{\pi_k}{\sum_j^K \pi_j p(x_n|\lambda_j)} \left(\frac{1}{x_n!}(x_n \lambda_k^{x_n - 1} exp(-\lambda_k) - \lambda_k^{x_n} exp(-\lambda_k) \right) \tag{10}$$

$$= \sum^N \frac{\pi_k}{\sum_j^K \pi_j p(x_n|\lambda_j)} \left(\frac{1}{x_n!}(x_n \lambda_k^{x_n - 1} exp(-\lambda_k)) - \lambda_k^{x_n} exp(-\lambda_k) \right) \tag{11}$$

$$= \sum^N \frac{\pi_k}{\sum_j^K \pi_j p(x_n|\lambda_j)} \left(\frac{1}{x_n!}(\frac{x_n}{\lambda_k} \lambda_k^{x_n} exp(-\lambda_k)) - \lambda_k^{x_n} exp(-\lambda_k) \right) \tag{12}$$

$$= \sum^N \left(\frac{x_n}{\lambda_k} r_{nk} - r_{nk} \right) = 0 \tag{13}$$

$$\text{with } N_k = \sum^N r_{nk} \tag{14}$$

$$\Rightarrow \frac{1}{\lambda_k} x_n r_{nk} = N_k \tag{15}$$

$$\Rightarrow \lambda_k = \frac{1}{N_k} \sum^N r_{nk} x_n \tag{16}$$

1.5

$$L(\boldsymbol{x}, \lambda) = ln(p(x)) + \lambda \left(\sum^K \pi_j - 1 \right) = 0 \tag{17}$$

$$= \sum^N \frac{\pi_k}{\sum_j^K \pi_j p(x_n|\lambda_j)} + \lambda \tag{18}$$

$$\text{with: } \frac{\pi_k}{\sum_j^K \pi_j p(x_n|\lambda_j)} = \frac{r_{nk}}{\pi_n} \tag{19}$$

$$L(\boldsymbol{x}, \lambda) = \sum^N r_{nk} + \lambda \pi_k = 0 \tag{20}$$

$$\Rightarrow N_k + \lambda \pi_k = 0 \tag{21}$$

$$\text{with: } \sum^K \pi_k \lambda = \sum^K (-N_k) \Rightarrow \lambda = -N \tag{22}$$

$$\Rightarrow \pi_k = \frac{N_k}{N} \tag{23}$$

1.6

set:

$$p(\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\lambda}|a, b, \alpha, K) =: p(\boldsymbol{P}|\boldsymbol{\Theta}) \tag{24}$$

now write:

$$p(\boldsymbol{P}|\boldsymbol{\Theta}) = D\left(\boldsymbol{\pi}, \frac{\alpha}{K}, ..., \frac{\alpha}{K}\right) \prod_j^K G(\lambda_j|a, b) \left(\prod_n^N \sum_k^K \pi_k \frac{1}{x_n!} \lambda_k^{x_n} exp(-\lambda_k) \right) \tag{25}$$

$$\ln p(\boldsymbol{P}|\boldsymbol{\Theta}) = \ln\left(D\left(\boldsymbol{\pi}, \frac{\alpha}{K}, ..., \frac{\alpha}{K}\right)\right) + \sum_{j}^{K} \ln G(\lambda_j|a,b) + \sum_{n}^{N} ln \sum_{k}^{K} \pi_k \frac{1}{x_n!}\lambda_k^{x_n} exp(-\lambda_k) \tag{26}$$

$$= \sum_j (a-1)\ln\lambda_j - b\lambda_j + \sum_j (\alpha/K - 1)\ln\pi_j \tag{27}$$

$$+ \sum_{n}^{N} ln \sum_{k}^{K} \pi_k \frac{1}{x_n!}\lambda_k^{x_n} exp(-\lambda_k) + C \tag{28}$$

