

Machine Learning: Exercises for Block V

Unsupervised Learning

Isabel Valera

Exercise 1: Weights for clustering

(Exercise 14.1 [2])

Show that weighted Euclidean distance

$$d_e^{(w)}(x_i, x_{i'}) = \frac{\sum_{j=1}^d w_j (x_{ij} - x_{i'j})^2}{\sum_{j=1}^d w_j}$$

satisfies

$$d_e^{(w)}(x_i, x_{i'}) = d_e(z_i, z_{i'}) = \sum_{j=1}^d (z_{ij} - z_{i'j})^2,$$

where

$$z_{ij} = x_{ij} \cdot \left(\frac{w_j}{\sum_{j=1}^d w_j} \right)^{1/2}.$$

Thus weighted Euclidean distance based on x is equivalent to unweighted Euclidean distance based on z .

Exercise 2: Mixture model

(Exercise 14.2 [2])

Consider a mixture model density in d -dimensional feature space,

$$p(x) = \sum_{k=1}^K \pi_k p_k(x),$$

where $p_k = \mathcal{N}(\mu_k, I_d \cdot \sigma^2)$ and $\pi_k \geq 0 \forall k$ with $\sum_k \pi_k = 1$. Here $\{\mu_k, \pi_k\}, k = 1, \dots, K$ and σ^2 are unknown parameters.

Suppose we have data $x_1, x_2, \dots, x_N \sim p(x)$ and we wish to fit the mixture model.

1. Write down the log-likelihood of the data.
2. (Optional) Derive an EM algorithm for computing the maximum likelihood estimates. coincides with K -means clustering.

Exercise 3: Gaussian Mixture Model

(Exercise 9.3 [1])

Consider a Gaussian mixture model in which the marginal distribution $p(\mathbf{z})$ for the K -dimensional binary latent variable \mathbf{z} (having a 1-of- K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0) is given by

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

with mixing coefficients π_k and the conditional distribution $p(\mathbf{x} | \mathbf{z})$ for the observed variable is given by

$$p(\mathbf{x} | \mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

Show that the marginal distribution $p(\mathbf{x})$, obtained by summing $p(\mathbf{z})p(\mathbf{x} | \mathbf{z})$ over all possible values of \mathbf{z} , is a Gaussian mixture of the form

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x} | \mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

with latent variables \mathbf{Z} .

Exercise 4: PCA

(Exercise 12.1 [1])

In this exercise, we use proof by induction to show that the linear projection onto an M -dimensional subspace that maximizes the variance of the projected data is defined by the M eigenvectors of the data covariance matrix $\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$, corresponding to the M largest eigenvalues. Here we consider a data set of observations $\{\mathbf{x}_n\}$ where $n = 1, \dots, N$ with $\bar{\mathbf{x}}$ being the sample set mean given by $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$. Suppose the results holds for some general value of M and show that it consequently holds for dimensionality $M + 1$. The idea for the induction step is as follows, first set the derivative of the variance of the projected data with respect to a vector \mathbf{u}_{M+1} defining the new direction in data space equal to zero. This should be done subject to the constraints that \mathbf{u}_{M+1} be orthogonal to the existing vectors $\mathbf{u}_1, \dots, \mathbf{u}_M$, and also that it be normalized to unit length. Use Lagrange multipliers to enforce these constraints. Then make use of the orthonormality properties of the vectors $\mathbf{u}_1, \dots, \mathbf{u}_M$ to show that the new vector \mathbf{u}_{M+1} is an eigenvector of \mathbf{S} . Finally, show that the variance is maximized if the eigenvector is chosen to be the one corresponding to eigenvector λ_{M+1} where the eigenvalues have been ordered in decreasing value. Hint: In Section 21.1 [1] the result was proven for $M = 1$.

