Introduction to Machine Learning and Artificial Intelligence, Summer 2023 (ET1550)

Assignment 5

1. K-means Clustering

Consider a small dataset representing the measurements of products in an industrial setting. Let's choose (2.0, 1.0) and (5.0, 6.0) as the initial cluster centers. Find the final cluster centers using the k-means algorithm for this dataset.

X1	X2
2.0	1.0
2.5	2.2
1.8	1.8
5.0	6.0
5.5	7.0
4.5	5.5

2. Cost Function of K-Means Clustering

In the previous question,

a) What is the amount of cost function using the two computed cluster centers?
b) What is the amount of cost function for the same dataset having three cluster centers of (4.75, 5.75), (2.1, 1.67) and (5.5, 7)?

3. Principal Component Analysis (PCA)

A normalized data point with six features is given as,

$$\mathbf{x}^{\mathrm{T}} = [0.15, -0.66, 1.58, -1.17, 0.96, -0.86]$$

In addition, the eigenvectors (normalized) corresponding to the first three principal components are,

$$\mathbf{v}_1{}^{\mathrm{T}} = [0.35, 0.5, 0.45, 0.4, 0.2, 0.4]$$

$$\mathbf{v}_2{}^{\mathrm{T}} = [-0.4, 0.2, -0.1, 0.6, -0.45, 0.45]$$

$$\mathbf{v}_3{}^{\mathrm{T}} = [-0.1, -0.1, -0.3, -0.2, 0.8, -0.4]$$

Apply Principal Component Analysis (PCA) to reduce the dimensionality of the data point from 6 to 3 using the given eigenvectors.

4. Explained Variance Using Principal Components

Consider a dataset consisting of 329 data points, each having 9 features. The eigenvalues of the covariance matrix of this dataset have been given as,

$$S = \begin{bmatrix} 0.3775 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0511 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0279 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0230 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0168 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0120 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0085 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0039 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0018 \end{bmatrix}$$

We want to reduce the dimensions of this dataset using principal component analysis. Examine the eigenvalues to determine the minimum number of principal components that should be considered to keep at least 90% of the variance in the data.