Homework 10, due November 17th, 11:59pm

November 10, 2023

In this homework, you are required to include in your report the code that you implemented. If you use some code from the web or package, also mention in your report the origin of the code.

1. The data `data_clust` is a Matlab file containing 13000 observations $x_i \in \mathbb{R}^{640}$ and their labels y_i. The data can be loaded in Python using `scipy.io.loadmat` and there is a similar package to load the data into R. We will perform k-means clustering using different initialization methods and evaluate the clustering result.

a) Perform PCA on the data matrix X and plot the projected points on the first two PCs. Use a different color for each class. (1 point)

b) Perform k-means clustering using one random initialization. Compute the 10×10 contingency matrix, which is a 2D histogram of the $(y_i, \hat{y}_i)$ combinations. See `sklearn.metrics.cluster.contingency_matrix` for details. Display the obtained contingency matrix as a grayscale image. (1 point)

c) The clustering result is the same up to a permutation of the labels. We will assign the cluster labels to the true labels by finding a permutation of the labels that maximizes the sum of the diagonal elements on the resulting contingency matrix. For that we will solve the linear sum assignment problem, see `scipy.optimize.linear_sum_assignment` for details. Use the contingency matrix to solve the linear sum assignment, but be aware that the linear sum assignment performs a minimization, but we want to maximize the assignment of clusters to labels. Display the contingency matrix obtained after the permutation of the labels given by the linear sum assignment, which should be close to a diagonal matrix. (1 point)

d) Repeat points b) and c) five times with different random initializations and display the 5 obtained original and permuted contingency matrices. In each case report the accuracy score, which is the sum of the diagonal elements of the permuted contingency matrix divided by the number of observations. Also report the obtained Adjusted Rand Index (`sklearn.metrics.adjusted_rand_score`) in each case, which should be pretty close to the accuracy. (2 points)

e) Repeat point d) with five different k-means++ initializations (see the `sklearn.cluster.KMeans` documentation for details). (1 point)

f) Implement the method for selecting k centers furthest from each other from page 29 of the LearningGM slides. Repeat point d) with five different instances of this initialization method. (2 points)

g) Report in a table the average accuracy scores and Adjusted Rand index scores obtained at points d), e), f), where each average is computed from the five corresponding initializations. Which initialization method obtains the largest average accuracy score and which obtains the largest average Adjusted Rand Index? (1 point)