دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

به نام خدا

دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)

دانشکده برق

یادگیری ماشین – نیم‌سال دوم ۱۴۰۱-۱۴۰۲

تمرین عملی لاجستیک رگرسیون و رگرسیون، سری سوم، درس یادگیری ماشین

Code for all programming assignments should be well documented. **A working program with no comments will receive only partial credit**. Documentation entails writing a description of each function/method, class/structure, as well as **comments throughout the code** to explain the program flow. Programming language for the assignment is **Python**.

Following libraries can be used when necessary:

- Matplotlib, NumPy, SciPy, Pandas and other basic libraries.
- libraries for calculating mean, variance and covariance.

Following libraries mustn't be used in any way:

- Libraries for making classifiers (You must implement classifiers from scratch).

Collaboration Policy

You are to complete this assignment individually. However, you may discuss the general algorithms and ideas with classmates, TAs, peer mentors and instructor in order to help you answer the questions. But we require you to:

- not explicitly tell each other the answers
- not to copy answers or code fragments from anyone or anywhere
- not to allow your answers to be copied
- not to get any code from the Web

If you have any questions regarding this assignment, please contact Mr. Janani and Mr. Aghdasian.

Telegram ID: @pooya_9877 @A_Aghdasian

Submit by 28th Ordibehesht 1402, 11.59pm

Question 1: Logistic Regression [30 point]

In this section, with the MIT-BIH Arrhythmia Dataset, you must implement the logistic regression algorithm to classify the heartbeat class. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias and myocardial infarction. These signals are preprocessed and segmented, with each segment corresponding to a heartbeat.

a) Implement the logistic regression algorithm with the gradient ascent algorithm. By changing the learning rate in the interval [1,0), check the effect of the learning rate on the convergence speed. Find the appropriate learning rate and training stopping point using a validation set that is randomly selected up to 20% of the training data. Then plot the accuracy curve on the training and test data. Also find the confusion matrix and report the training duration.

b) Implement a GNB algorithm and compare the results with 1(a). To avoid zero probability, consider the prior distribution as a Gaussian distribution (hint: Be sure to use the logarithm). Also find the confusion matrix and report the training duration.

c) Implement the logistic regression algorithm with the regularization form by choosing the appropriate parameter, and compare it with result of question 1(a) and 1(b). Also find the confusion matrix and report the training duration.

d) For logistic regression (the best result) and GNB, plot the training and test curve based on different training samples (classification accuracy rate based on the number of training samples). For this purpose, start the amount training data from a minimum of a quarter of the number of each class, and take the step of increasing about one eighth of the amount of data of that class.

Question 2: Logistic Regression [30 point]

a) For training of the logistic regression of question 1(a) use 5-fold-cross validation. Then test it on the entire test data and finally report the average classification accuracy on the training and test data. Repeat this act of randomly dividing the data for training 5 times and determine the accuracy values each time, and finally get the average accuracy. Compare all the results with result of question 1(a) and 1(b).

b) Do the previous step this time for the regularized form of logistic regression. Compare all the results with results of question 1(a) and 1(b) and 2(a).

Attach the written programs along with the results and figures and analyzes requested in each step. Also, do a discussion about the results and the effect of overfitting to training data and comparison between GNB and logistic regression in different sizes of training data (speed and accuracy).

Question 3: linear Regression [40 point]

The purpose of this section is to implement an algorithm to estimate data based on the linear regression algorithm. The attached data is related to the average house price in different areas of California. This data includes 8 numerical features and one nominal feature. The number of samples is 20640. And the house prices are in the value_house_median column.

a) **Pre-processing the data**: To implement the regression algorithm, you must first prepare the data. For this purpose, keep the following points in mind:

- **Missing values:** There are several strategies for dealing with this type of data, such as eliminating that feature, removing samples with missing values, and so on. In this exercise replace the missing values with the average of the rest of the data.
- **Scaling data:** Numerical features are in different ranges, for example, the age of buildings is in the range of 1-50 years and the population of areas is in the range of 1-35000, convert the range of features to 0-1. (This will help the gradient descent algorithm work better and also makes the features comparable due to the same scaling)

- **Dummy variable indicator method:** As you know, the regression algorithm works with numerical values, use this method to convert non-numerical values into numbers, in such a way that for each unique value in the non-numerical feature, a variable is added to the dataset in the form of a column, and the binary value 1 or 0 indicates whether or not it belongs to that category. For more information, please search about it.

b) Split the data randomly into three parts, 60% - 20% - 20% respectively for training, validation and test set.

c) Run the linear regularization algorithm with regularization and early stopping as follows:

- For 5 learning rate 0.1, 0.3, 0.5, 0.7 and 1.
- For 4 regularization coefficients 0, 0.1, 1, 10.
- Plot the loss curves for the training and validation data in one plot.(Totally 20 plots.)

d) Finally, choose the best parameters and report the cost value in the test data with these parameters.

e) We want to extract more information from the data, go back to the scaling data level, calculate and add the following two features for each sample. Then scale the data.

- Population_per_household = $\frac{population}{households}$
- rooms_per_household = $\frac{total\_rooms}{households}$

plot the training process (loss values for train and validation data) with the best previous parameters (learning rate and regularization coefficient) and report the loss value for the test set.

What to submit:
• Code
• A short write-up about your implementation with results and your observations from each result.

Good luck 😊