CS4342 Assignment #6
Total points; 70

Conceptual and Theoretical Questions

1. This problem involves hyperplanes in two dimensions.
(a) Sketch the hyperplane $1 + 3X_1 - X_2 = 0$. Indicate the set of points for which $1 + 3X_1 - X_2 > 0$, as well as the set of points for which $1 + 3X_1 - X_2 < 0$.
(b) On the same plot, sketch the hyperplane $-2 + X_1 + 2X_2 = 0$. Indicate the set of points for which $-2 + X_1 + 2X_2 > 0$, as well as the set of points for which $-2 + X_1 + 2X_2 < 0$.
Points: 5

2. We have seen that in p = 2 dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$.We now investigate a non-linear decision boundary.
(a) Sketch the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) On your sketch, indicate the set of points for which $(1 + X_1)^2 + (2 - X_2)^2 > 4$,
as well as the set of points for which $(1 + X_1)^2 + (2 - X_2)^2 \leq 4$.
(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

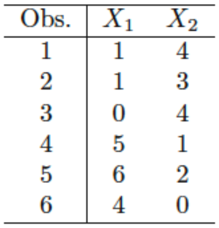
and to the red class otherwise. To what class is the observation (0, 0) classified? (−1, 1)? (2, 2)? (3, 8)?
(d) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_2, X_1^2, and X_2^2.
Points:5

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

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}.$$

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.
(a) 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 occurs, as well as the observations corresponding to each leaf in the dendrogram.
(b) Repeat (a), this time using single linkage clustering.
(c) Suppose that we cut the dendogram obtained in (a) such that two clusters result. Which observations are in each cluster?
(d) Suppose that we cut the dendogram obtained in (b) such that two clusters result. Which observations are in each cluster?
(e) 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 (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.
Points: 5

4. In this problem, you will perform K-means clustering manually, with K = 2, on a small example with n = 6 observations and p = 2 features. The observations are as follows.



Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the observations.
(b) Randomly assign a cluster label to each observation. Report the cluster labels for each observation.
(c) Compute the centroid for each cluster.
(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.
(e) Repeat (c) and (d) until the answers obtained stop changing.
(f) In your plot from (a), color the observations according to the cluster labels obtained.
Points: 5

Applied Questions

1. In this problem, you will use support vector approaches in order to predict whether a given car gets high or low gas mileage based on the Auto data set.
(a) Create a binary variable that takes on a 1 for cars with gas mileage above the median, and a 0 for cars with gas mileage below the median.
(b) Fit a support vector classifier to the data with the linear kernel, in order to predict whether a car gets high or low gas mileage. Report the cross-validation error. Comment on your results.
(c) Now repeat (b), this time using SVMs with radial and polynomial basis kernels, with different values of gamma and degree. Comment on your results.
(d) Make some plots to back up your assertions in (b) and (c).
Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- Check https://scikit-learn.org/0.18/auto_examples/svm/plot_iris.html

Points: 25

3. Consider the USArrests data. We will now perform hierarchical clustering on the states.
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.
(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.
Hints:

- Check https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html

Points: 25