Leren — Homework 3

Chapter 6-8, Alpaydin

Tim Stolp (11848782)

Deadline: 23:59, Sunday, December 2nd, 2018

This is the fifth week's assignment for Leren. This assignment covers chapters 6, 7 & 8 of Alpaydin. Please take note of the following:

- You are expected to hand in your solutions in LaTeX;
- This problem set is an individual assignment;
- The deadline for this assignment is Sunday, December 2 at 23:59.

1 Chapter 6: Dimensionality Reduction

1.1 Principal Component Analysis (PCA)

Suppose we have a dataset of N vectors $\{\boldsymbol{x}^i\}_{i=1}^N$ of dimension D. We can represent the entire dataset as a N by D matrix $\boldsymbol{X}$ (row i is $\boldsymbol{x}^i$). We wish to project this data onto a lower-dimensional subspace and choose PCA as our technique for dimensionality reduction. Consider the following procedure for PCA:

Step 1 Data preprocessing to get $\tilde{\boldsymbol{X}}$ from $\boldsymbol{X}$

Step 2 Compute the sample covariance $\boldsymbol{C}$ of $\tilde{\boldsymbol{X}}$, i.e. compute the (unbiased) estimator of $\boldsymbol{\Sigma} = \text{cov}(\tilde{\boldsymbol{X}})$.

Step 3 Solve the eigenvalue problem $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^T$ (spectral decomposition of $\boldsymbol{C}$), where $\boldsymbol{V}$ is a column matrix of eigenvectors $\boldsymbol{v}_k$ and $\boldsymbol{L}$ is a diagonal matrix of eigenvalues λ_k, i.e. $\boldsymbol{L}_{kl} = \lambda_k \delta_{kl}$, where $\delta_{kl} = 1$ if $k = l$, otherwise $\delta_{kl} = 0$.

Step 4 Pick the K eigenvectors with the K largest eigenvalues $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}$.

Step 5 Project data onto K-dimensional subspace.

Answer the following questions:

(a) Let's assume that the variance of each of the D variables is approximately the same such that rescaling the variance is not required. What other data preprocessing step (**Step 1**) is required for PCA? Provide an expression on how this preprocessing can be done. Hint: you can give the expression for $\tilde{\boldsymbol{x}}^i$ (i.e. the i-th row of the preprocessed data $\tilde{\boldsymbol{X}}$).

To further pre-process the data we have to center the data around the origin. This is done by substracting the mean vector from each vector in $\tilde{\boldsymbol{X}}$.
$\tilde{\boldsymbol{x}}^i = x^i - \bar{x}$

(b) What is the dimensionality of $\boldsymbol{C}$?

D

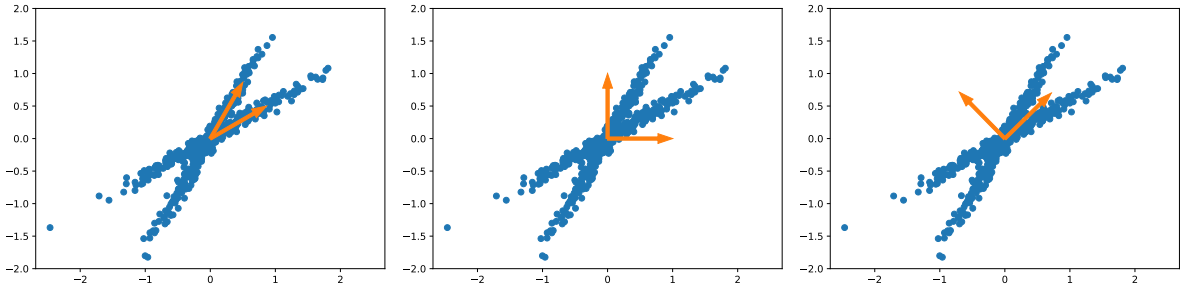
(c) How do we construct $\boldsymbol{W}$? Reminder: $\boldsymbol{W}$ is the linear projection that maps data vectors $\tilde{\boldsymbol{x}}^i$ onto the K-dimensional principal subspace, i.e. $\boldsymbol{z}^i = \boldsymbol{W}^T \tilde{\boldsymbol{x}}^i$.

