EEE-361 HOMEWORK 5

May 1, 2021

In this assignment you will investigate the performance of different kernel methods for k-means clustering. Recall that in k-means clustering the goal is to minimize the sum of distances of the data points from their nearest centroids. Following the terminology of the book, given the d-dimensional data points $P = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $k \in \mathbb{N}$, we want to find the partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of P that minimizes the sum of distances of all data points from their corresponding centroids, over all partitions, i.e.

$$\underset{\mathcal{P}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in P_i} \|\boldsymbol{x}_j - \boldsymbol{c}_i\|^2 \ ,$$

where the centroids $\boldsymbol{c}_i$ are formally defined as $\frac{1}{|P_i|} \sum_{\boldsymbol{x}_j \in P_i} \boldsymbol{x_j}$. If the distances are scaled, we are dealing with weighted k-means clustering. In this case the distances are $d(\boldsymbol{x}, \boldsymbol{c}_i) = w_i \|\boldsymbol{x} - \boldsymbol{c}_i\|^2$ and the centroids are defined as $\boldsymbol{c}_i = \frac{w_i}{|P_i|} \sum_{\boldsymbol{x} \in P_i} \boldsymbol{x}$.

Recall that if we know the dot products across all data points, we can find all possible distances. Therefore, a matrix $\boldsymbol{K}$ which captures the dot products would be enough. This matrix may be defined as a kernel matrix and data points are here mapped to a feature space in which the distances differ from those in the Euclidean space. The three kernels to compare are the following ones:

$$\begin{aligned} \textbf{Polynomial} \quad & \boldsymbol{K}_{il} = (\boldsymbol{x}_i \cdot \boldsymbol{x}_l + c)^d \\ \textbf{Gaussian} \quad & \boldsymbol{K}_{il} = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2 / 2\sigma^2) \\ \textbf{Sigmoid} \quad & \boldsymbol{K}_{il} = \tanh(c\boldsymbol{x}_i \cdot \boldsymbol{x}_l + \theta) \end{aligned}$$

Note that here the overall sum to be minimized is as given in equation (18) in page 253 of the book.

In this homework we will test the classification performance of k-means clustering technique in the classification of handwritten digits in the MNIST database which can be found in the following link: `http://yann.lecun.com/exdb/mnist/`. In this database, there are training and test data with corresponding labels in separate files.

First, on the training data investigate how well you can identify the number of classes by using standard k-means and kernel based k-means algorithms. Then, for $k = 10$, try to choose the number of iterations until termination and the parameters of your kernel based approaches to obtain a high accuracy of classification on the training data. Support your findings with relevant plots. Report the set of parameters and number of iterations until termination as well as the classification accuracy on the training data. Then, use those k-means classifiers on the test data and report the classification accuracy. Comment on your observations.