1.7

$$\frac{\partial \ln p(\boldsymbol{P}|\boldsymbol{\Theta})}{\partial\lambda_k} = \frac{1}{G(\lambda_j|a,b)}\frac{\partial G(\lambda_k|a,b)}{\partial\lambda_k} + \sum^{N}\left(\frac{x_n}{\lambda_k}r_{nk} - r_{nk}\right) \tag{29}$$

$$\frac{\partial G(\lambda_k|a,b)}{\partial\lambda_k} = \frac{\partial}{\partial\lambda_k}\left(\frac{1}{\Gamma(a)}b^a\lambda_k^{a-1}exp(-b\lambda)\right) \tag{30}$$

$$= \frac{1}{\Gamma(a)}b^a(a-1)\lambda_k^{a-2}exp(-b\lambda_k) - b\lambda_k^{a-1}exp(-b\lambda_k) \tag{31}$$

$$= \frac{1}{\Gamma(a)}b^a\lambda_k^{a-2}exp(-b\lambda)(a-1-b\lambda_k) \tag{32}$$

$$= G(\lambda_k|a,b)\frac{a-1-b\lambda_k}{\lambda_k} \tag{33}$$

$$= G(\lambda_k|a,b)\left(\frac{a-1}{\lambda_k} - b\right) \tag{34}$$

$$\frac{\partial \ln p(\boldsymbol{P}|\boldsymbol{\Theta})}{\partial\lambda_k} = \frac{1}{G(\lambda_k|a,b)}\frac{\partial}{\partial\lambda_k}G(\lambda_k|a,b) + \sum^{N}\left(\frac{x_n}{\lambda_k}r_{nk} - r_{nk}\right) \tag{35}$$

$$= \frac{1}{G(\lambda_k|a,b)}G(\lambda_k|a,b)\left(\frac{a-1}{\lambda_k} - b\right) + \sum^{N}\left(\frac{x_n}{\lambda_k}r_{nk} - r_{nk}\right) \tag{36}$$

$$= \frac{a-1}{\lambda_k} - b + 1\lambda_k\sum^{N} x_n r_{nk} - N_k = 0 \tag{37}$$

$$\Rightarrow \lambda_k = \frac{a-1+\sum^N x_n r_{nk}}{b+N_k} \tag{38}$$

1.8

$$
\begin{aligned}
L(\boldsymbol{x},\lambda) &= ln(p(x)) + \lambda\left(\sum^{K}\pi_j - 1\right) = 0 && (39)\\
\frac{\partial \ln D(\boldsymbol{\pi},\alpha/K,...,\alpha/K)}{\partial \pi_k} &= \frac{\partial}{\partial \pi_k}\left(C\frac{\alpha}{K}\prod_{j=1}^{K}\pi_j^{\alpha/K-1}\right) && (40)\\
&= \frac{1}{\pi_j^{\alpha/K-1}}\pi_j^{\alpha/K-2}(\alpha/K-1) && (41)\\
&= \frac{\alpha-K}{K\pi_k} && (42)\\
\frac{\partial L(x,\lambda)}{\partial \pi_k} &= \frac{\alpha-K}{K\pi_k} + \sum_{n=1}^{N}\frac{\frac{1}{x_n!}\lambda_k^{x_n}exp(-\lambda_k)}{\sum_j^K \pi_j\frac{1}{x_n!}\lambda_j^{x_n}exp(-\lambda_j)} + \lambda && (43)\\
&= \frac{\alpha-K}{K\pi_k} + \sum_{n=1}^{N}\frac{r_{nk}}{\pi_k} + \lambda = 0 && (44)\\
&= \lambda\pi_k = \frac{K-\alpha}{K} - N_k && (45)\\
&\Rightarrow \lambda = \sum_{j=1}^{K}\left(\frac{K-\alpha}{K} - N_k\right) && (46)\\
&= K - \alpha - N && (47)\\
\pi_k &= \frac{K-\alpha-KN_k}{K(K-\alpha-N)} && (48)
\end{aligned}
$$

1.9

Algorithm 1 EM

initialize $\boldsymbol{\pi}$ randomly
initialize λ randomly

//Repeat until convergence
while Δ(log joint) $> \varepsilon$ **do**
  // E Step
  for all k,n **do**

    $r_nk = \frac{\pi_k \lambda^{x_n k} exp(-\lambda_k)}{\sum_j^K \pi_j \lambda_j^{x_n} exp(-\lambda_j)}$
  // M step
  for all k **do**

