SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Intro to Big Data Science: Assignment 4

Due Date: May 10, 2024

✏ **Exercise 1**

Log into “cookdata.cn”, and enroll the course “数据科学导引”. Finish the online exercise there.

✏ **Problem 2** (Support Vector Machine (SVM)) Soft-Margin Linear SVM. Given the following dataset aligning on the x-axis (See the figure below), which consists of 4 positive data points $\{0,1,2,3\}$ and 3 negative data points $\{-3,-2,-1\}$. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft-margin linear SVM can be formalized as the following constrained quadratic optimization problem. In this formulation, C is the regularization parameter, which balances the size of margin (i.e., smaller $\|\mathbf{w}\|_2^2$) vs. the violation of the margin (i.e., smaller $\sum_{i=1}^{m}\xi_i$).

$$\begin{aligned}&\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|_2^2+C\sum_{i=1}^{n}\xi_i\\ &\text{s.t.}\ \ y_i(\mathbf{w}^T\mathbf{x}_i+b)\geqslant 1-\xi_i,\quad \xi_i\geqslant 0,\quad i=1,\ldots,n\end{aligned}$$

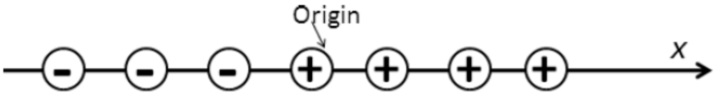


Figure 1: The data set.

1. If $C=0$, which means that we only care the size of the margin, how many support vectors do we have?

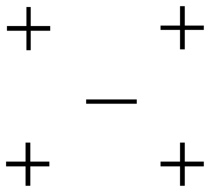
2. if $C \to \infty$, which means that we only care the violation of the margin, how many support vectors do we have?

3. Properties of Kernel:

   a) Using the definition of kernel functions in SVM, prove that the kernel $K(\mathbf{x}_i,\mathbf{x}_j)$ is symmetric, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors for i-th and j-th examples.

   b) Given n training examples $(\mathbf{x}_i,\mathbf{x}_j)$ for $(i,j = 1,\ldots,n)$, the kernel matrix A is an $n \times n$ square matrix, where $A(i,j) = K(\mathbf{x}_i,\mathbf{x}_j)$. Prove that the kernel matrix A is semi-positive definite.

Exercise 3 Consider training an AdaBoost classifier using decision stumps on the five-point data set (4 "+" samples and 1 "-" sample):



1. Which examples will have their weights increased at the end of the first iteration? Circle them.
2. How many iterations will it take to achieve zero training error? Explain by doing some computation using the above algorithm.
3. Can you add one more sample to the training set so that AdaBoost will achieve zero training error in two steps? If not, explain why.

Exercise 4 (Hierarchical Clustering)

Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

1. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion (merge) occurs, as well as the observations corresponding to each leaf in the dendrogram.
2. Repeat 1, this time using single linkage clustering.

3. Suppose that we cut the dendrogram obtained in 1 such that two clusters result. Which observations are in each cluster?
4. Suppose that we cut the dendrogram obtained in 2 such that two clusters result. Which observations are in each cluster?
5. It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in 1, for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

✏ **Exercise 5** In this problem, you need to show that the within-cluster point scatter (or in other words, the sum-of-squared errors (SSE)) is **non-increasing** when the number of clusters increases.

Consider a data set $\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ that contains N observations. Each sample $\mathbf{x}_i$ is a d-dimensional vector of continuous-valued attributes. You are performing K-means clustering.

1. Suppose all the N samples are grouped into a **single** cluster. Let μ be the centroid of the cluster. Express the total sum-of-squared errors SSE_T in terms of $\mathbf{x}_i$, μ and N. And show that SSE_T can be decomposed into d separate terms, one for each attribute, i.e., $SSE_T = \sum_{j=1}^{d} SSE_j$.
2. Now, suppose all the N observations are grouped into two clusters, C_1 and C_2. Let μ_1 and μ_2 be their corresponding cluster centroids while n_1 and n_2 are their respective cluster sizes ($n_1 + n_2 = N$). Express the sum-of-squared errors for each cluster, $SSE^{(j)}$ ($j = 1$ or 2), in terms of $\mathbf{x}_i$, n_j, and μ_j. You need to expand the quadratic term, $(a-b)^2 = a^2 - 2ab + b^2$, and simplify the expression.
3. By rewriting your expression for SSE_T in terms of $\mathbf{x}_i$, n_1, n_2, μ_1, μ_2 and N, show that $SSE_T \geqslant SSE^{(1)} + SSE^{(2)}$.