Assignment 3

ML Class: CS 6375.501

November 3, 2020

1 Assignment Policies for CS 6375

The following are the policies regarding this assignment.

1. This assignment needs be done individually by everyone.
2. You are expected to work on the assignments on your own. If I find the assignments of a group of (two or more) students very similar, the group will get zero points towards this assignment. You may possibly also be reported to the judiciary committee.
3. Please use Python for writing code. You can submit the code as a Jupyter notebook
4. For the theory questions, please use Latex
5. This Assignment is for 30 points.
6. This will be due on November 14th.

2 Questions

1. **PCA and SVMs (8 points):** For this question, we will use PCA for feature selection on a SVM classifier. We will use the UCI Spambase dataset: http://archive.ics.uci.edu/ml/datasets/Spambase/.
   (a) Perform PCA on the training data to reduce the dimensionality of the data set (ignoring the class labels for the moment). What are the top six eigenvalues of the data covariance matrix?
   (b) For each $k \in \{1, 2, 3, 4, 5, 6\}$, project the training data into the best k dimensional subspace (with respect to the Frobenius norm) and use the SVM with slack formulation to learn a classifier for each $c \in \{1, 10, 100, 1000\}$. Report the error of the learned classifier on the validation set for each k and c pair.
   (c) What is the error of the best k/c pair on the test data? How does it compare to the best classifier (with the same possible c choices) without feature selection?
   (d) If you had to pick a value of k before evaluating the performance on the validation set (e.g., if this was not a supervised learning problem), how might you pick it?
2. **Clustering (12 Points):** There are three parts to this question. The first one is to implement k-means and the other two are theory questions.
   (a) Part 1: (6 points) Implement k-means and k-mediods algorithms from scratch. To test your algorithm, just use the features from the digits dataset from scikit-learn:

`sklearn.datasets.load_digits`

The digits dataset consists of 10 classes (digits from 0 to 9). Run your k-means and k-mediods on this dataset with $k = 10$, and visualize the obtained clustering. Compare the performance of k-means and k-mediods clustering and check if the k-means and k-mediods both reduce the clustering loss.

(b) Part 2: (2 points) Recall that in the k-means algorithm, we used the squared L2 distance as the distance measure for clustering. Also recall that there are two main steps in k-means, the first is the assignment step where given means μ_i for $i \in 1, \cdots, k$, we assign the points to the clusters. The second step involves recomputing the means, and recall that in the case of the L2 distance measure, the means step was: $\mu_i = \sum_{j \in C_i} x_j / |C_i|$. Prove that this is this is indeed the mean – recall that the mean is the point with the minimum average distance to all the points in the particular cluster. In other words, given a distance function $d(x, \mu)$, and a set of points $x_1, \cdots, x_n$, the mean can be expressed as:

$$\mu^* = \text{argmin}_\mu \sum_{i=1}^{n} d(x_i, \mu) \tag{1}$$

(c) Part 3: (4 points) Following from the previous part, note that we can compute two means, the left mean and the right mean as follows. The right-mean of a set of points $x_1, \cdots, x_n$ is a point μ^* minimizing equation eq-mean-right, while the left-mean minimizes the following equation:

$$\mu^* = \text{argmin}_\mu \sum_{i=1}^{n} d(\mu, x_i) \tag{2}$$

Next, write down the assignment step and the means step for the distance metrics shown below.

- (left-mean) $D(x, y) = x \log \frac{x}{y}$
- (right-mean) $D(x, y) = x \log \frac{x}{y}$
- k-mediods clustering. Also, what is the complexity of k-mediods clustering and compare it with k-means.
- $D(x, y) = \|x - y\|_1$

3. **VC Dimension (10 points):** In this question you will compute the VC dimension for the following hypothesis spaces. In each case, also compute number of samples to guarantee that an optimal learning algorithm will attain an accuracy of 80% with a probability of atleast 0.95.

(a) (3 points) Consider a binary classification for data points in R^2 with a hypothesis space of a circle of fixed radius r such that any point inside the circle is labeled as positive and outside as negative. Prove it. How does the VC dimension change if the points inside are negative and outside as positive? Also what will be the VC dimension of a hypothesis space which is a union of the above two hypothesis spaces?

(b) (3 points) Instead of a circle, what if we have rectangles in R^2. What is the VC dimension if the points inside are positive and outside as negative? What about if you were allowed to choose whether points inside or outside were classified as positive examples?

(c) (4 points) Consider a binary classification problem for data points in R^2 with hypothesis space consisting of exactly k distinct linear separators. If $k = 3$, what is the smallest and largest VC dimension of this hypothesis space? Prove it by giving specific examples whose VC dimension match your upper and lower bound.