دانشگاه شیراز
Shiraz University

Artificial Intelligence (June 2020)

Homework#4: Linear Regression, Weighted Linear Regression, and Classification

Due date: 18th July 2020

In order to do this assignment, you have to go through linear and logistic regression theories and concepts. In this assignment, you're provided with 2 identical datasets, one dataset is normal(Dataset1) and other dataset have some outliers(Dataset2). The datasets include information about height, weight and gender (label) of some random people. You should ignore third column (label column) and only work with height, and weight features.

Part A: Linear Regression

Dataset: ***Dataset1.txt, Dataset2.txt***

In this assignment, you need to implement closed form solution (that contains least square objective function), Gradient Descent algorithm (Batch or Stochastic with enough epochs) for linear regression on heights (X) and weights (Y). Please perform the following tasks:

- Normalize the datasets so that the values of each feature change between 0 (for min. value of the feature) and 1 (for max. value of the feature). Note that this task is very important for the desired results of the upcoming tasks.

- Train each model separately on the normalized datasets and plot the datasets alongside with the obtained regression model. For these plots, the X axis should be the height feature and the Y axis should be the weight feature.

- One of these datasets have some outliers(Dataset2). Does it affect the robustness of the model? Explain. (think about effect outlier on regression)
- Plot $J(\Theta)$ in terms of $\Theta1$ in [-2:2] and $\Theta2$ in [-2:2]. (in 2D and 3D figure) (*what is $\Theta1$ and $\Theta2$? Think about line equation. $\Theta1$ can be slope and $\Theta2$ can be y-intercept and in linear regression you should estimate line parameters)*

- Explain what does the normalization process do? When would it be useful to normalize the data?
- In your class you learned about sigmoid function. Explain when we use this function in learning and what is the output of that. Plot sigmoid function (binary sigmoid).

Part B: Weighted Linear Regression

Dataset: ***Dataset2.txt***

The effect of outliers on linear regression method have been analyzed in the previous part. Here, you have to apply weighted linear regression on the dataset which includes outliers.

- Derive the closed form solution for weighted linear regression(WLR).
- Propose a weighting function which decreases the effect of outliers with a formula. Explain why it could be appropriate.
- Normalize the dataset (similar to the previous part).
- Apply weighted linear regression using your suggested weighting function and find:
  - The closed form solution.
  - gradient descent (batch or stochastic) solution (with enough epochs).
- Plot the outlier dataset and the models obtained from both the previous and the current part on the same figure. Compare the results of parts A and B. (2 figures)
- When and how does WLR work better than simple linear regression?

Part C: Classification(Logistic Regression)

Dataset: Iris https://archive.ics.uci.edu/ml/datasets/Iris

Section A (Binary Classification)

- The Iris dataset consists of 4 features and 3 classes. Use only the first and second features (remove the third and fourth columns) and also delete the instances of the 'Iris-versicolor' class to reduce the data to 2 classes with 2 features.

- Consider the first 80% of the data in each class for train and the rest 20% for test
- Report the training and testing errors, and the equation of the decision boundary.
- Also, plot the decision boundary along with the samples of the two classes with different colors all in one plot.

Notes:

- **Pay extra attention to the due date. It will not extend.**
- **Be advised that submissions after the deadline would not grade.**
- **Your implementation should be functional.**
- **Prepare your full report in PDF format and include the figures and results.**
- **Do not use sklearn or any similar library for regression and logistic regression and write your own code.**
- **The allowed programming languages are any language and feel free**
- **Feel free for using sklearn in python for load iris or split train and test dataset**
- **Submit your assignment using a zipped file with the name of "StdNum_FirstName_LastName.zip" to compuscien@gmail.com with AI2020-HW#4 subject.**
- **Maybe this homework has an online presentation and attention that**