Modern Statistics and Big Data Analysis, Exercises 5

Deadline: End of Friday 17 November.

1. (3.5 points) On Virtuale under "Data sets" you can find a data set named `glass.dat`. The data set can be read into R as follows:

   ```
   glass <- read.table("glass.dat",header=TRUE)
   ```

   Some methods require the matrix version as input:

   ```
   mglass <- as.matrix(glass)
   ```

   The dataset gives measurements taken on splinters of glass. There are 9 variables:

   RI refractive index

   Na Sodium (natrium; unit measurement: weight percent in corresponding oxide, same as the following variables)

   Mg Magnesium

   Al Aluminum

   Si Silicon

   K Potassium (kalium)

   Ca Calcium

   Ba Barium

   Fe Iron (ferrum)

   All these variables can be used.

   The task is to cluster the glass splinters into different types of glass. Specifically, these are from crime scenes, and it is of interest how many different types of glass there are, and which splinters belong together. The term "type of glass" can refer for example to different types of windows, headlamps, bowls etc.

   Find a good clustering, i.e., one of which you think that it captures in the best possible manner a really meaningful grouping. Try out at least two clusterings, of which at least one is based on a Gaussian mixture, and at least one is from a different approach.

   Compare the clusterings and comment on how meaningful and useful you think they are. Select one clustering that you prefer. Discuss in particular whether the Gaussian mixture is a good method for these data in your view, and what might be potential problems with it.

   Produce at least one visualisation each for at least two clusterings.

   Interpret the clusters of the chosen clustering (you can use all given variables and your visualisations).

Comment on other aspects of the data set that you find out as far as you think they could be relevant.

There is no single correct or best solution that I have in mind and want to see here. I'm very open to your suggestions. If you have doubts about the one you are proposing, please write these down. It's more important to have something that appropriately reflects the data than something that looks "strong" if in fact it isn't.

Note: This is a slightly modified past exam question; in the exam I just asked for at least two clusterings and didn't specify that there should be a Gaussian mixture clustering.

2. (3 points) Consider the M-step of the EM-algorithm for one-dimensional Gaussian mixtures (see p. 234/235 of the slides for the p-dimensional version). Consider maximising

$$E_\eta = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}(\log \pi_k + \log \varphi_{a_k,\sigma_k^2}(x_i)) \tag{1}$$

by choice of $\pi_1, \ldots, \pi_K, a_1, \ldots, a_K, \sigma_1^2, \ldots, \sigma_K^2$, all one-dimensional, where φ_{a,σ^2} is the density of a Gaussian distribution with mean a and variance σ^2. I omitted the $(t-1)$ of $p_{ik}^{(t-1)}$ from the slides for ease of notation.

(a) Show that for $k = 1, \ldots, K$, regardless of $a_1, \ldots, a_K, \sigma_1^2, \ldots, \sigma_K^2$, (1) is maximised by choosing π_k as

$$\pi_k^* = \frac{1}{n} \sum_{i=1}^{n} p_{ik}.$$

Hint: This is a constrained maximisation problem and can be solved by the method of Lagrange multipliers
(see `https://en.wikipedia.org/wiki/Lagrange_multiplier`).
The constraint is

$$\sum_{k=1}^{K} \pi_k = 1 \Leftrightarrow \sum_{k=1}^{K} \pi_k - 1 = 0.$$

This means that you need to take partial derivatives w.r.t. π_k, $k = 1, \ldots, K$, and λ of

$$\sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}(\log \pi_k + \log \varphi_{a_k,\sigma_k^2}(x_i)) - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right),$$

set them all to zero, solve for $\lambda, \pi_1, \ldots, \pi_K$, and show that this leads to the choices of π_k given above. You can use $\sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik} = n$. (Give an argument why this holds.)

(b) Show that for $k = 1, \ldots, K$, regardless of $\pi_1, \ldots, \pi_K$, (1) is maximised by choosing a_k as

$$a_k^* = \frac{1}{\sum_{i=1}^{n} p_{ik}} \sum_{i=1}^{n} p_{ik} x_i,$$

and σ_k^2 as

$$\sigma_k^{2*} = \frac{1}{\sum_{i=1}^{n} p_{ik}} \sum_{i=1}^{n} p_{ik}(x_i - a_k)^2.$$

Hint: Look up and use a proof that derives the ML-estimators (mean and ML-variance, which has a factor $\frac{1}{n}$ rather than $\frac{1}{n-1}$ used for the sample variance) for a Gaussian distribution, and introduce the weights p_{ik} there.

3. (3.5 points) On the Virtuale page I have uploaded the paper
   Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. "A density-based algorithm for discovering clusters in large spatial databases with noise". In Simoudis, E., Han, J. and Fayyad, U.M. (eds.) Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) . Palo Alto, CA: AAAI Press, 226-231.

   Read the paper and answer the following questions (don't be too worried about not understanding everything in the paper):

   (a) Summarize in your own words what the DBSCAN method does.

   (b) What are the advantages, according to the authors, of their DBSCAN method compared with other clustering methods, particularly those that you already know? Do you think that the authors' arguments are convincing?

   (c) Find out how to run DBSCAN in R and apply it to the Glass data from question 1 (you may also try it out on other datasets). You will need to make some decisions (and/or experiments) about tuning parameters. Comment on the results.