Assignment 4 - CSC/DSC 265/465 - Spring 2018 - Due May 1

Q1: We wish to fit the model

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n, \tag{1}$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and x_i is a predictor variable. The function $g(x)$ has the following properties:

(i) There are two knots $\xi_1 < \xi_2$.
(ii) $g(x)$ is continuous at the knots.
(iii) $g(x)$ possesses a continuous first derivative at the knots.
(iv) $g(x)$ is a first order polynomial $g(x) = a_0 + b_0 x$ for $x < \xi_1$.
(v) $g(x)$ is a second order polynomial $g(x) = a_1 + b_1 x + c_1 x^2$ for $x \in (\xi_1, \xi_2)$.
(vi) $g(x)$ is a first order polynomial $g(x) = a_2 + b_2 x$ for $x > \xi_2$.

(a) How many linear constraints are imposed on the parameters $(a_0, b_0, a_1, b_1, c_1, a_2, b_2)$ by properties (i)-(vi)? Write these explicitly.
(b) Assume the knots ξ_1, ξ_2 are known, but the parameters $(a_0, b_0, a_1, b_1, c_1, a_2, b_2)$ are to be estimated. How many degrees of freedom does this estimation problem possess (that is, how many free parameters are required to completely define $g(x)$)?

Q2: For this problem use data set `fgl` from the `MASS` package. This is a forensic application. The observations consist of fragments of broken glass. The `type` column gives the type of glass. The `RI` column gives refractive index. See description from `help(fgl)`. The remaining 8 columns are percentages by weight of various oxides. The row totals of these 8 columns are approximately 100%.

In this problem the ability of hierarchical clustering to distinguish between types of glass based on forensic samples of chemical composition and refractive index will be examined.

(a) For this exercise we will only consider 3 glass types: window float glass (`WinF`), window non-float glass (`WinNF`) and vehicle headlamps (`Head`). Create a subset of the data from only these glass types. Then standardize each column to zero mean and unit variance.
(b) Using the function `hclust` plot dendograms for hierarchical clusterings using agglomeration methods `single`, `complete` and `average`. Generally, do the observations appear to cluster by class `gr` in any of the dendograms? Substitute single character labels when plotting the histograms, since `WinF` and `WinNF` will be difficult to distinguish visually.
(c) There are various ways to quantify the ability of a hierarchal clustering to accurately distinguish classes. Suppose we create a single clustering of size $k =$ `c.size`, using `cutree(hfit,k=c.size)`. Suppose one sample from each of the 3 types of glass is chosen at random. Let α_k be the probability that the 3 observations are in different clusters. Suppose p_{js} is the proportion of samples of glass type $j \in \{1, 2, 3\}$ in cluster $s \in \{1, \ldots, k\}$. Give an expression for α_k in terms of the proportions p_{js}.
(d) The proportions p_{js} can be easily estimated by cross-tabulating glass type and cluster membership. Write an `R` program that estimates and plots α_k for each of the hierarchical clusterings created in Part (b). Superimpose the three plots on a single graph, and use the range $k = 1, \ldots, 10$. In general, how do the agglomeration methods compare in terms of accuracy?
(e) One way to assess whether or not α_k is significantly large is to use a permutation procedure. Suppose the original types are contained in the vector `gr`. Then, create a new class vector `gr.perm` by randomly permuting the original class vector `gr` (you can use function `sample()`). Create a new sequence α'_k, $k = 1, \ldots, 10$ with the same procedure used in Part (d), except that `gr` is replaced by `gr.perm`. Do the permutation 25 times, superimposing all α_k and α'_k sequences on the same plot. Make sure the sequence types are easily distiguishable (say, use green for α_k and gray for each α'_k). Do this for each of the hierarchical clusterings created in Part (b). Use separate plots for each, but use the `ylim=c(0,1)` option when plotting so that the scales will be comparable. Which clusterings are compatible with the actual glass types?

Q3: This problem will make use of the `biopsy` data set from the `MASS` library (this data was used in Question 3 of Assignment 3). See `help(biopsy)` for details.

