Machine Learning 1 - Homework 5

Selene Baez

1 PCA

Suppose we have a dataset of N vectors $\{\mathbf{x}_n\}$ of dimension D. We can write the entire dataset as a D by N matrix $\mathbf{X}$ (column n is x_n). We may wish to perform PCA on this data in the original data space, or in *kernel*-space using kernel-PCA. In the latter case, the data are projected into *feature* space $\boldsymbol{\phi}$, such that $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ is M-dimensional feature space representation of x_n. Consider the procedure for PCA (which can be generalized to kernel-PCA):

Step 1 Center $\mathbf{X}$, producing a center data matrix $\hat{\mathbf{X}}$.

Step 2 Compute sample covariance $\mathbf{S}$ of the centered dataset.

Step 3 Solve the eigen-value problem $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ is a column matrix of eigen-vectors and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigen-values λ_k, ie $\boldsymbol{\Lambda}_{kl} = \lambda_k \delta_{kl}$, where $\delta_{kl} = 1$ iff $k = l$.

Step 4 Pick eigen-vectors with largest eigen-values $\{\mathbf{u}_1, \dots \mathbf{u}_K\}$.

Step 5 Project data onto K-dimensional manifold.

Answer the following questions:

(a) Provide an expression for $\hat{\mathbf{x}}_n$.
Solution:

$$\hat{\mathbf{x}}_n = \mathbf{x}_n - \frac{1}{N}\sum_{m}^{M} \mathbf{x}_m = \mathbf{x}_n - \bar{\mathbf{x}}$$

Where $\bar{\mathbf{x}}$ is the mean of the vector $\mathbf{x}$

(b) Prove that the average of $\hat{\mathbf{x}}_n$ (over N data vectors) is the 0 vector.
Solution:

$$\sum_{n}^{N} \hat{\mathbf{x}}_n = \sum_{n}^{N} (\mathbf{x}_n - \bar{\mathbf{x}}) = \sum_{n}^{N} \mathbf{x}_n - N\bar{\mathbf{x}}$$

Using the definition of $\bar{\mathbf{x}}$

$$\sum_{n}^{N} \hat{\mathbf{x}}_n = \sum_{n}^{N} \mathbf{x}_n - N\frac{1}{N}\sum_{m}^{M} \mathbf{x}_m = \sum_{n}^{N} \mathbf{x}_n - \sum_{m}^{M} \mathbf{x}_m$$

Since the following is true: $\sum_{n}^{N} \mathbf{x}_n = \sum_{m}^{M} \mathbf{x}_m$

Then we get :

$$\sum_{n}^{N} \hat{\mathbf{x}}_n = 0$$

(c) Provide an expression for $\mathbf{S}$ in terms of $\hat{\mathbf{X}}$.
Solution:

$$\mathbf{S} = \frac{1}{N}\sum_{n}^{N} \hat{\mathbf{x}}_n \hat{\mathbf{x}}_n^T = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T$$

(d) What is the dimensionality of $\mathbf{S}$?
Solution:
Since $\mathbf{S}$ depends on the dot product of $\mathbf{X}$ with its transpose, then the dimensionality is D by D

(e) What is the expression for the linear projection $\mathbf{L}$ that maps data vectors $\hat{\mathbf{x}}_n$ onto a K-dimensional sub-space, $y_n = \mathbf{L}\hat{\mathbf{x}}_n$, such that it has zero mean and identity covariance. Prove that the average over N of y_n is 0. Prove that the covariance of y_n is the identity. What is this operation called?
Solution:
a) Linear projection expression:

$$\mathbf{L} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T$$

b) Average over N

$$\sum_n^N y_n = \sum_n^N \mathbf{L}\hat{\mathbf{x}}_n = \sum_n^N \left(\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\hat{\mathbf{x}}_n\right) = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T \sum_n^N \hat{\mathbf{x}}_n$$