The k columns of $\boldsymbol{W}$ are constructed by taking the k leading eigenvectors of $\boldsymbol{V}$.

(d) How do we obtain the PCA reconstruction $\hat{\boldsymbol{x}}^i$ of $\boldsymbol{x}^i$ from the projected data point $\boldsymbol{z}^i$? Provide an expression.

$\hat{\boldsymbol{x}}^i = \boldsymbol{z}^i \boldsymbol{V}^{\boldsymbol{T}}$

(e) Consider the three different proposals of principal components for the same 2-dimensional dataset in Figure 1 (principal components displayed in orange). Identify the true principal components that would be found by applying PCA. For each proposal briefly explain your decision why these are (not) the correct principal components.



Option 1 Option 2 Option 3

Figure 1: Proposals of principal components on a 2D dataset

Option 1 is wrong because the eigenvectors are not orthogonal.
Option 2 is wrong because the eigenvectors don't point in the direction with the biggest variance.
Option 3 is correct because it points in the direction of the biggest variance and they are orthogonal.

(f) Consider the data in Figure 1. Which of the following set of eigenvalues corresponds to the true principal components? Briefly explain your reasoning on each of the possible eigenvalue pairs.

- $l_1 = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}$
- $l_2 = \begin{bmatrix} 0.56 \\ 0.04 \end{bmatrix}$
- $l_3 = \begin{bmatrix} 0.46 \\ -0.14 \end{bmatrix}$

l_1 is wrong because these values would mean that the spread of the data is the same in both directions, which it is not.

l_2 is correct because these values correspond to the right spread of data
l_3 is wrong because the second value can't be negative because the covariance matrixis a positive semidefinite matrix

(g) Suppose we project the data from Figure 1 to the principal component with the largest eigenvalue. What will be the dimension of $\boldsymbol{W}$?

1

1.2 Singular Value Decomposition (SVD)

In this exercise, we want to draw the connection between Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). Let $\tilde{\boldsymbol{X}}$ be the centered data matrix of size $N \times D$. SVD factorizes $\tilde{\boldsymbol{X}}$ into three components $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ where $\boldsymbol{U}$ is a $N \times N$ orthogonal matrix, $\boldsymbol{S}$ is a $N \times D$ rectangular diagonal matrix and $\boldsymbol{V}$ is a $D \times D$ orthogonal matrix. An orthogonal matrix $\boldsymbol{Q}$ has the property $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix.

(a) The sample covariance matrix $\boldsymbol{C}$ is given by $\boldsymbol{C} = \frac{\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}}{N-1}$ where we apply element-wise division by the scalar $(N-1)$. Show that $\boldsymbol{C}$ can be decomposed as $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^T$ where $\boldsymbol{V}$ is the same orthogonal matrix as in the SVD of $\tilde{\boldsymbol{X}}$.

$\boldsymbol{C} = \frac{\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}}{N-1} = \frac{\boldsymbol{V}\boldsymbol{S}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T}{N-1} = \frac{\boldsymbol{V}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{V}^T}{N-1} = \boldsymbol{V}\frac{\boldsymbol{S}^T\boldsymbol{S}}{N-1}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^T$

(b) Using your result from (a), provide the expression for $\boldsymbol{L}$ (in terms of the decomposition factors of $\tilde{\boldsymbol{X}}$). What is the dimension of $\boldsymbol{L}$?

