CS57300 Spring 2018 Homework — Homework 1 — Last update: February 5, 2018

https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2018/index.html

Due: February 7th 6:00am, 2018

Instructions and Policy: Each student should write up their own solutions independently, no copying of any form is allowed. You MUST to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.

YOU MUST INCLUDE YOUR NAME IN THE HOMEWORK

You need to submit your answer in PDF. LaTeX is typesetting is encouraged but not required. Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.

Q0 (0pts correct answer, -1,000pts incorrect answer: (0,-1,000) pts): A correct answer to the following questions is worth 0pts. An incorrect answer is worth -1,000pts, which carries over to other homeworks and exams, and can result in an F grade in the course.

(1) Student interaction with other students / individuals:

(a) I have copied part of my homework from another student or another person (plagiarism).

(b) Yes, I discussed the homework with another person but came up with my own answers. Their name(s) is (are) ______________________________

(c) No, I did not discuss the homework with anyone

(2) On using online resources:

(a) I have copied one of my answers directly from a website (plagiarism).

(b) I have used online resources to help me answer this question, but I came up with my own answers (you are allowed to use online resources as long as the answer is your own). Here is a list of the websites I have used in this homework:

____

(c) I have not used any online resources except the ones provided in the course website.

Q1 (6 pts): Classification using Python

Please download the dataset at
`https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2018/data/creditcard.csv`
and answer the following questions. This dataset contains features of a credit card transition and whether or not the transaction was reported as fraudulent. You are **not** allowed to redistribute this dataset.

Train a logistic regression classifier using the function `sklearn.linear_model.LogisticRegression` in the sklearn library, with L2 regularization and $C = 10^{-8}$ over this dataset. Note that you should remove "Time" from the set of features. The goal is to predict "Class". Use the following code to report the accuracy of your logistic classifier over the positive (Class=1) and negative (Class=0) classes.

```
from sklearn.metrics import confusion_matrix


CM = confusion_matrix(target_train, pred_train)
print("Accuracy of positive class:",CM[1,1]/(CM[1,0]+CM[1,1]))
print("Accuracy of negative class:",CM[0,0]/(CM[0,0]+CM[0,1]))
```

Answer the following questions:

(1) (1pt) Explain the main reason why the accuracy over the training data for the positive class is smaller than that of the negative class. (PDF)

(2) (1pt) Describe **two** effective solutions to increase the accuracy of the positive class. (PDF)

(3) (1pt) Describe the effect of increasing or decreasing C on the classifier accuracy over the different classes. Derive the relationship between C and σ^2 of slide 17, Lecture 7 slides. Use Bayes theorem to explain why one of the classes is affected more by the regularization than the other class. (PDF)

(4) (1pt) Describe how we can best fix the accuracy issue observed in Q1(1) due to the class imbalance. Write the new logistic log-likelihood function of the dataset. Show all the necessary steps used to perform the calculation. (PDF)
Hint: Revisit the data selection bias (Lecture 5, slide 33)..

(5) (1pt) Note that the log-likelihood derived above can be implemented in the original `sklearn.linear_model.LogisticRegression` just using the parameter `class_weight` (you are **not** allowed to use `class_weight = "balanced"` to solve this problem, you must perform your own class weight computation). Report the new positive and negative class accuracies. **(PDF + uploaded code using turnin as `hw1q1_5.py`)**.

(6) (1pt) Redo Q1(4) now considering the SVM objective (score) function (Lecture 6, slide 14). Write the new objective function for SVM. Show all the necessary steps used to perform the calculation. (PDF)

Q2 (2 pts): Linear regression with priors

Given a dataset $\{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^n$ of n samples from an unknown population, where x is a set of real-valued features and y is a real-valued target variable. A linear model to predict the labels y takes the form

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

(1) (1pt) Consider the distribution of $\boldsymbol{\varepsilon}$ to be Gaussian $N(0, \lambda\mathbf{I})$, what is the distribution $P[y_i|\mathbf{x}_i, \boldsymbol{\beta}]$?

(2) (1pt) Assume $\boldsymbol{\beta} = \text{Normal}(0, \sigma\mathbf{I})$. Give the maximum a posteriori (MAP) estimate equation of $\boldsymbol{\beta}$ given $\{y_i, \mathbf{x}_i\}_{i=1}^n$ (you need to write the estimator as: $\arg\max_a f(a)$, with the correct variable a and the correct score function $f(a)$ for this problem).

Q3 (2 pts): Decision Boundary and Classification

Consider a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$ for a binary classification task with labels y={+1,-1} and $\mathbf{x} \in \mathcal{R}^p$. Let us define the mean of two classes as

$$\boldsymbol{\mu}_+ = \frac{1}{n_+} \sum_{y_i=+1} \mathbf{x}_i$$

$$\boldsymbol{\mu}_- = \frac{1}{n_-} \sum_{y_i=-1} \mathbf{x}_i$$

where $n_+ = \sum_{i=1}^n I_{\{y_i=+1\}}$ and $n_- = \sum_{i=1}^n I_{\{y_i=-1\}}$.

For a new data point $\mathbf{x}$ we check the **Euclidean distance** of the new point from $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ and assign the label of the class for which this distance is smaller.

Question: Find the decision boundary between the two classes. (Hint: It is of the form $\mathbf{w}^T\mathbf{x}$ + b. Express $\mathbf{w}$ and b in terms of $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$)