Since we know that: $\sum_n^N \hat{\mathbf{x}}_n = 0$, then:

$$\sum_n^N y_n = 0$$

c) Covariance
Projection for i^{th} entry is:

$$y_n^{(i)} = \lambda_i^{-1/2}\mathbf{U}^{(i)T}\hat{\mathbf{x}}_n$$

Then the ij^{th} entry of the covariance matrix is:

$$\begin{aligned} c_{ij} &= \frac{1}{N}\sum_n^N \left(y_n^{(i)}y_n^{(j)}\right) = \frac{1}{N}\sum_n^N \left(\lambda_i^{-1/2}\mathbf{U}^{(i)T}\hat{\mathbf{x}}_n * \hat{\mathbf{x}}_n^T\lambda_j^{-1/2}\mathbf{U}^{(j)}\right) \\ &= \left(\lambda_i^{-1/2}\mathbf{U}^{(i)T}\right)\mathbf{S}\left(\lambda_j^{-1/2}\mathbf{U}^{(j)}\right) = \left(\lambda_i^{-1/2}\mathbf{U}^{(i)T}\right)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\left(\lambda_j^{-1/2}\mathbf{U}^{(j)}\right) \end{aligned}$$

Breaking it down into cases:

$$\begin{cases} \frac{\lambda_i}{\lambda_i} = 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Which equals $\mathbf{I}$
d) The operation is called sphering

2 Mixture Models

Consider a data distribution whose underlying generating process is a mixture of Poisson distributions, but we do not know the parameters of the mixture model. In this question you are asked to derive the update equations for the general Poisson mixture model.

The Poisson distribution is:

$$P(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)$$

where $x = 0, 1, 2, ...$ (non-negative integers), $\lambda > 0$ is the 'rate' of the data; the expected value of x is λ. A mixture representation assumes the following:

$$P(x_n) = \sum_{k=1}^{K} \pi_k P(x_n|\lambda_k)$$

where $P(x_n|\lambda_k)$ is a Poisson distribution with rate λ_k and x_n is a single data observation. To answer the following questions assume we are given a dataset $\{x_1, x_2, \ldots, x_N\}$. Make sure that the constraint $\sum_k \pi_k = 1$ is satisfied (i.e. think of the log-likelihood or log-joint as f (an objective to maximize) and $\sum_k \pi_k - 1 = 0$ as $g = 0$ (a constraint that must hold)).

(a) Write down the likelihood (as usual) for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.
Solution:
Plugin in the Poisson distribution in the mixture representation:

$$P(x_n) = \sum_{k=1}^{K} \pi_k P(x_n|\lambda_k) = \sum_{k=1}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp -\lambda_k$$

Assuming i.i.d for the variables:

$$\mathcal{L} = \prod_{n}^{N} p(x_n) = \prod_{n}^{N} \sum_{k}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda)$$

(b) Write down the log-likelihood (as usual) for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.
Solution:
Taking the log of the previous:

$$\ell = \log \left(\prod_{n}^{N} \sum_{k}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k) \right) = \sum_{n}^{N} \log \sum_{k}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)$$

(c) Find the expression for the responsibilities r_{nk}.
Solution:
The responsibility is defined as:

$$r_{nk} = p(z_{nk}|x_n) = \frac{p(x_n, z_{nk})}{\sum_j^K p(x_n, z_{nj})} = \frac{\pi_k P(x_n|\lambda_k)}{\sum_j^K \pi_k P(x_n|\lambda_k)} = \frac{\pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)}{\sum_j^K \pi_j \frac{1}{x_n!} \lambda_j^{x_n} \exp(-\lambda_j)}$$