$\boldsymbol{L} = \frac{\boldsymbol{S}^T\boldsymbol{S}}{N-1}$
$\boldsymbol{L}$ is $\boldsymbol{N} \times \boldsymbol{D}$

(c) What are the eigenvectors and eigenvalues of $\boldsymbol{C}$? Hint: eigenvectors $\boldsymbol{v}$ and eigenvalues λ of the matrix $\boldsymbol{A}$ satisfy the equation $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$

The eigenvectors of $\boldsymbol{C}$ are the columns of $\boldsymbol{V}$.
The eigenvalues of $\boldsymbol{C}$ are the diagonals of $\boldsymbol{L}$.

2 Chapter 7: Clustering

For the next two problems, consider a dataset $\mathcal{X}_c = \{\mathbf{x}^i\}_{i=1}^N$ consisting of $N = 5$ data-points. The main goal of the next two problems is to cluster the dataset $\mathcal{X}_c$ reported in Figure **??** into $k = 2$ groups.

2.1 Agglomerative Hierarchical Clustering

Given a measure of distance between clusters d, the procedure to perform agglomerative hierarchical clustering proceeds as follows:

- Initialize one cluster for each data-point

$$\mathcal{C}^i = \{\mathbf{x}^i\} \qquad \forall \mathbf{x}^i \in \mathcal{X}_c$$

- While the total number of clusters is greater then k

1. Compute the distance between each pair of clusters according to d
2. Merge the two clusters $\mathcal{C}^i$ and $\mathcal{C}^j$ that have minimal distance

$$\mathcal{C}^{i,j} = \mathcal{C}^i \cup \mathcal{C}^j$$

For this problem, d will consider the minimum euclidean distance between the elements of the clusters. More formally, given two clusters $\mathcal{C}^i$ and $\mathcal{C}^j$, the distance between them is defined as:

$$d(\mathcal{C}^i, \mathcal{C}^j) = \min \left\{ \|\mathbf{y} - \mathbf{z}\|_2 : \mathbf{y} \in \mathcal{C}^i, \mathbf{z} \in \mathcal{C}^j \right\} \tag{1}$$

Where $\|\mathbf{y} - \mathbf{z}\|_2$ represents the euclidean distance between $\mathbf{y}$ and $\mathbf{z}$.

(a) Perform agglomerative hierarchical clustering on the dataset $\mathcal{X}_c$ for $k = 2$. Which datapoints are part of the final two clusters? Show the iterative procedure by specifying which clusters are merged at each step.
(The figure is left out because it crashed overleaf (online LaTeX).)

First calculate all distances.

	x_1	x_2	x_3	x_4	x_5
x_1	0				
x_2	$\sqrt{2}$	0			
x_3	$\sqrt{13}$	$\sqrt{5}$	0		
x_4	$\sqrt{5}$	1	2	0	
x_5	$\sqrt{18}$	$\sqrt{8}$	1	$\sqrt{5}$	0

The smallest distances are between x_2 and x_4, and x_3 and x_5 so we combine those together. The distances of the new clusters to the old clusters is the smallest value of the combined old clusters.

	x_1	x_2x_4	x_3x_5
x_1	-		
x_2x_4	$\sqrt{2}$	-	
x_3x_5	$\sqrt{13}$	2	-

Same steps. $\sqrt{2}$ is smallest so combine x_2x_4 and x_1

	$x_1x_2x_4$	x_3x_5
$x_1x_2x_4$	-	
x_3x_5	2	-

Then we are left with 2 clusters (K=2): $x_1x_2x_4$ and x_3x_5

(b) Consider a distance d_2 defined as the minimal squared euclidean distance between elements of two clusters:

$$d_2(\mathcal{C}^i, \mathcal{C}^j) = \min \left\{ \|\mathbf{y} - \mathbf{z}\|_2^2 : \mathbf{y} \in \mathcal{C}^i, \mathbf{z} \in \mathcal{C}^j \right\}$$

What would change if we perform the clustering procedure considering d_2 instead of d? Justify your answer.

Nothing would change because even after squaring the distances the smallest stays the smallest. The ratios don't change.

