# Exercises unsupervised learning

## 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 $z$.

Solution:

$$d_e(z_{i'}, z_{i'}) = \sum_{j=1} (z_{ij} - z_{i'j})^2 = \sum_{j=1} \left[ x_{ij} \left( \frac{w_j}{\sum_{j=1}^{d} w_j} \right)^{1/2} - x_{i'j} \left( \frac{w_j}{\sum_{j=1}^{d} w_j} \right)^{1/2} \right]^2 =$$

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

## 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 $\mathbf{S} = \frac{1}{\mathbf{N}} \sum_{\mathbf{n=1}}^{\mathbf{N}} (\mathbf{x_n} - \overline{\mathbf{x}}) (\mathbf{x_n} - \overline{\mathbf{x}})^{\mathrm{T}}$, corresponding to the $M$ largest eigenvalues. Here we consider a data set of observations $\{\mathrm{x}_n\}$ where $n = 1, \ldots, N$ with $\overline{\mathbf{x}}$ being the sample set mean given by $\overline{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$. 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 $\mathbf{u}_1, \ldots, \mathbf{u}_M$, 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 $\mathbf{u}_1, \ldots, \mathbf{u}_M$ 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 $\lambda_{M+1}$ where the eigenvalues have been ordered in decreasing value. Hint: In Section 21.1 [1] the result was proven for $M = 1$.

Solution:

Solution:

Assume that it holds for a general $M$.

That is, the $M$ eigenvectors $u_1, u_2, \dots, u_M$ associated with the $M$ biggest eigenvalues, form a basis that projects the data to an $M$-dimensional space that maximizes

$$Tr(U^T X^T X U) \quad \text{where } X = [x_1\ x_2 \cdots x_N]^T,\ U = [u_1\ u_2 \cdots u_M]$$

and we assume that $\frac{1}{N}\sum_n x_n = \bar{x} = \vec{0}$ without loss of generality.

We look for a vector $v$ that maximizes

$$Tr([U\ v]^T X^T X [U\ v]) \quad \text{subject to } \|v\| = 1 \text{ and } \langle v, u_k \rangle = 0\ \forall k.$$

$$Tr([U\ v]^T X^T X [U\ v]) = \sum_{kijn} u_{ki} x_{ni} x_{nj} u_{kj} + \sum_{ijn} v_i x_{ni} x_{nj} v_j$$

The problem is

$$\max_v\ v^T X^T X v$$
$$\text{s.t}\quad \|v\| = 1$$
$$v^T u_k = 0 \quad k = 1, 2, \dots, M$$

Using Lagrange multipliers: $L = v^T \underbrace{X^T X}_{S} v + \lambda_{M+1}(1 - v^T v) + \sum_k \eta_k u_k^T v$

$$\frac{\partial L}{\partial v} = 2Sv - 2\lambda_{M+1} v + \sum_k \eta_k u_k = 0$$

And multiply on the left by $u_m^T$:

$$2u_m^T S v - 2\lambda_{M+1} u_m^T v + \sum_k \eta_k u_m^T u_k = 2u_m^T S v + \eta_m = 0$$

Because $S$ is symmetric $2u_m^T S v = 2v^T S u_m$ and therefore:

$$2v^T S u_m + \eta_m = 2v^T \lambda_m u_m + \eta_m = 2\lambda_m v^T u_m + \eta_m = \eta_m$$

So that $\eta_1 = \eta_2 = \cdots = \eta_M = 0$. and substituting into $\frac{\partial L}{\partial v}$ we obtain

$$\frac{\partial L}{\partial v} = 0 \Rightarrow \quad Sv = \lambda_{M+1} v$$

Therefore $v$ must be an eigenvector with eigenvalue $\lambda_{M+1}$, and so the projected variance in the direction of $v$, $v^T S v = \lambda_{M+1}$ is maximized by choosing $v$ as the (unit) eigenvector associated with the largest eigenvalue that has not been picked yet.

## 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 $m$ the component is given by $z_{0m} = \sum_{j=1}^{N} \alpha_{jm} K(x_0, x_j)$.

Solution:

For simplicity, let us assume that the projected data set has zero mean, so that $K = \frac{1}{N} \sum_n \phi(x_n)\phi(x_n)^T = \phi(x)^T\phi(x)$. Note that $k_{ij} = k(x_i, x_j)$.

We seek the eigenvectors of $K$, $Ku_k = \lambda_k u_k$, with eigen-decomposition of $K = UD^2U^T$ with $D$ diagonal and $U$ orthogonal.
For a square matrix, the number of eigenvalues is equal to the rank of that matrix since rank($K$) = no. of non-zero values in $D$ as $|\det(U)| = 1$.
Therefore number eigenvalues = rank($K$).

In matrix form, eigenvectors $Z = UD$ hold $KZ = UD^2U^TUD = UD^3 = ZD^2$

$$KZ = KUD = \left(\sum_k k(x_i, x_k) u_{kj} d_j\right)_{ij} = \left(z_{ij} d_j^2\right)_{ij}$$

So the $i$-th component of the $m$-th eigenvector, $z_{im}$, is written as

$$z_{im} = \sum_k k(x_i, x_k) \frac{u_{km}}{d_m} = \sum_j k(x_i, x_j) \alpha_{jm} \quad \text{with } \alpha_{jm} = \frac{u_{jm}}{d_m}.$$

Note that for any $z_m$ we have $Kz_m = d_m z_m = \frac{1}{N} \sum_n \phi(x_n)\phi(x_n)^T z_m d_m$

Therefore $z_m = \frac{d_m}{N} \sum_n \phi(x_n) \left[\phi(x_n)^T z_m\right] = \sum_n \phi(x_n) \alpha_{nm}$ and

$$\phi(x_0)^T z_m = \sum_n \phi(x_0)^T \phi(x_n) \alpha_{nm} = \sum_n k(x_0, x_n) \alpha_{nm}.$$