(d) Find the expression for λ_k that maximizes the log-likelihood.
Solution:
We find the partial derivative:

$$\begin{aligned}
\frac{\partial \ell}{\partial \lambda_k} &= \sum_n^N \frac{1}{\sum_j^K \pi_j \frac{1}{x_n!} \lambda_j^{x_n} \exp(-\lambda_j)} \pi_k \frac{1}{x_n!} \left(x_n \lambda_k^{x_n - 1} \exp(-\lambda_k) - \lambda_k^{x_n} \exp(-\lambda_k) \right) \\
&= \sum_n^N \frac{\pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)}{\sum_j^K \pi_j \frac{1}{x_n!} \lambda_j^{x_n} \exp(-\lambda_j)} \left(x_n \lambda_k^{-1} - 1 \right)
\end{aligned}$$

Substituting r_{nk} into the equation.

$$\frac{\partial \ell}{\partial \lambda_k} = \sum_n^N r_{nk} \left(x_n \lambda_k^{-1} - 1 \right) \tag{1}$$

Set derivative to 0 and solve for λ_k

$$\begin{aligned}
\sum_n^N r_{nk} \left(x_n \lambda_k^{-1} - 1 \right) &= 0 \\
\sum_n^N r_{nk} x_n \lambda_k^{-1} - \sum_n^N r_{nk} &= 0 \\
\lambda_k &= \frac{\sum_n^N r_{nk} x_n}{\sum_n^N r_{nk}}
\end{aligned}$$

(e) Find the expression for π_k that maximizes the log-likelihood.
Solution:
Since we have the constraint that $\sum_k \pi_k = 1$, then the objective function to minimize is:

$$f = \sum_n^N \log \sum_k^K \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k) + \mu(\sum_k^K \pi_k - 1)$$

Finding the partial derivative, substituting r_{nk}:

$$\frac{\partial f}{\partial \pi_k} = \sum_n^N \frac{\frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)}{\sum_j^K \pi_j \frac{1}{x_n!} \lambda_j^{x_n} \exp(-\lambda_j)} + \mu = \sum_n^N r_{nk} + \pi_k \mu \tag{2}$$

Set derivative to 0 and solve for π_k

$$\sum_n^N r_{nk} + \pi_k \mu = 0$$

$$\pi_k = -\frac{\sum_n^N r_{nk}}{\mu}$$

Simplifying, using $N_k = \sum_n^N r_{nk}$

$$\begin{aligned}\sum_k \pi_k &= -\sum_k \frac{\sum_n^N r_{nk}}{\mu} \\ 1 &= -\frac{\sum_n^N N_k}{\mu} = -\frac{N}{\mu} \\ \mu &= -N\end{aligned}$$

Therefore:

$$\pi_k = \frac{N_k}{N}$$

(f) Now assume priors for π_k and λ_k. $p(\lambda_k|a,b) = \mathcal{G}(\lambda_k|a,b)$ (a Gamma prior) and $p(\pi_1,\ldots,\pi_k) = \mathcal{D}(\pi_1,\ldots,\pi_k|\alpha/K,\ldots,\alpha/K)$ (a Dirchlet distribution). These distributions are defined in the appendix of Bishop. Write down the log-joint distribution
$\log p(\mathbf{x}_1,\ldots,\mathbf{x}_N,\{\pi_k\},\{\lambda_k\}|a,b,\alpha,K)$.
Solution:
Log-prior for λ_k.

$$\begin{aligned}\log p(\lambda_k|a,b) &= \log \frac{b^a \lambda_k^{a-1} e^{-b\lambda_k}}{\Gamma(a)} = \log b^a + \log \lambda_k^{a-1} - b\lambda_k - \log\Gamma(a) \\ &= C_\lambda + \log\left((a-1)\lambda_k\right) - b\lambda_k \\ \text{where: } C_\lambda &= \log b^a - \log\Gamma(a)\end{aligned}$$

Log-prior for $\{\pi_k\}$.

