Machine Learning 1 - Practice exercise 4

1 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 $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.

Solutions

$$\mathsf{likelihood} = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \tfrac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)$$

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

Solutions

$$L = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \tfrac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k)$$

(c) Find the expression for the responsibilities r_{nk}.

Solutions

$$r_{nk} = \frac{\pi_k P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)}$$

(d) Find the expression for λ_k that maximizes the log-likelihood.

Solutions

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_k} &= \sum_n \pi_k \frac{\partial P(x_n|\lambda_k)/\partial \lambda_k}{\sum_l \pi_l P(x_n|\lambda_l)} = 0 \\
\partial P(x_n|\lambda_k)/\partial \lambda_k &= P(x_n|\lambda_k)(x_n \lambda_k^{-1} - 1) \\
\frac{\partial L}{\partial \lambda_k} &= \sum_n r_{nk}(x_n \lambda_k^{-1} - 1) = 0 \\
\sum_n r_{nk} x_n \lambda_k^{-1} &= \sum_n r_{nk} \\
\lambda_k &= \sum_n r_{nk} x_n / N_k \\
N_k &= \sum_n r_{nk}
\end{aligned}$$

(e) Find the expression for π_k that maximizes the log-likelihood.

Solutions

Constraint: $\sum_k \pi_k = 1$.

$$
\begin{aligned}
L &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \frac{1}{x_n!} \lambda_k^{x_n} \exp(-\lambda_k) + \beta(\sum_k \pi_k - 1) \\
\frac{\partial L}{\partial \pi_k} &= \sum_n \frac{P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)} + \beta = 0 \\
&= \sum_n r_{nk} + \pi_k \beta = 0 \\
\pi_k &= -\sum_n r_{nk}/\alpha \\
\sum_k \pi_k &= -\sum_k \sum_n r_{nk}/\beta \\
\beta &= -N \\
\pi_k &= \sum_n r_{nk}/N \\
r_{nk} &= \frac{\pi_k P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)}
\end{aligned}
$$

(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
$\log p(\mathbf{x}_1,\ldots,\mathbf{x}_N,\{\pi_k\},\{\lambda_k\}|a,b,\alpha,K)$. You can collect all the terms that do not depend on the data or the parameters as a constant C.

Solutions

$$
\begin{aligned}
\hat{L} &= L + \sum_k \log \mathcal{G}(\lambda_k|a,b) + \log \mathcal{D}(\{\pi_k\}|\alpha,K) \\
&= L + \sum_k (a-1)\log\lambda_k - b\lambda_k + \sum_k (\alpha/K - 1)\log\pi_k + C
\end{aligned}
$$

where C is a constant that does not depend on the data or the parameters (i.e. the log normalizing constants).

(g) Find the expression for λ_k that maximizes the log-joint.

Solutions

$$\begin{aligned}
\frac{\partial \hat{L}}{\partial \lambda_k} &= \sum_n r_{nk}(x_n \lambda_k^{-1} - 1) + (a-1)\lambda_k^{-1} - b \\
&= \lambda_k^{-1}(\sum_n r_{nk} x_n + a - 1) - (\sum_n r_{nk} + b) \\
\lambda_k &= \frac{\sum_n r_{nk} x_n + a - 1}{N_k + b}
\end{aligned}$$

(h) Find the expression for π_k that maximizes the log-joint.

Solutions

$$\begin{aligned}
\frac{\partial \hat{L}}{\partial \pi_k} &= \sum_n \frac{P(x_n|\lambda_k)}{\sum_l \pi_l P(x_n|\lambda_l)} + \beta + (\alpha/K - 1)\pi_k^{-1} = 0 \\
&= \sum_n r_{nk} + \pi_k \beta + \alpha/K - 1 = 0 \\
\pi_k &= -\frac{\sum_n r_{nk} + \alpha/K - 1}{\beta} \\
\sum_k \pi_k &= -\frac{\sum_k \sum_n r_{nk} + \alpha - K}{\beta} \\
\beta &= -(N + \alpha - K) \\
\pi_k &= \frac{\sum_n r_{nk} + \alpha/K - 1}{N + \alpha - K}
\end{aligned}$$

(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.

Solutions

(a) Randomly assign data to K clusters (i.e. set hard values for r_{nk}). Compute π_k and λ_k from the initial assignments.

(b) Update (soft) r_{nk} (E-step).

(c) Update π_k using MLE or MAP estimates from above (M-step for π_k).

(d) Update λ_k using MLE or MAP estimates from above (M-step for λ_k).

(e) Compute L (or $\hat{L}$) and repeat b-c-d until $\Delta L < \epsilon$ (or $\Delta \hat{L} < \epsilon$).

2 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 a 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 $\hat{\mathbf{X}}$ with column n given by $\hat{\mathbf{x}}_n$.

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

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

Step 4 Pick the eigenvectors $\{\mathbf{u}_1, \ldots \mathbf{u}_K\}$ with the largest eigenvalues.

Step 5 Project the data onto the K-dimensional manifold spanned by the eigenvectors of step 4.

Answer the following questions:

(a) Provide an expression for $\hat{\mathbf{x}}_n$ in terms of the original $\mathbf{x}_n$.

Solutions

$$\begin{aligned} \hat{\mathbf{x}}_n &= \mathbf{x}_n - \frac{1}{N}\sum_m \mathbf{x}_m && (1)\\ &= \mathbf{x}_n - \bar{\mathbf{x}} && (2) \end{aligned}$$

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

Solutions

$$\begin{aligned} \sum_n \hat{\mathbf{x}}_n &= \sum_n (\mathbf{x}_n - \bar{\mathbf{x}}) && (3)\\ &= \sum_n \mathbf{x}_n - N\bar{\mathbf{x}} && (4)\\ &= \sum_n \mathbf{x}_n - N\frac{1}{N}\sum_m \mathbf{x}_m && (5)\\ &= \sum_n \mathbf{x}_n - \sum_m \mathbf{x}_m && (6)\\ &= \mathbf{0} && (7) \end{aligned}$$

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

Solutions

$$\begin{aligned} S &= \frac{1}{N}\sum_n \hat{\mathbf{x}}_n \hat{\mathbf{x}}_n^T && (8)\\ &= \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T && (9) \end{aligned}$$

(d) What is the shape of $\mathbf{S}$?

Solutions

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, $\mathbf{y}_n = \mathbf{L}\hat{\mathbf{x}}_n$, such that it has zero mean and identity covariance. Prove that the average over N of $\mathbf{y}_n$ is $\mathbf{0}$. Prove that the covariance of $\mathbf{y}_n$ is the identity. What is this operation called?

Solutions

- The full projection is $\mathbf{y}_n = \Lambda^{-1/2}\mathbf{U}^T\hat{\mathbf{x}}_n$, i.e. $\mathrm{L} = \Lambda^{-1/2}\mathbf{U}^T$.
- $\sum \mathbf{y}_n = \mathbf{L}\sum_n \hat{\mathbf{x}}_n$; since $\sum_n \hat{\mathbf{x}}_n = 0$, $\sum \mathbf{y}_n = \mathbf{0}$.
- Examine the covariance between the ith and jth projection.

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

  In the last line, the result is a dot product between a vector of all zeros except λ_i at index i, with a vector of zeros except 1 at index j. Therefore: for $i \neq j$, $C_{ij} = 0$; for $i = j$, $C_{ij} = \frac{\lambda_i}{\lambda_i^{1/2}\lambda_i^{1/2}} = 1$.

- Whitening or sphering.