2.2 K-Means Clustering

For this problem, let $\mathbf{b}$ be an indicator vector for data point $\mathbf{x}^i$ such that $b_i = 0$ if $\mathbf{x}^i$ is a member of the first cluster and 1 if $\mathbf{x}^i$ belongs to second one. Let $\mathbf{m}_1$ and $\mathbf{m}_2$ be the means of the two clusters respectively. The most common form of the K-Means algorithm proceeds as follows

- Initialize $\mathbf{m}_1$ and $\mathbf{m}_2$.
- Repeat the following two steps until the values of $\mathbf{m}_1$ and $\mathbf{m}_2$ are unchanged with respect to the previous iteration

  1. Determine clusters for each sample $\mathbf{x}^i$ by determining labels b_i.

$$b_i = \begin{cases} 0, & ||\mathbf{x}^i - \mathbf{m}_1||^2 \leq ||\mathbf{x}^i - \mathbf{m}_2||^2 \text{ i.e. } \mathbf{x}^i \text{ belongs to cluster 1} \\ 1, & \text{otherwise} \end{cases}$$

  2. Recompute $\mathbf{m}_1$ and $\mathbf{m}_2$ using the updates

$$\mathbf{m}_1 = \frac{\sum_{i=1}^{N}(1-b_i)\mathbf{x}^i}{N - \sum_{i=1}^{N} b_i} \qquad \mathbf{m}_2 = \frac{\sum_{i=1}^{N} b_i \mathbf{x}^i}{\sum_{i=1}^{N} b_i}$$

(a) Perform the K-means clustering algorithm on the dataset $\mathcal{X}_c$ (Figure **??**) by initializing $\mathbf{m}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{m}_2 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$. Show each iteration of the algorithm by specifying the updates for the values of the assignments $\mathbf{b}$ and the means $\mathbf{m}_1$ and $\mathbf{m}_2$ until convergence.

Calculate $||\mathbf{x}^i - \mathbf{m}_1||^2$ and $||\mathbf{x}^i - \mathbf{m}_2||^2$ b_i for each x^i and then choose b_i.

	$\|\|\mathbf{x}^i - \mathbf{m}_1\|\|^2$	$\|\|\mathbf{x}^i - \mathbf{m}_2\|\|^2$	b_i
x_1	5	10	0
x_2	1	4	0
x_3	2	5	0
x_4	2	1	1
x_5	5	4	1

Then update the means:

$$\mathbf{m}_1 = \frac{x_1+x_2+x_3}{5-2} = \frac{\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix}}{3} = \begin{bmatrix} 2 \\ 2\frac{1}{3} \end{bmatrix}$$

$$\mathbf{m}_2 = \frac{x_4+x_5}{2} = \frac{\begin{bmatrix} 3 \\ 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 4 \end{bmatrix}}{2} = \begin{bmatrix} 3\frac{1}{2} \\ 3 \end{bmatrix}$$

Then update the table:

	$\|\|\mathbf{x}^i - \mathbf{m}_1\|\|^2$	$\|\|\mathbf{x}^i - \mathbf{m}_2\|\|^2$	b_i
x_1	$2\frac{7}{9}$	$10\frac{1}{4}$	0
x_2	$\frac{1}{9}$	$3\frac{1}{4}$	0
x_3	$3\frac{7}{9}$	$1\frac{1}{4}$	1
x_4	$1\frac{1}{9}$	$1\frac{1}{4}$	0
x_5	$6\frac{7}{9}$	$\frac{1}{4}$	1

Update means:

$$\mathbf{m}_1 = \frac{x_1+x_2+x_4}{5-2} = \frac{\begin{bmatrix}1\\1\end{bmatrix}+\begin{bmatrix}2\\2\end{bmatrix}+\begin{bmatrix}3\\2\end{bmatrix}}{3} = \begin{bmatrix}2\\1\frac{2}{3}\end{bmatrix}$$

$$\mathbf{m}_2 = \frac{x_3+x_5}{2} = \frac{\begin{bmatrix}3\\4\end{bmatrix}+\begin{bmatrix}4\\4\end{bmatrix}}{2} = \begin{bmatrix}3\frac{1}{2}\\4\end{bmatrix}$$

