STA414/2104: Practice Problems 7

March, 2018

These practice problems are not for credit. Students may complete independently, in groups, or however they like. Treat these as representative of what might be on the tests.

Questions involving coding may be completed in any language. If you would like help regarding your language of choice from the course team, use R or Python. There will not be any *code*-related questions on the tests, but you will be asked about computational algorithms.

1. Consider the code example regarding the preprocessing of the `ames` housing data discussed in class.
   (a) I ignored missing values. Modify the preprocessing pipeline to include a treatment for them. Suggestions (none of these are that good) are:
   - Impute using the (univariate) mean
   - Impute using the (univariate) median
   - Impute using a KNN algorithm (mean of the K closest other observations)

   (b) I told you that `model.matrix` ignores missing values, so explain why there are still `NA` coefficients in the linear regression model we fit. It may help to consider the `corrplot` we created above.

2. Consider the code example regarding the preprocessing of the `ames` housing data discussed in class. Choose a subset of features using forward stepwise selection by AIC and BIC. Comment on any differences you see between the features selected by the two metrics.

3. *KS Statistic.* Suppose we have a training set $(t_1, \ldots, t_N)$ of binary targets $t_n \in \{0, 1\}$, and we build a binary classification model that yields a soft classification for each point, $\hat{t}_n \in (0, 1)$. Let T_0 refer to the random variable representing a predicted probability for the subset of the training observations for which $t_n = 0$, with CDF $F_0(x) = P(T_0 < x)$, and similarily for the subset of the training observations for which $t_n = 1$. The KS statistic is then

$$KS = \max_x \big| F_0(x) - F_1(x) \big|$$

Consider a strictly monotone function $g(\cdot)$ and define $Y_0 = g(T_0), Y_1 = g(T_1)$ with corresponding CDFs $G_0(x) = P(Y_0 < x)$, $G_1(x) = P(Y_1 < x)$. The KS statistic for Y is then

$$KS_+ = \max_x \big| G_0(x) - G_1(x) \big|$$

Prove that $KS = KS_+$, that is, the KS statistic is invariant under strictly monotone transformations of the predicted probabilities.

4. For the ROC curve as defined in lecture, prove the statement that

$$P(\hat{t}_1 > \hat{t}_2 | t_1 = 1, t_2 = 0)$$

for any randomly selected pair of observations where one is positive and the other is negative.

5. Consider the binary logistic regression model discussed in lecture: for training data $(\mathbf{x}_n, t_n), n = 1 \ldots N$, the distribution of the targets is taken to be $t_n \sim Binom(1, p_n)$ and the predictions are given by

$$\hat{p}_n = y(\mathbf{x}_n, \mathbf{w}) = \frac{1}{1 + \exp\left(-\mathbf{x}_n' \mathbf{w}\right)}$$

   (a) For a **saturated** model, the log-likelihood is technically undefined. But, if we take $0 \log 0$ to be 0 (which is not that odd of a thing to do, since $\lim_{\epsilon \to 0} \epsilon \log \epsilon = 0$), show that the log-likelihood for the saturated model is 0.

   (b) Taking the log-likelihood for the saturated model to be zero as above, show that the deviance for a **null** model is

$$-2 * \left((\#t_n = 1) \times \log \bar{t} + (\#t_n = 0) \times \log(1 - \bar{t}\right)$$

   where $\bar{t}$ is the sample mean of the targets.

   (c) Show that the deviance for a general model is

$$-2 \sum_{n=1}^{N} t_n \log \hat{p}_n + (1 - t_n) \log\left(1 - \hat{p}_n\right)$$

   (d) Write down the algorithm, in detail, for performing forward stepwise selection in logistic regression based on AIC or BIC.