Exercise 5: (Linear) PCA

(Exercise 12.27 [1])

Show that the conventional linear PCA algorithm is recovered as a special case of kernel PCA if we choose the linear kernel function given by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$

Exercise 6: Kernel PCA 1

(Exercise 14.16 [2])

Consider the kernel principal component procedure outlined in Section 14.5.4 [2]. Argue that the number M of principal components is equal to the rank of \mathbf{K} , which is the number of non-zero elements in \mathbf{D} . Show that the m th component \mathbf{z}_m (m th column of \mathbf{Z}) can be written (up to centering) as $z_{im} = \sum_{j=1}^N \alpha_{jm} K(x_i, x_j)$, where $\alpha_{jm} = u_{jm}/d_m$. Show that the mapping of a new observation x_0 to the m th component is given by $z_{0m} = \sum_{j=1}^N \alpha_{jm} K(x_0, x_j)$.

Exercise 7: Kernel PCA 2

(Exercise 14.17 [2])

Show that with first principal component function $g_1(x) = \sum_{j=1}^N c_j K(x, x_j)$, the solution to

$$\max_{g_1 \in \mathcal{H}_K} \text{Var}_{\mathcal{T}} g_1(X) \text{ subject to } \|g_1\|_{\mathcal{H}_K} = 1$$

is given by $\hat{c}_j = u_{j1}/d_1$.

We denote principal component functions $g_m \in \mathcal{H}_K$, with \mathcal{H}_K the reproducing kernel Hilbert space generated by K ; $\text{Var}_{\mathcal{T}}$ refers to the sample variance over training data \mathcal{T} . Here \mathbf{u}_1 is the

first column of \mathbf{U} in

$$\tilde{\mathbf{K}} = (\mathbf{I} - \mathbf{M})\mathbf{K}(\mathbf{I} - \mathbf{M}) = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$$

with principal components variables \mathbf{Z} of a data matrix \mathbf{X} , inner-product (Gram) matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{M} = \mathbf{1}\mathbf{1}^T/N$. Further, d_1 is the first diagonal element of \mathbf{D} . Show that the second and subsequent principal component functions are defined similarly. Hint: See Section 5.8.1 [2].

Exercise 8: (Optional) EM algorithm

(Exercise 9.4 [1])

Suppose we wish to use the EM algorithm to maximize the posterior distribution over parameters $p(\boldsymbol{\theta} | \mathbf{X})$ for a model containing latent variables, where \mathbf{X} is the observed data set. Show that the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized is given by $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$ where $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ is defined by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$$

for some general parameter value $\boldsymbol{\theta}$. The current estimate for the parameters is denoted $\boldsymbol{\theta}^{\text{old}}$.

Exercise 9: (Optional) EM for PCA

(Exercise 12.15 [1])

Derive the M-step equations

$$\begin{aligned} \mathbf{W}_{\text{new}} &= \left[\sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}}) \mathbb{E}[\mathbf{z}_n]^T \right] \left[\sum_{n=1}^N \mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] \right]^{-1} \\ \sigma_{\text{new}}^2 &= \frac{1}{ND} \sum_{n=1}^N \left\{ \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 - 2\mathbb{E}[\mathbf{z}_n]^T \mathbf{W}_{\text{new}}^T (\mathbf{x}_n - \bar{\mathbf{x}}) \right. \\ &\quad \left. + \text{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] \mathbf{W}_{\text{new}}^T \mathbf{W}_{\text{new}}) \right\} \end{aligned}$$

for the probabilistic PCA model by maximization of the expected complete-data log likelihood function given by

$$\begin{aligned} \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2)] &= - \sum_{n=1}^N \left\{ \frac{D}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \text{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T]) \right. \\ &\quad \left. + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} \mathbb{E}[\mathbf{z}_n]^T \mathbf{W}^T (\mathbf{x}_n - \boldsymbol{\mu}) \right. \\ &\quad \left. + \frac{1}{2\sigma^2} \text{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] \mathbf{W}^T \mathbf{W}) \right\}. \end{aligned}$$

Notation: We denote the dataset as $\mathbf{X} = \{\mathbf{x}_n\}$, the $D \times M$ weight matrix as \mathbf{W} and the n th row of the matrix \mathbf{Z} as \mathbf{z}_n . The predictive distribution $p(\mathbf{x})$ is governed by the parameters $\boldsymbol{\mu}$, \mathbf{W} , and σ^2 , where $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\sigma^2 \in \mathbb{R}$.

References

- [1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.
- [2] J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*. Springer series in statistics New York, 2001.