(a) Prepare the data by first removing the `ID` column, then removing records with missing values using the `na.omit()` function. The new data set should have $n = 683$ records. Column 10 is now the `class` variable, containing the tumor class (`benign` or `malignant`). Columns 1-9 now contain quantitative tumor features with which to discrimate between tumor types. In this analysis, instead of normalizing the features to zero mean and unit variance, subject each feature to a log transformation (use the natural logarithm).

(b) Calculate K-means cluster solutions based on the 9-log transformed features. Use $K = 1, \ldots, 25$. For each K calculate $R^2 = 1 - SS_{within}/SS_{total}$, then plot R^2 against K. Between which two values of K does the greatest increase in R^2 occur? How does this relate to the true number of clusters?

(c) Calculate the principal components of the 9 log transformed features using the `prcomp()` function. Use centering but not scaling, that is, use options `center=T` and `scale.=F`. Create a pairwise plot (using function `pairs()`) for the first 4 principal components, using separate coloring for each class (the classes need not be labeled). Then create a scree plot. This can be done using the command `plot(pr.fit)`, where `pr.fit` is the principal components object created by the `prcomp()` function. How do the various plots suggest that most of the discriminating information regarding tumor type is contained by the first principal component?

(d) Finally, calculate a LASSO fit using response $Y_i = 1$ if `class` is `malignant` and $Y_i = 0$ otherwise (use the `binomial` model). Use cross-validition with function `cv.glmnet()` and options `family='binomial'` and `alpha=1`. Do this using the original log-transformed features as predictors, then using the 9 principal components calculated in Part (c) as predictors. Examine the coefficients for the `fit$lambda.1se` solution for each set of predictors. Do these conform to what you see in Part (c)? (Note that since cross-validation is random, repeated fits will yield different coeffients. However, the overall conclusion should be the same).

Q4: We wish to fit a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and x_i is a predictor variable in the range $[1, 10]$. We consider the following six models

M1 $g(x) = \beta_1 x$, where β_1 is to be estimated.

M2 $g(x) = \beta_0 + \beta_1 x$, where β_0, β_1 are to be estimated.

M3 $g(x) = \beta_1 \sqrt{x}$, where β_1 is to be estimated.

M4 $g(x) = \beta_0 + \beta_1 \sqrt{x}$, where β_0, β_1 are to be estimated.

M5 $g(x)$ is a continuous piecewise linear spline with 1 knot at $\xi = 4$.

M6 $g(x)$ is a cubic spline with 2 knots at $\xi = 3, 6$.

The relevant SSE values are given in the following table. The sample size is $n = 91$. Which model is preferred based on the AIC score and on the BIC score (use form $n \log(SSE/n) + C$ for each)? Does this model minimize SSE among those considered?

Q5: **[For Graduate Students]** Consider the matrix representation of the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X}$ is a $n \times q$ matrix, $\boldsymbol{\beta}$ is a $q \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms. The least squares solution is expressed using the coefficient vector $\boldsymbol{\beta}$ which minimizes the error sum of squares

$$SSE[\boldsymbol{\beta}] = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Model	SSE
M1	74.007
M2	3.441
M3	9.258
M4	2.811
M5	2.935
M6	2.744

(a) By setting each partial derivative $\partial SSE[\boldsymbol{\beta}]/\partial\beta_j$ to zero, $j = 1, \ldots, q$, verify that the least squares solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

(b) Next, recall that the ridge regression coefficients are the obtained by minimizing

$$\Lambda = SSE[\beta] + \lambda \sum_{j=1}^{q} \beta_j^2$$

for a fixed constant $\lambda \geq 0$. By setting each partial derivative $\partial SSE[\boldsymbol{\beta}]/\partial\beta_j$ to zero, $j = 1, \ldots, q$, show that the ridge regression solution is

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I_q)^{-1}\mathbf{X}^T\mathbf{y},$$

where I_q is the $q \times q$ identity matrix.