Homework 2

MSA 8150 - Machine Learning for Analytics (Spring 2020)
Instructor: Alireza Aghasi
Due date: Thursday, Feb 27th (1 PM)

February 17, 2020

Please revise the homework guidelines reviewed in the first lecture. Specifically note that:

- Start working on the homework early
- Late homework **is not accepted** and will receive zero credit.
- Each student must write up and turn in their own solutions
- **(IMPORTANT)** If you solve a question together with another colleague, each need to write up your own solution and **need to list the name of people who you discussed the problem with on the first page of the material turned in**
- The homework should be manageable given the material lectured in the class. The long questions are to help clarifying the problem.

Q1. Consider taking n independent samples $x_1, x_2, \cdots, x_n$ from a so called "generalized logistic distribution" with the following p.d.f:

$$f(x) = \frac{\alpha e^{-x}}{(1+e^{-x})^{\alpha+1}}, \qquad \alpha > 0.$$

The parameter α is a positive quantity that is a free model parameter.

(a) We do not know α when taking the independent samples. Show that the maximum likelihood estimate of α is the following:

$$\alpha_{ML} = \frac{n}{\sum_{i=1}^{n} \log\left(1+e^{-x_i}\right)} \tag{1}$$

(b) Consider $n = 10$ and the independent samples to be as follows:

$$x_1 = 0.5377,\ x_2 = 1.8339,\ x_3 = -2.2588,\ x_4 = 0.8622,\ x_5 = 0.3188,$$
$$x_6 = -1.3077,\ x_7 = -0.4336,\ x_8 = 0.3426,\ x_9 = 3.5784,\ x_{10} = 2.7694.$$

Use Equation (1) to calculate α_{ML} for this sample set. We next want to see how reliable our estimate of α_{ML} is, so we need a method to estimate the standard error for α_{ML}. Use the Bootstrap technique with $B = 10,000$ resamplings to estimate the standard deviation for your calculated α_{ML}. Attach your programming code as well.

(c) For a classification problem with only one feature x the label value is binary (the values of the response only take two values: $y = 1$ and $y = 2$). We use a QDA approach and after fitting the QDA parameters (as explained in the slides) we obtain the following parameters:

$$\pi_1 = \frac{1}{3},\quad \pi_2 = \frac{2}{3},\quad \mu_1 = -3,\quad \mu_2 = 1,\quad \sigma_1 = 1,\quad \sigma_2 = 2.$$

Find the points x which are on the decision boundary (here there would be only two points).

Hint: Note that all you need to do is to find the points x such that $\pi_1 f_1(x) = \pi_2 f_2(x)$. You will end up with a quadratic equation, and then you would need to find the two solutions.

Q2. Along with the homework package, you can access a data file "AutoHighLow.csv", which reports the Mile per Gallon (mpg01) along with some other car features for various automobile instances. The "mpg01" is a categorical variable which will be treated as the response in our problem. All other variables: "cylinders, displacement, horsepower, weight, acceleration, year, origin" will be treated as the features. The data are already randomly blended.

(a) **Validation Set Approach.** Split the data into two sets, the first 350 rows as the training set and the next 42 rows as the validation set (representative of your test set). Train classification models which predict mpg01 using the other features (cylinders, displacement, horsepower, weight, acceleration, year, origin).

– Use logistic regression for your classification. Report the p-values associated with the intercept and all the features. Based on the p-values, which features might be problematic? Also, report your accurate classification rate by applying your model to the validation set.

Note: your accurate classification rate is the ratio of the number of correct predictions to the total number of samples – which is 42 in this example. For instance if `Y.t` is a vector that contains the actual mpg01 values in your validation set and `Y.hat.t` contains your predicted labels for this set, the accurate classification rate can be obtained via:

```
acc.cls.rate = mean(Y.t == Y.hat.t)
```

– Apply LDA and QDA, and again report your accurate classification rate with reference to the validation set.

– Apply the K-nearest neighbor classification (you can see examples of KNN in lecture 4 R code example) with $K = 1, 5, 10$ neighbors and report your accurate classification rate for each K (again by predicting the validation set responses).

– Among logistic regression, LDA, QDA and KNN with $K = 1$, $K = 5$ and $K = 10$, which model seemed the most accurate one?

(b) In part (a) all features are treated as numerical quantities. You might realize that the variable "origin" only takes values 1, 2 or 3. Repeat all the steps of part (a), this time treating the feature "origin" as a categorical variable. (Hint: if you look at mod3 in LinearRegressionReview.R file of lecture 4, you see an example of handling categorical features)

(c) **Cross Validation.** In lecture 4 and specifically the R example (CVExample.R) you learned how to apply the LOOCV and K-Fold CV to regression problems. In this homework we would like to apply the LOOCV and K-Fold CV to our logistic regression, LDA and QDA models. Using $K = 10$ folds, LOOCV fit the models and report the classification accuracy for the 3 models (logistic regression, LDA and QDA). For this question use $mpg01$ as the response variable and all the other variables as features (consider all variables to be numerical). You may notice that there is not much difference between the classification accuracies of the three models when it comes to LOOCV and K-fold CV evaluations.

Hint: if you like to see an example of K-fold CV for a classification problem and get some idea on how to modify it for the LOOCV, you may find the R "caret" package and the following links useful:
`https://www.r-bloggers.com/evaluating-logistic-regression-models/`
`https://www.rdocumentation.org/packages/caret/versions/6.0-78/topics/trainControl`
Note that for part (c) you would need to use the entire data and you can determine the classification accuracy as one of the outputs of the "train" function in R.