    $\lambda_k = \frac{a-1+\sum^N x_n r_{nk}}{b+N_k}$

    $\pi_k = \frac{K-\alpha-KN_k}{K(K-\alpha-N)}$
  compute log joint

2

2.1

$$\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \frac{1}{N}\sum_d^N \boldsymbol{x}_d \tag{49}$$

$$= \boldsymbol{x}_n - \overline{\boldsymbol{x}} \tag{50}$$

2.2

$$\frac{1}{N}\sum_n^N \hat{\boldsymbol{x}}_n = \frac{1}{N}\sum_n^N \left(\boldsymbol{x}_n - \frac{1}{N}\sum_d^N \boldsymbol{x}_d\right) \tag{51}$$

$$\sum_n^N \hat{\boldsymbol{x}}_n = \sum_n^N \left(\boldsymbol{x}_n - \frac{1}{N}\sum_d^N \boldsymbol{x}_d\right) \tag{52}$$

$$= \sum_n^N \boldsymbol{x}_n - N\frac{1}{N}\sum_d^N \boldsymbol{x}_d \tag{53}$$

$$= \sum_n^N \boldsymbol{x}_n - \sum_d^N \boldsymbol{x}_d \tag{54}$$

$$\sum_n^N \boldsymbol{x}_n = \sum_d^N \boldsymbol{x}_d \tag{55}$$

$$\Rightarrow \frac{1}{N}\sum_n^N \hat{\boldsymbol{x}}_n = 0 \tag{56}$$

2.3

$$\boldsymbol{S} = \frac{1}{N}\sum^N \left((\boldsymbol{x}_n - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_n)^T\right) \tag{57}$$

$$= \frac{1}{N}\sum_n \hat{\boldsymbol{x}}_n \hat{\boldsymbol{x}}_n^T \tag{58}$$

$$= \frac{1}{N}\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T \tag{59}$$

2.4

The dimensions of S must be DxD (follows directly from the definition$\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T$)

2.5

write the projection as:

$$\boldsymbol{y}_n = \Lambda_M^{-1/2} U_m^T \hat{\boldsymbol{x}}_n \tag{60}$$

So that $L := \Lambda_M^{-1/2} U_m^T$. Zero mean follows directly, since the transformation through L is linear:

$$\begin{aligned}
\overline{\boldsymbol{y}}_n &= \frac{1}{N}\sum_n^N (\boldsymbol{y}_n) && (61)\\
&= \frac{1}{N}\sum_n^N \left(\Lambda_M^{-1/2} U_m^T \overline{\boldsymbol{x}}_n\right) && (62)\\
&= \frac{1}{N}\sum_n^N \left(\Lambda_M^{-1/2} U_m^T \mathbf{0}\right) && (63)\\
&= \mathbf{0} && (64)
\end{aligned}$$

Identity Covariance. We use $\boldsymbol{y}_n = \Lambda_M^{-1/2} U_m^T \hat{\boldsymbol{x}}_n$, $\boldsymbol{U}_M^T \boldsymbol{U}_M = \boldsymbol{I}$, $\boldsymbol{S} = U_M \boldsymbol{\Lambda} \boldsymbol{U}_M^T$ and $(\Lambda_M^{-1/2})^T = (\Lambda_M^{-1/2})$ because it is a diagonal matrix.

$$\begin{aligned}
\boldsymbol{C} &= \frac{1}{N}\sum_n^N \boldsymbol{y}_n \boldsymbol{y}_n^T && (65)\\
&= \frac{1}{N}\sum_n^N \boldsymbol{\Lambda}_M^{-1/2} \boldsymbol{U}_M^T (\boldsymbol{x}_n - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_n)^T \boldsymbol{U}_M (\Lambda_M^{-1/2})^T && (66)\\
&= \boldsymbol{\Lambda}_M^{-1/2} \boldsymbol{U}_M^T \boldsymbol{S} \boldsymbol{U}_M \boldsymbol{\Lambda}_M^{-1/2} && (67)\\
&= \boldsymbol{\Lambda}_M^{-1/2} \boldsymbol{U}_M^T \boldsymbol{U}_M \boldsymbol{\Lambda} \boldsymbol{U}_M^T \boldsymbol{U}_M \boldsymbol{\Lambda}_M^{-1/2} && (68)\\
&= \boldsymbol{\Lambda}_M^{-1/2} \boldsymbol{I} \boldsymbol{\Lambda} \boldsymbol{I} \boldsymbol{\Lambda}_M^{-1/2} && (69)\\
&= \boldsymbol{\Lambda}_M^{-1/2} \boldsymbol{\Lambda} \boldsymbol{\Lambda}_M^{-1/2} && (70)\\
&= \boldsymbol{I} && (71)
\end{aligned}$$

The process is called whitening or sharpening (consisting of centering and de-correlate of the features and unit standard deviation by rescaling).