CSC 3520
Machine Learning
Florida Southern College

Assignment 5: Support Vector Machines and Clustering
Due: Friday, December 7, 2018

1. Support Vector Machines (50 points)

Suppose we are given the following positively-labeled data in $\mathbb{R}^2$:

(2,2), (2,-2), (-2,-2), (-2,2)

and the following negatively-labeled data in $\mathbb{R}^2$:

(1,1), (1,-1), (-1,-1), (-1,1)

(a) Plot the data (by hand). Use 'o' for positive samples and 'x' for negative samples. Is there a linear hyperplane that perfectly separates the positive and negative samples in this 2D space?

(b) Suppose we decide to implement a nonlinear SVM with the following kernel:

$$\Phi\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases}\begin{pmatrix}4-x_2+|x_1-x_2|\\4-x_1+|x_1-x_2|\end{pmatrix} & \text{if } \sqrt{x_1^2+x_2^2} > 2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & \text{otherwise}\end{cases}$$

On a separate graph, plot the transformed data after applying this kernel function. Use the same marking convention for positive/negative samples. Is there a linear hyperplane that perfectly separates the positive and negative samples in this 2D space?

(c) Manually draw the SVM decision boundary on the plot of the transformed data. Identify the margin with dashed lines. Give the equation of the decision boundary.

(d) How many support vectors are there? Identify them on the plot by circling them.

(e) Given several test data samples in (un-transformed) 2D space, use the SVM decision boundary to assign a positive (+) or negative (-) label to each sample.

x_1	x_2	label
0	0	
1.5	1.5	
0	4	
2	0	

2. Clustering (50 points)

Given a dataset (data.txt) comprising 1,500 samples in 2D space with corresponding labels [0, 1], write a Python script (called clustering.py) that implements agglomerative hierarchical clustering. Show all plots in your submitted pdf.

(a) Before implementing the clustering algorithm, plot the data points using distinguishable colors for the two labeled classes. Is this data linearly separable?

(b) Use Euclidean distance and the single-linkage metric to generate two clusters via hierarchical clustering. Visualize the results by plotting the data points with two separate colors, one for each cluster.

(c) What is the accuracy of this clustering when compared to the target labels?

(d) Repeat (b) and (c), but with the complete-linkage similarity metric.

(e) Compare the results of single-linkage versus complete-linkage in at least 2-3 sentences. Why is the accuracy different?

(f) Now, suppose we want four clusters instead of two. Repeat (b) for both single-linkage and complete-linkage using $k=4$. Are the results surprising? Explain your reasoning in at least 2-3 sentences.