University of Michigan, Dept of Statistics

Stat 503, Instructor: Long Nguyen

Homework Assignment 5
(due Tuesday, April 17, 2018)
in class or Canvas dropbox by 1pm

1. **Clustering analysis** The crabs data set (`crabs.txt` on Canvas) describes the morphological measurements of crabs of the species Leptograpsus variegatus collected in Australia. The variables are: Species (1=blue crabs, 2 = orange crabs); Sex (1=male, 2=female); FL (frontal lobe size measured in mm); RW (rear width in mm); CL (carapace length in cm); CW (carapace width in cm); BD (body depth in cm).

   (a) Using the last five variables from the crabs dataset (ignore Species and Sex), perform clustering on the data using (i) hierarchical clustering method (compare at least three different distance measures, or more if you like), (ii) K-means, and (iii) mixture modeling. Try these methods on different values of k, and then use the BIC criterion (for the mixture model) to pick the best value of k.

   (b) Compare results between K-means and model-based clustering using the k you have chosen, and also compare the clusters you find with the Species and Sex variables, and report how well the clusters agree with those pre-defined categories for each method. Include silhouette plots and comment on the quality of clustering.

   Hints: Useful R functions for this homework include `kmeans()`, `pam()`, `agnes()`, `silhouette()`, `Mclust()`. Consult also the clustering chapter in Michailidis and the lab materials.

2. **Mixture models** This exercise helps you get acquainted with the EM algorithm, which is the most popular model fitting algorithm for latent variable models nowadays.

   (a) By mimicking the EM algorithm for Gaussian mixtures, derive the EM algorithm for parameter estimation of a mixture of two Gamma distributions, given an n-iid sample $X_1, \ldots, X_n$. The Gamma mixture density takes the following form

   $$p(x|\pi, \beta_1, \beta_2) = \pi \mathrm{Gam}(x|\beta_1, 1) + (1-\pi)\mathrm{Gam}(x|\beta_2, 1).$$

   (b) This exercise will help you appreciate how broadly applicable the EM algorithm can be (and how far removed are we from the (initial) K-means algorithm): Derive the EM algorithm for parameter estimation of a mixture of two distributions of different types, given an n-iid sample $X_1, \ldots, X_n$:

   $$p(x|\pi, \mu, \beta) = \pi \mathrm{Norm}(x|\mu, 1) + (1-\pi)\mathrm{Gam}(x|\beta, 1).$$

   In the above $\mathrm{Norm}(x|\mu, 1)$ denotes the normal density with mean μ and unit variance, while $\mathrm{Gam}(x|\beta, 1)$ denotes the gamma density with rate β and unit shape (which becomes an exponential density): for $x > 0$

   $$\mathrm{Gam}(x|\beta, 1) = \beta e^{-\beta x}.$$

   (c) **Bayesian estimation** Taking a Bayesian approach on the mixture model of part (b), propose prior distributions for parameters π, μ and β. Rewrite this model as a latent variable model (which includes the latent variables $Z_1, \ldots, Z_n \in \{0, 1\}$). Describe a Gibbs sampling algorithm by deriving the conditional distribution of Z_i given everything else, and then giving the conditional distribution of each of the parameters π, μ and β, given $(Z_i, X_i)_{i=1}^n$ and all remaining parameters.

Instructions The solution to problem 2 can be either typed up or written by hand. The solution to data analysis questions (Problem 1) may be written as a data analysis report. The report needs to be clear, concise, and to the point. There should be no graphs or tables that are not commented on in the text. Please include your R code (at least the main parts) in a *separate* appendix at the end of the report.