$$\begin{aligned}\log p(\pi_1,...,\pi_k) &= \log C(\alpha) \prod_{k=1}^K \pi_k^{\alpha/K-1} = \log C(\alpha) + \log\left(\sum_{k=1}^K \pi_k^{\alpha/K-1}\right) \\ &= C_\pi + (\alpha/K-1)\sum_{k=1}^K \log \pi_k \\ \text{where: } C_\pi &= \log C(\alpha)\end{aligned}$$

Thus, the joint-likelihood is:

$$
\begin{aligned}
\log p(\{x_n\}, \{\pi_k\},\{\lambda_k\}|a, b, \alpha, K) = \sum_n^N \log \sum_k^K \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k) \\
+ \sum_k \left(C_\lambda + \log\left((a-1)\lambda_k\right) - b\lambda_k\right) \\
+ C_\pi + (\alpha/K - 1) \sum_{k=1}^K \log \pi_k
\end{aligned}
$$

(g) Find the expression for λ_k that maximizes the log-joint.
Solution:
Finding the partial derivative, using Equation 1 for the loglikelihood partial derivative w.r.t. λ_k:

$$
\frac{\partial}{\partial \lambda_k} \log p(\{x_n\}, \{\pi_k\}, \{\lambda_k\}|a, b, \alpha, K) = \sum_n^N r_{nk} \left(x_n \lambda_k^{-1} - 1\right) + \frac{a-1}{\lambda_k} - b
$$

Set derivative to 0 and solve for λ_k

$$
\begin{aligned}
\sum_n^N r_{nk} \left(x_n \lambda_k^{-1} - 1\right) + \frac{a-1}{\lambda_k} - b = 0 \\
\lambda_k \left(\sum_n^N r_{nk} + b \right) &= \sum_n^N r_{nk} x_n + a - 1 \\
\lambda_k &= \frac{\sum_n^N r_{nk} x_n + a - 1}{N_k + b}
\end{aligned}
$$

(h) Find the expression for π_k that maximizes the log-joint.
Solution:
Finding the partial derivative, using Equation 2 for the loglikelihood partial derivative w.r.t. π_k:

$$\frac{\partial}{\partial \pi_k} \log p(\{x_n\}, \{\pi_k\}, \{\lambda_k\}|a, b, \alpha, K) = \sum_n^N r_{nk} + \pi_k \mu + \alpha/K - 1$$

Set derivative to 0 and solve for π_k

$$\sum_n^N r_{nk} + \mu \pi_k + \alpha/K - 1 = 0$$

$$\begin{aligned}
\pi_k &= -\frac{N_k + \alpha/K - 1}{\mu} \\
\sum_k \pi_k &= -\frac{\sum_k (N_k + \alpha/K - 1)}{\mu} \\
1 &= -\frac{N + \alpha - K}{\mu} \\
\mu &= -N + K - \alpha
\end{aligned}$$

Therefore:

$$\pi_k = -\frac{N_k + \alpha/K - 1}{-N + K - \alpha} = \frac{N_k + \alpha/K - 1}{N - K + \alpha}$$

(i) Write down an iterative algorithm using the above update equations (similar to the ones derived in class for the Mixture of Gaussians); include initialization and convergence check steps.
Solution:

(a) Initialize hyperparameters a, b, α, K.

(b) Initialize $\boldsymbol{\theta}$ ($\boldsymbol{\lambda}$, $\boldsymbol{\pi}$) according to random assignment of data to distribution k

(c) Repeat until convergence ($\Delta \mathcal{L} < \epsilon$)

a) E-step

for all i $\in$ N **do**

for all k $\in$ K **do**

$$r_{nk} = \frac{\pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)}{\sum_j^K \pi_j \frac{1}{x_n!} \lambda_j^{x_n} \exp(-\lambda_j)}$$

end for

end for

b) M-step

for all i $\in$ N **do**

for all k $\in$ K **do**

$\pi_k = \frac{N_k+1-\alpha/K}{N+K-\alpha}$

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

end for

end for

b) Compute $\mathcal{L}$ and $\Delta\mathcal{L}$