Homework 2

MSA 8150 - Machine Learning for Analytics (Spring 2019)
Instructor: Alireza Aghasi
Due date: Wednesday Feb 27th (1 PM)

February 15, 2019

Please revise the homework guidelines reviewed during the first session. 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 totally manageable given the material lectured in the class. The long questions are to help clarifying the problem, the majority of the solutions are short

Q1. Consider taking n independent samples $x_1, x_2, \cdots, x_n$ from the following distribution

$$f(x) = \frac{\theta^4 x^3 \exp(-\theta x)}{6}, \quad \text{where} \;\; x > 0, \;\; \theta > 0.$$

(a) We do not know θ when taking the independent samples. By maximizing the log-likelihood, show that the maximum likelihood estimate of θ is the following:

$$\theta_{ML} = \frac{4n}{\sum_{i=1}^{n} x_i} \tag{1}$$

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

$$x_1 = 0.3057, \; x_2 = 0.7227, \; x_3 = 1.1566 \; x_4 = 2.8622, \; x_5 = 1.3588,$$
$$x_6 = 0.5377, \; x_7 = 0.4336, \; x_8 = 0.3426, \; x_9 = 3.5784, \; x_{10} = 2.7694.$$

Use Equation (1) to calculate θ_{ML} for this sample set. Use the Bootstrap technique with $B = 50,000$ resamplings to estimate the standard deviation for your calculated θ_{ML}. Attach your programming code as well.

(c) In part (b), how much is your standard deviation estimate if you use $B = 1000$. Do you see a lot of difference between the two values?

Q2. For a classification problem with two features (that is, our feature vector $\boldsymbol{x}$ has two components, i.e., $\boldsymbol{x} = (x_1, x_2)^\top$) the label value is binary (the values of the response only take two values). We use an LDA approach and after fitting an LDA model we obtain the following parameters:

$$\pi_1 = \frac{1}{3}, \; \pi_2 = \frac{2}{3}, \; \boldsymbol{\mu}_1 = \begin{pmatrix} -3 \\ 2 \end{pmatrix}, \; \boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 2 \end{pmatrix}.$$

Show that the line that separates the two classes (the decision boundary) is the following line:

$$12x_2 - 27x_1 = 31.5 + \log(2) \tag{2}$$

Hint. The decision boundary is where $\mathbb{P}(y = 1|\boldsymbol{x}) = \mathbb{P}(y = 2|\boldsymbol{x})$, where for the LDA happens where $\pi_1 f_1(\boldsymbol{x}) = \pi_2 f_2(\boldsymbol{x})$.

Q3. The goal of this question is predicting the heart health of patients in a hospital. In the homework package, you can access the data file HeartData.csv, which consists of 13 features and one response variable (num). The features represent some measurements of the patients' health atributes and num is an indication of the heart health. If num = 0, the heart is healthy, and if num = 1, it reports an issue. Below we summarize a brief description of each feature:
age: Age of patient
sex: Sex, 1 for male
cp: chest pain
trestbps: resting blood pressure
chol : serum cholesterol

fbs: fasting blood sugar larger 120mg/dl (1 true)
restecg: resting electroc. result (1 anomality)
thalach: maximum heart rate achieved
exang: exercise induced angina (1 yes)
oldpeak: ST depression induc. ex.
slope: slope of peak exercise ST
ca: number of major vessel
thal: no explanation provided, but probably thalassemia (3 normal; 6 fixed defect; 7 reversable defect)
num: diagnosis of heart disease (angiographic disease status)

(a) Consider splitting the data into a a training and test set. Samples 1 to 200 form the training set and samples 201 to 297 form the test set. Try the following classification models to predict "num" in terms of the other features in the dataset:

- Use logistic regression for your classification. Report the p-values associated with the intercept and all the features. Which features have large p-values? Use the test data to estimate the accuracy of your model.
- Apply LDA and QDA, and again report your model accuracies using the test data.
- 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 test accuracies.
- Among logistic regression, LDA, QDA and KNN with different number of neighbors, which model seemed the most accurate one?

(b) In part (a) we treated all features as numerical quantities. As you may see, some features take categorical values. Repeat all the steps of part (a), this time treating the feature "sex", "cp", "fbs", "slope", "exang", "ca" and "thal", as a categorical variables. (instances of handling categorical features are available in LinearRegressionReview.R file of lecture 4 material)

(c) 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 10-fold CV and LOOCV fit the models and report the classification accuracy for the 3 models (logistic regression, LDA and QDA). For this question use *num* as the response variable and all the other variables as features (again consider the features in part (b) as categorical). Note that for this part you would need to use the entire data when performing the K-fold CV and LOOCV (no more test/training splitting). (Hint: you may find the R "caret" package and function "trainControl" useful for this problem)

Is there much difference between the test accuracies of the three models when it comes to LOOCV and K-fold CV?