Homework 1

Due January 15th, 2020 by 11:59pm

Instructions: Upload your answers to the questions below to Canvas. Submit the answers to the questions in a PDF file and your code in a (single) separate file. Be sure to comment your code to indicate which lines of your code correspond to which question part. You may use python packages such as scikit-learn to complete this homework.

There are 2 self study assignments and 3 exercises in this homework. The reading assignments are not graded.

Reading Assignments

- Read pages 1-10 of "Kernel Methods in Machine Learning" by T. Hofmann, B. Scholkopf, A. J. Smola.
- Study the scikit-learn tutorial.

The following exercises are based on the ones in *An Introduction to Statistical Learning* and *The Elements of Statistical Learning*.

Exercise 1

In this problem, you will perform k-means clustering manually, with $K = 2$, on a small example with $n = 8$ observations and $p = 2$ features. The observations are as follows.

Obs.	1	2	3	4	5	6	7	8
X_1	1	1	0	0.5	4	6	5	5.5
X_2	3	3.5	4	4.2	1	0.5	0	1.2

Table 1: Observations.

(a) Plot the observations.

(b) Randomly assign a cluster lable to each observation. You can use the *choice()* function in the *Python* package *numpy* to do this. 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.

Exercise 2

In this problem, you will generate simulated data, and then perform PCA and k-means clustering on the data.

(a) Generate a simulated data set with 25 observations in each of three classes (i.e. 75 observations total), and 50 variables (features).

Hint: There are a number of functions in the package numpy that you can use to generate data.

(b) Perform PCA on the 75 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.

(c) Perform k-means clustering of the obervations with $K = 3$. How well do the clusters that you obtained in k-means clustering compare to the true class labels?

Hint: Be careful how you interpret the results: k-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.

(d) Perform k-means clustering with $K = 2$. Describe your results.

(e) Now perform k-means clustering with $K = 4$, and describe your results.

(f) Now perform k-means clustering with $K = 3$ on the first two principal component score vectors, rather than on the raw data. That is, perform k-means clustering on the 75×2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

(g) Using the *scale()* function in the package *scikit-learn*, perform k-means clustering with $K = 3$ on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (c)? Explain.

Exercise 3

Generate data with three features, with 30 data points in each of three classes as follows:

$$\begin{aligned}
\theta_1 &= U(-\pi/8, \pi/8) \\
\phi_1 &= U(0, 2\pi) \\
x_1 &= \sin\theta_1 \cos\phi_1 + W_{11} \\
y_1 &= \sin\theta_1 \sin\phi_1 + W_{12} \\
z_1 &= \cos\theta_1 + W_{13}
\end{aligned}$$

$$\begin{aligned}
\theta_2 &= U(\pi/2 - \pi/4, \pi/2 + \pi/4) \\
\phi_2 &= U(-\pi/4, \pi/4) \\
x_2 &= \sin\theta_2 \cos\phi_2 + W_{21} \\
y_2 &= \sin\theta_2 \sin\phi_2 + W_{22} \\
z_2 &= \cos\theta_2 + W_{23}
\end{aligned}$$

$$\begin{aligned}
\theta_3 &= U(\pi/2 - \pi/4, \pi/2 + \pi/4) \\
\phi_3 &= U(\pi/2 - \pi/4, \pi/2 + \pi/4) \\
x_3 &= \sin\theta_3 \cos\phi_3 + W_{31} \\
y_3 &= \sin\theta_3 \sin\phi_3 + W_{32} \\
z_3 &= \cos\theta_3 + W_{33}
\end{aligned}$$

Here $U(a, b)$ indicates a uniform variable on the range $[a, b]$ and W_{jk} are independent normal variables with mean 0 and standard deviation 0.5. Hence the data lie near the surface of a sphere in three clusters centered at $(1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$.

Write a program to run k-means on these data. How does the clustering evolve as the k-means algorithm run? What do you observe? Vary the number of clusters. What do you observe?