Assignment 4 - CSC/DSC 265/465 - Spring 2017 - Due May 9

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. We set

$$g(x) = \left\{ \begin{array}{lll} a_1x^3 + b_1x^2 + c_1x + d_1 & ; & x < \xi \\ a_2x^3 + b_2x^2 + c_2x + d_2 & ; & x \geq \xi \end{array} \right. , \tag{2}$$

where ξ is fixed, and the polynomial coefficients a_j, b_j, c_j, d_j, $j = 1, 2$ are to be estimated. However, the coefficients must be constrained so that $g(x)$ is continuous, and possesses continuous first and second derivatives, at ξ.

(a) Give precisely the linear constraints on the coefficients required for the given continuity conditions. How can these constraints be used to determine the model degrees of freedom?

(b) Show that the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 (x_i - \xi)_+^3 + \epsilon_i, \;\; i = 1, \ldots, n, \tag{3}$$

is equivalent to (1)-(2). HINT: First show this is true for $\xi = 0$.

Q2: For this problem use data set `Cars93` from the `MASS` package. There is data on 93 makes of automobile, including `Manufacturer`, `Model`, `Type`, `Origin`, and miscellaneous technical data (`Wheelbase`, `RPM`, etc). The purpose of this analysis is to show how hierarchical clustering may be useful in exploratory analysis.

(a) Select the features for this analysis using the following column indices:

```
xf = Cars93[,c(5,7,8,12,13,14,15,17,19,20,21,22,25)]
```

Standardize each column to zero mean and unit variance. Then, create a class vector `gr` from variable `Man.trans.avail`. This identifies whether or not manual transmission is available.

(b) Using the function `hclust` plot dendograms for hierarchical clusterings using agglomeration methods `single`, `complete` and `average`. Label the observations by `gr`. Do the observations appear to cluster by class `gr` in any of the dendograms?

(c) There are a number of quantities which may be used to determine whether or not a clustering conforms well to a known class variable. Suppose we create a single clustering of size `c.size` (using `cutree(hfit,k=c.size)`. Suppose `gr` contains exactly two classes. Let α_k be the probability that two observations, one randomly chosen from each class, are in the same cluster, given k clusters. This can be easily estimated by cross-tabulating class and cluster frequencies. Smaller values of α_k suggest that the cluster conforms to the class.

   (i) Show that $\alpha_{k+1} \leq \alpha_k$ for $k \geq 1$.

   (ii) Show that α_k approaches 0 as k approaches sample size n.

(d) For each of the hierarchical clusterings, plot α_k for $k = 1, \ldots, 93$ (plot all three on the same graph). Does one agglomeration method have smaller α_k for most cluster sizes k?

(e) One way to assess whether or not α_k is significantly small is to use a permutation procedure. Suppose the class vector is randomly permuted. This should eliminate any association between class and cluster. To see this, using the agglomeragtion method selected in part (d), plot again α_k for $k = 1, \ldots, 25$. 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, $k = 1, \ldots, 25$, using the same procedure, except that `gr` is replaced by `gr.perm`. Do the permutation 10 times, superimposing all α_k and α'_k sequences on the same plot. Make sure the sequence types are easily distiguishable (say, use black for α_k and gray for each α'_k). Does the plot suggest that there is a statistically significant association between cluster and class?

(f) Finally, calculated a LASSO fit using `gr` as response and the feature matrix `xf` (use the `binomial` model). Examine the coefficients for the `fit$lambda.min` solution. Do the selected variables seem related to class? In other words, what type of cars tend to have manual transmission?

Q3: A classifier requires variation within a set of features, which can be quantified and analyzed in various ways. The purpose of a classifier can be thought of as to determine what portion of the variation can be explained by class. Sometimes, it is possible to identify, then remove, feature variation which we know will not be explained by class.

For this problem use data set `crabs` from the `MASS` package. This contains data on 200 crabs. There are 5 morphological measurements (columns 4-8). The variable `sp` identifies the crab by species (`B` for blue, `O` for orange). The variable `sex` identifies the sex (`M` or `F`).

(a) By combining `sp` and `sex` we can identify 4 classes of crab in total. Create a class variable `gr` which does this.

(b) Create a pairwise plot using all 5 morphological features (leave them all in their original units of millimeters *mm*). Color each class separately (they need not be labeled). How would you characterize feature variation attributable to class? What other form of variation is there which is not explainable by class?

(c) Calculate the principle components for the 5 morphological features. Using centering but not scaling. Create a pairwise plot using all 5 principle components, using separate coloring for each class (again, classes need not be labeled). What form of variation does the first principle component capture. What subset of principle components appears to best capture variation due to class?

(d) Create a function that inputs a feature matrix `X`, number of classes `K` and class vector `gr`, and which performs the following steps:

   (i) Calculates a K-means cluster solution based on the input `X` and `K`.

   (ii) Draws pairwise plots using all features, and superimposes the centers output with clustering solution (see lecture code). Classes need not be visually distinguished, but this won't be discouraged, as long as the centers are clearly distinguishable.

   (iii) Calculates $R^2 = 1 - SS_{within}/SS_{total}$.

   (iv) Calculates the classification error rate. Because K-means clustering is an unsupervised learning algorithm, we need to define 'error' carefully. Assign to each true class the highest frequency cluster among observations of that class. Take that cluster to be a correct prediction, then calculate classification error accordingly.

(e) Apply the function of part (d) to the original feature matrix X, the (centered but unscaled) principal components P, and to the feature matrix $P^{(-1)}$ defined by principle components 2,3,4 and 5 .

(f) It can be shown that the principle components are an orthonormal transformation R of the original data, which is *isometric*, or distance-preserving:

$$\|x - y\| = \|xR - yR\|.$$

What role does this fact play in the results of part (e)?

(g) Suppose we wish to construct a classifier for $(species, sex)$, and we can choose between any of the three feature matrices used in part (e). Which have the highest R^2 and which have the lowest classification error? What form of variation is the K-means solution for feature matrix X attempting to explain? What form of variation is the K-means solution for feature matrix $P^{(-1)}$ attempting to explain?

(h) **(This part need not be handed in)** In this analysis, no scaling was used either for the original feature matrix X or for the principle components. Here, all variance is in the same units (standard deviation in millimeters). When features are in different units, scaling should generally be used. However, when features are in the same units, it may be that the unscaled feature variances are already proportional to the information content, and so shouldn't be changed. Repeat part (e) using various scaling methods. For example, X may be scaled, the principle components can use the scaling option, and the principal components themselves can be scaled (or rescaled, as apporpriate). Can you improve the classification error in this way?