Homework 3

AI 539 - Machine Learning for Non-AI Majors
Instructor: Alireza Aghasi
Due date: See Canvas

April 25, 2024

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. The goal of this problem is to implement your own version of logistic regression, and compare it to the output of the Python package.

(a) Load the data file `BinaryData.csv` and perform a simple logistic regression in the programming language of your choice, predicting the class y based on x. Report the values of β_0 and β_1.

(b) Now lets work on implementing our own version of logistic regression, and understand its basics. To start, consider the function

$$f(z) = \alpha \log\left(1 + e^{-z}\right) + (1-\alpha)\log\left(1+e^{z}\right), \qquad\qquad 0 \le \alpha \le 1,$$

where α is a known coefficient between 0 and 1. Show that $z = \log(\frac{\alpha}{1-\alpha})$ is a stationary point (point of zero derivative) for $f(z)$.

(c) Show that $f(z)$ is convex (the second derivative test might be the easiest).

(d) Now that you know $f(z)$ is convex, is $z = \log(\frac{\alpha}{1-\alpha})$ a minimizer or a maximizer? Why?

(e) Plot $f(z)$ for $\alpha = 0.3$, and the values of z between -3 and 3.

(f) In the class we learned that sum of convex functions is convex. Furthermore, we showed that if $f(z)$ is convex, $f(\beta_0 + \beta_1 x)$ is also convex. This is an indication that the logistic loss

$$L(\beta_0,\beta_1) = \sum_{i=1}^{n} y_i \log\left(1 + e^{-\beta_0-\beta_1 x_i}\right) + (1-y_i)\log\left(1+e^{\beta_0+\beta_1 x_i}\right).$$

is convex. Now, derive an expression for

$$\frac{\partial L}{\partial \beta_0} = \ldots, \qquad , \frac{\partial L}{\partial \beta_1} = \ldots.$$

Simplify the expressions in a way that the end results only involve `sigmoid` functions and not the log or exp functions. Expressions like $\sum_{i=1}^{n} w_i \texttt{sigmoid}(w_i') + w_i''$, where w_i, w_i', w_i'' are expressions in terms of the problem parameters.

(g) Use the data file `BinaryData.csv` in part (a) and set up $L(\beta_0,\beta_1)$ for the x_i and y_i in the dataset. Write a gradient descent (GD) scheme to minimize $L(\beta_0,\beta_1)$ in Matlab or Python. For your scheme use a learning rate of $\eta = 0.01$, and run the GD for 500 iterations. As the initial values for β_0 and β_1 you can use zero (clearly, since the problem is convex, the initialization does not matter and we will converge to the global minimizer no matter where we start). Attach all your code and results.

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:

$$12x_2 - 27x_1 = 31.5 + \log(2) \tag{1}$$

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)

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.

– Among logistic regression, LDA, and QDA which model(s) seems the most accurate one(s)?

Q4. A study analyzes the data on law school admission, and the goal is to examine the correlation between LSAT score and the first year GPA. For each of 15 law schools, we have the pair of data points (LSAT, GPA) as

$$\begin{array}{lll} (576, 3.93), & (580, 3.07), & (653, 3.12) \\ (635, 3.30), & (555, 3.00), & (575, 2.74) \\ (558, 2.81), & (661, 3.43), & (545, 2.76) \\ (578, 3.03), & (651, 3.36), & (572, 2.88) \\ (666, 3.44), & (605, 3.13), & (594, 2.96). \end{array}$$

(a) Calculate the correlation coefficient between LSAT and GPA.

(b) Pick the programming language of your choice, and use bootstrapping to estimate the standard deviation of the correlation coefficient. Use $B = 1000$ bootstrap resamples. Also plot a histogram of the results (use 20 bins).