STAT0032: Exercise Sheet #8

The exercises in this sheet focus on assorted questions on unsupervised learning. As in the previous sheet, a couple of the questions are from James et al., "An Introduction to Statistical Learning" (ISLR).

1. Provide a concrete example of a problem where you can get a bad estimate of a tail area probability $P(X > c)$ for some c, where X follows a non-Gaussian distribution but where you made the wrong assumption that X is Gaussian.

2. (Computer Implementation) (Adapted from Wasserman, Chapter 20) Consider the forensic glass data available in the MASS package of R (under the name of FGL; also available as GLASS.DAT in Moodle).

   (a) Estimate the density of the first variable (refractive index) using a histogram and a kernel density estimator. DO NOT LET R CHOOSE THE SMOOTHING FOR YOU. Instead, without seeing R's choice, play with different binwidths/bandwidths and visualize it until it looks "reasonable" to you. Using argument `breaks` for the histogram, and `bw` for the kernel density estimator.

   (b) Allow R (or implement cross-validation yourself) to select the amount of smoothing in both methods. How close do they get to what you chose in (a)?

   (c) Construct confidence intervals for your estimators by reusing the corresponding code from Week8Code.R, plotting them. Comment how well they might accommodate for the choice you made in part (a).

   (d) Notice that this dataset has different classes of glass. Do density estimation separately in each one of them. In which ways they look similar/dissimilar? How do they look like compared to the aggregated data in part (c)? How do they compare to a Gaussian model applied to different classes?

3. A pair of variables (X_1, X_2) follows a bivariate Gaussian with zero mean, unit variance and correlation coefficient of 0.4. A different pair of variables (Y_1, Y_2) follows a bivariate Gaussian distribution with zero mean, unit variance and correlation coefficient of zero.

   (a) Is $P(-1 \leq X_1 \leq 2)$ equal, less than or more than $P(-1 \leq Y_1 \leq 2)$? Explain your reasoning.

   (b) Express $P(0 \leq X_1 \leq 3 \text{ and } 0 \leq X_2 \leq 3)$ as an integral over the joint pdf of X_1 and X_2.

   (c) Is $P(0 \leq X_1 \leq 3 \text{ and } 0 \leq X_2 \leq 3)$ equal, less than or more than $P(0 \leq Y_1 \leq 3 \text{ and } 0 \leq Y_2 \leq 3)$? Explain your reasoning.

4. (Exercise 3 of Chapter 10, ISLR.) 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. You can use the `sample()` command in R 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.

5. (Exercise 5 of Chapter 10, ISLR.) In words, describe the results that you would expect if you performed K-means clustering of the eight shoppers in Figure 10.14, on the basis of their sock and computer purchases, with $K = 2$. Give three answers, one for each of the variable scalings displayed. Explain.

Figure 10.14 from ISLR is given below:

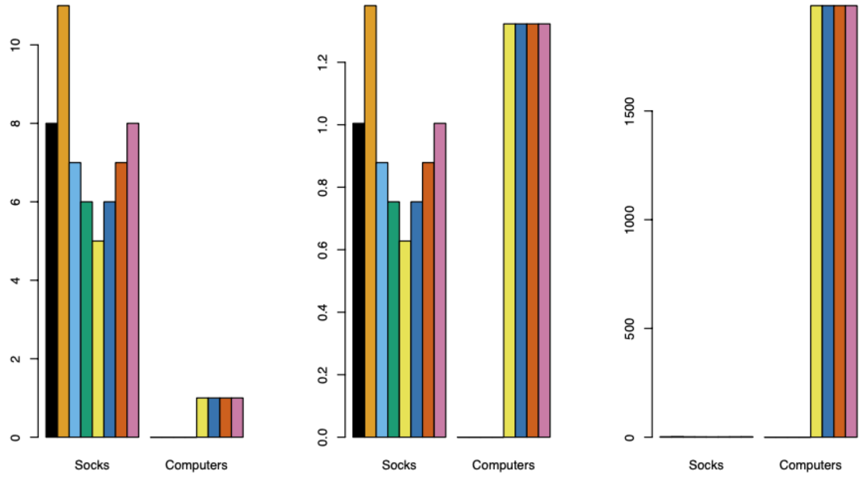

FIGURE 10.14. *An eclectic online retailer sells two items: socks and computers.* Left: *The number of pairs of socks, and computers, purchased by eight online shoppers is displayed. Each shopper is shown in a different color. If inter-observation dissimilarities are computed using Euclidean distance on the raw variables, then the number of socks purchased by an individual will drive the dissimilarities obtained, and the number of computers purchased will have little effect. This might be undesirable, since (1) computers are more expensive than socks and so the online retailer may be more interested in encouraging shoppers to buy computers than socks, and (2) a large difference in the number of socks purchased by two shoppers may be less informative about the shoppers' overall shopping preferences than a small difference in the number of computers purchased.* Center: *The same data is shown, after scaling each variable by its standard deviation. Now the number of computers purchased will have a much greater effect on the inter-observation dissimilarities obtained.* Right: *The same data are displayed, but now the y-axis represents the number of dollars spent by each online shopper on socks and on computers. Since computers are much more expensive than socks, now computer purchase history will drive the inter-observation dissimilarities obtained.*

6. (Exercise 6 of Chapter 10, ISLR.) A researcher collects expression measurements for 1000 genes in 100 tissue samples. The data be written as a 1000×100 matrix, which we call X, in which each row represents a gene and each column a tissue sample. Each tissue sample was processed on a different day, and the columns of X are ordered so that the samples that were processed earliest are on the left, and the samples that were processed later are on the right. The tissue samples belong to two groups: control (C) and treatment (T). The C and T samples were processed in a random order across the days. The researcher wishes to determine whether each gene's expression measurements differ between the treatment and control groups.

   As a pre-analysis (before comparing T versus C), the researcher performs a principal component analysis of the data, and finds that the first principal component (a vector of length 100) has a strong linear trend from left to right, and explains 10% of the variation. The researcher now remembers that each patient sample was run on one of two machines, A and B, and machine A was used more often in the earlier times while B was used more often later. The researcher has a record of which sample was run on which machine.

   (a) Explain what it means that the first principal component "explains 10% of the variation".

   (b) The researcher decides to replace the (i, j)th element of X with

   $$x_{ij} - z_{i1}\phi_{j1}$$

   where z_{i1} is the ith score, and ϕ_{j1} is the jth loading, for the first principal component. He will then perform a two-sample t-test on each gene in this new data set in order to determine whether its expression differs between the two conditions. Critique this idea, and suggest a better approach.

   (c) Design and run a small simulation experiment to demonstrate the superiority of your idea.

7. Explore (using algebra) the relationship between the maximum likelihood of the Gaussian mixture model (GMM) and the K-means optimisation.

8. (Computer Implementation) Exercise 8 of Chapter 12, ISLR.

9. (Computer Implementation) Exercise 10 of Chapter 12, ISLR.

10. (Computer Implementation) Use k-means with the dataset UCL.DAT provided in the Moodle page. What would you do to make it work better?