Update table:

	$\|\|\mathbf{x}^i - \mathbf{m}_1\|\|^2$	$\|\|\mathbf{x}^i - \mathbf{m}_2\|\|^2$	b_i
x_1	$1\frac{4}{9}$	$15\frac{1}{4}$	0
x_2	$\frac{1}{9}$	$6\frac{1}{4}$	0
x_3	$6\frac{4}{9}$	$\frac{1}{4}$	1
x_4	$1\frac{1}{9}$	$4\frac{1}{4}$	0
x_5	$9\frac{4}{9}$	$\frac{1}{4}$	1

After this update you can see that the b_i values have not changed so we would get the same results for the means so there is no need to iterate further. x_1, x_2 and x_4 belong to cluster 1. x_3 and x_5 belong to cluster 2.

(b) Repeat the K-means clustering algorithm by initializing $\mathbf{m}_1 = \begin{bmatrix}1\\2\end{bmatrix}$ and $\mathbf{m}_2 = \begin{bmatrix}3\\3\end{bmatrix}$ instead. Do we obtain the same result? Explain your findings.

Same steps as (a) but with different initialized means:

	$\|\|\mathbf{x}^i - \mathbf{m}_1\|\|^2$	$\|\|\mathbf{x}^i - \mathbf{m}_2\|\|^2$	b_i
x_1	1	8	0
x_2	1	2	0
x_3	8	1	1
x_4	4	1	1
x_5	13	2	1

Update means:

$$\mathbf{m}_1 = \frac{x_1+x_2}{5-3} = \frac{\begin{bmatrix}1\\1\end{bmatrix}+\begin{bmatrix}2\\2\end{bmatrix}}{2} = \begin{bmatrix}\frac{3}{2}\\\frac{3}{2}\end{bmatrix}$$

$$\mathbf{m}_2 = \frac{x_3+x_4+x_5}{3} = \frac{\begin{bmatrix}3\\4\end{bmatrix}+\begin{bmatrix}3\\2\end{bmatrix}+\begin{bmatrix}4\\4\end{bmatrix}}{3} = \begin{bmatrix}\frac{10}{3}\\\frac{10}{3}\end{bmatrix}$$

Update table:

	$\|\|\mathbf{x}^i - \mathbf{m}_1\|\|^2$	$\|\|\mathbf{x}^i - \mathbf{m}_2\|\|^2$	b_i
x_1	$\frac{1}{2}$	$10\frac{8}{9}$	0
x_2	$\frac{1}{2}$	$3\frac{5}{9}$	0
x_3	$8\frac{1}{2}$	$\frac{5}{9}$	1
x_4	$2\frac{1}{2}$	$1\frac{8}{9}$	1
x_5	$12\frac{1}{2}$	$\frac{8}{9}$	1

The values of b_i have not changed so we are done iterating. The results are different this time because x_4 now belongs to cluster 2 instead of cluster 1.

3 Chapter 8: K-Nearest Neighbors

For this problem we will consider a dataset $\mathcal{X}_k = \{\mathbf{x}^i\}_{i=1}^N$ consisting of $N = 7$ data-points of two different classes. The goal of this exercise is to use the k-nearest neighbors algorithm to classify new observations. Figure 2 visualizes the examples $\{\mathbf{x}^i\}_{i=1}^N$ together with the instances $\mathbf{t}^1$ and $\mathbf{t}^2$ for which the label is unknown.

