Assignment 3 - CSC/DSC 265/465 - Spring 2017 - Due April 18

Q1: The purpose of the AIC score is to minimize prediction error, rather than to identify the correct model. The two goals are obviously related, but they are not identical. This can be seen using simple linear regression.

Suppose we are given model

$$y = \beta_0 + \beta_1 x + \epsilon, \ \ \epsilon \sim N(0, \sigma^2).$$

Suppose, given paired observations (y_i, x_i), $i = 1, \ldots, n$, we can calculate least-squares coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$. We want to predict, for some fixed predictor value x a new observation $Y_x \sim N(\beta_0 + \beta_1 x, \sigma^2)$ from the model which is independent of the observations used in the estimates. We consider two predictors:

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i \approx Y_x,$$

or

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x \approx Y_x.$$

In one sense, $\hat{y}_x$ is the correct choice, unless $\beta_1 = 0$ (which we don't rule out). One approach is to test against null hypothesis $H_0 : \beta_1 = 0$, then choose $\hat{y}_x$ if we reject H_0. The other approach is to try to minimize prediction error directly.

Here, the square-error risk is:

$$MSE(y') = E\left[(Y_x - y')^2\right],$$

where y' is whatever predictor (that is, $\bar{y}$ or $\hat{y}_x$) we are considering.

(a) Express $MSE(y')$ in terms of σ^2, and the bias and variance of y'. Assume Y_x and y' are independent. Note that in this case $bias(y') = E[y'] - E[Y_x]$.

(b) Derive $MSE(y')$ for $y' = \bar{y}$ and $y' = \hat{y}_x$.

(c) Give conditions on $\beta_0, \beta_1, \sigma^2, SS_X = \sum_{i=1}^{n}(x_i - \bar{x})^2$ under which $MSE(\bar{y}) < MSE(\hat{y}_x)$. Is it possible that $\bar{y}$ could have smaller prediction error even if $\beta_1 \neq 0$?

Q2: Relationships between the size of two physiological components Y, X of a species of animal often obey a power relationship

$$Y = KX^r,$$

where K and r are two fixed constants. Of course, the value r is not necessarily $r = 1$, but will depend on the size measure, and any number of scaling principles. Suppose we have paired observations (X_i, Y_i), $i = 1, \ldots, n$. Then K and r may be estimated using simple linear regression, after taking the double-log transformation:

$$\log Y_i = \log K + r \log X_i, \tag{1}$$

that is, we have intercept and slope $\beta_0 = \log K$, $\beta_1 = r$.

For this problem use data set `Animals` from the `MASS` package, which contains X = average body (kg) and Y = brain (g) weights for 28 species of land animals. We may expect Y to be positively associated with an animal's cognitive abilities. However, we also expect X and Y to be positively associated for reasons having nothing to do with cognitive abilities. Thus, the *encephalization quotient* (EQ) measures the relative brain size after controlling for body size (`en.wikipedia.org/wiki/Encephalization_quotient`).

(a) Construct a scatter-plot of $\log Brain$ against $\log Body$. Use $\log Body$ as the horizontal axis. Instead of plotting symbols, plot the actual name of the species (here, the `text` command may be used). Does there seem to be a linear trend on the double-log scale? Identify the three most obvious outliers. How do they differ from the remaining species?

(b) Fit the model (1), with and without the outliers, and superimpose each fitted line on the scatter-plot. Give the estimates of K and r for each fit.

(c) The *encephalization quotient* (EQ) can be formally defined as the ratio of the actual brain mass to the predicted brain mass based on the species size. If model (1) is used to predict brain mass, show that for species i

$$EQ \approx \exp(e_i)$$

where e_i is the residual from the regression fit for that species.

(d) After removing the outliers, rank the species by their EQ. How would the EQ of the outlier species rank?

Q3: For this problem use data set `biopsy` from the `MASS` package. Biopsies of 699 breat tumours were examined and classified as `benign` or `malignant`. Nine attributes (features) were scored on a scale of 1 to 10, higher scores signifying evidence that the tumour is malignant (use `help(biopsy)` for details).

We consider how to consolidate the features into a single quantitative predictor for malignancy. The simplest method is to sum the features to yield a single score S on a scale of 9 to 90, which the use of a 10 point scale suggests. This would be equivalent to a linear model in which all coefficients β_i are equal, other than the intercept. On the other hand, it is always possible that a combination of model selection and unequal coefficients will yield a strictly better predictor.

(a) Remove observations with missing values using the `na.omit` function. There should be $n = 683$ observations remaining. The 9 biopsy features can then be used to create 683×9 feature matrix X. Label the ith feature (column of X) F_i.

(b) Calculate the principal components of the 9 features. Set option `center = TRUE` and `scale. = FALSE`. That is, we don't rescale the original 10 point scale. The principal components are given by

$$P = \bar{X}A$$

where $\bar{X}$ is the $n \times 9$ feature matrix after centering by column (ie. by subtracting the column means), P is the $n \times 9$ matrix of principal components, and A is the 9×9 matrix of loadings. So, for example, the first principal component is a linear combination of centered features, with coefficients given by the first column of A. Label the ith centered feature (column of $\bar{X}$) $\bar{F}_i$. Then $\bar{F}_i = F_i - m_i$, where m_i is an $n \times 1$ constant column vector, with all elements equal to the mean of F_i.

(c) Create pairwise plots using all principal components. Use separate colors for the `benign` and `malignant` classes. Create a *scree* plot. Comment on what you see.

(d) Examine the loadings matrix A.

   (i) How does the 1st principal component resemble sum S?

   (ii) What feature has the largest (in magnitude) loading for the 2nd principal component?

   (iii) What feature has the largest (in magnitude) loading for the 3rd principal component?

(e) A principal components analysis can be insightful even if the principal components are not explicitly used in an analysis. For example, let B be an 9×9 matrix which has all diagonal entries equal to 1, and all other entries equal to 0, except for column 9, in which all entries equal 1/9. Create new $n \times 9$ feature matrix W:

$$W = XB$$

Based on your answers to part (d), which columns of W correspond approximately to the first 3 principal components?

(f) Using function `cv.glmnet()` create a LASSO fit using W as feature matrix and `class` as a binary response. This means setting option `family='binomial'` and `alpha=1` (why?). Other options can use default settings. Give the coefficients for the model corresponding to the smoothing parameter λ set to `lambda.1se` (what does this mean?). How is this solution related, approximately, to the principal components?

(g) Create the same type of LASSO fit as for part (f), using the original feature matrix X (this need not be centered). Which features have nonzero coefficients? Calculate the predicted values for the fits using W and X as the feature matrix (these need not be transformed by the logistic function). Create a scatter plot of the predicted values (each point represents the pair of predicted values for each observation). Superimpose the identity using `abline(0,1)`. Also, for each fit, use the response/predicted value pair to calculate the AUC statistic for the ROC curve. How similar are the two predictive models?

(h) If two predictive models are similar, we might prefer the simpler or more interpretable model. Consider the model of part (f), based on feature matrix W. We can set $\bar{S} = S/9$, the average of the features on the original 10 point scale.

(i) Interpret the coefficients of the model of part (f). What role does $\bar{S}$ play?

(ii) Identify all observations for which at least one feature F_i has score 10. How many of each class are among these?

(iii) Create a list of 9 vectors, where the ith vector is the subset of all values $\bar{S}$ for observations with $F_i = 10$. Construct a single side-by-side boxplot using these vectors. Superimpose horizontal lines at the $25th$, $50th$, $75th$ percentiles of $\bar{S}$ separately for each class (clearly indicate the classes). How does this plot explain the feature selection of the model of part (f)?