STATG006: 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 CHAPTER6.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?

   (e) Comment on the implications of (d) if we were to do classification using Bayes' rule as discussed in Question 3 of Exercise Sheet #5.

3. (Computer Implementation) This concerns reproducing the some of the experiments of Chapter 6 and understanding its steps.

   (a) Explain how function GAUSSIFY from CHAPTER6.R works.

   (b) After applying GAUSSIFY to the gene expression data of Sachs et al., perform some visual check of whether the Gaussian copula is appropriate for the different pairs of variables in this data. Explain your reasoning and conclusion.

4. 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 coefficien 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.

5. This exercise concerns the relationship between partial correlation and regression.

   (a) State the relationship between the coefficients $(\beta_0^\star, \beta_1^\star, \ldots, \beta_p^\star)$ obtained by

   $$(\beta_0^\star, \beta_1^\star, \ldots, \beta_p^\star) = \text{argmin}_{(\beta_0,\ldots,\beta_p)} E(Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p)^2$$

   and the correlation matrix of the joint distribution of $(X_1, \ldots, X_p, Y)$.

   (b) How does this help you to understand the difficulties that model selection for regression can face when some covariates are conditionally strongly correlated given the other covariates?

   (c) If you know that all variables in $(X_1, \ldots, X_p)$ are mutually independent, how would use you this information to write a method that is guaranteed to find the "true" covariates that should be used in a regression model for Y given $(X_1, \ldots, X_p)$?

6. Exercise 3 of Chapter 10, ISLR.

7. Exercise 5 of Chapter 10, ISLR.

8. Exercise 6 of Chapter 10, ISLR.

9. Prove the relationship between the "simplified" maximum likelihood of the mixture of Gaussian model and K-means.

10. (Computer Implementation) Exercise 8 of Chapter 10, ISLR.

11. (Computer Implementation) Exercise 10 of Chapter 10, ISLR.

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

13. (For Fun) See how well you can do at `http://guessthecorrelation.com/`.