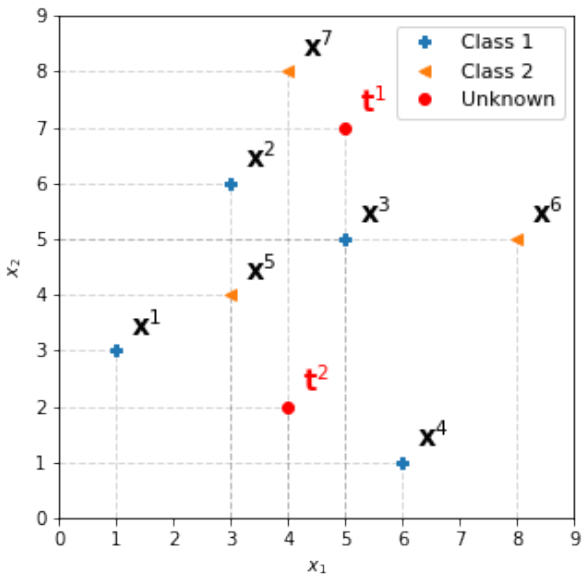


Figure 2: Two-class dataset $\mathcal{X}_k$.

In the most common version of the k-nearest neighbors algorithm, the class label of a new data-point is determined by the class assigned to the majority of its neighbors according to a specified distance measure. Answer the following questions:

(a) Consider the euclidean distance as the measure to determine the proximity of the data-points. Which is the class label associated to the test point $\mathbf{t}^1$ by considering $k = 1$, $k = 3$ and $k = 5$ neighbors? Justify your answers by specifying the neighbors of $\mathbf{t}^1$ for the different values of k.

Calculate the euclidean distances between the points and the test point $\mathbf{t}^1$ and write down each point's respective class:

	t_1	Class
x_1	5,66	1
x_2	2,24	1
x_3	2	1
x_4	6,08	1
x_5	3,61	2
x_6	3,61	2
x_7	1,41	2

For each k you pick the points with smallest distance to the test point.
For K = 1 that is x_7 so we choose class 2.
For K = 3 that are x_2, x_3, x_7 which are classes 1,1,2. The majority here is class 1 so we choose class 1.
For k = 5 that are x_2, x_3, x_5, x_6x_7 which are classes 1,1,2,2,2. The majority is class 2 so we choose class 2.

(b) The cosine similarity is a popular measure to compare correspondences between vectors. Based on this notion, it is possible to define a measure of distance known as the cosine distance. Given two data-points $\mathbf{x}^i$ and $\mathbf{x}^j$, the cosine distance is defined as:

$$d_{cos}(\mathbf{x}^i, \mathbf{x}^j) = 1 - \frac{(\mathbf{x}^i)^T \mathbf{x}^j}{\|\mathbf{x}^i\|_2 \|\mathbf{x}^j\|_2}$$

By using the cosine distance to determine the neighbors, report the class label assigned to the test point $\mathbf{t}^2$ for $k = 1$, $k = 3$ and $k = 5$. Specify the neighbors of $\mathbf{t}^2$ and its predicted label for each mentioned value of k.

Calculate the cosine similarities between the points and the test point $\mathbf{t}^2$ and write down each point's respective class:

	t_2	Class
x_1	0,293	1
x_2	0,200	1
x_3	0,051	1
x_4	0,044	1
x_5	0,106	2
x_6	0,005	2
x_7	0,200	2

Use same method as in question (a).

K = 1 gives x_6 which gives class 2.
K = 3 gives x_3, x_3, x_6 which gives class 1.
K = 5 gives x_3, x_4, x_5, x_6 and either x_2 or x_7, depending on which point you choose you get class 1 for x_2 and class 2 for x_7.

(c) What problem arises if we select $k = 7$ to classify $\mathbf{t}^1$ and $\mathbf{t}^2$? Can we solve the problem by choosing a different distance metric? Explain your reasoning.

With K = 7 you look at the class of all points. This will always give class 1 as the answer because there are more points in total in class 1. Choosing a different distance metric will not change anything because you are already taking every point into account.