Homework Assignment # 3

Due: Wednesday, November 16, 2016, 11:59 p.m.
Total marks: 100

Question 1. [60 MARKS]

In this question, you will implement several binary classifiers: naive Bayes, logistic regression and a neural network. An initial script in python has been given to you, called `script_classify.py`, and associated python files. You will be running on a physics dataset, with 8 features and 100,000 samples (called `susysubset`). The features are augmented to have a column of ones, in `dataloader.py` (not in the data file itself). Baseline algorithms, including random predictions and linear regression, are used to serve as sanity checks. We should be able to outperform random predictions, and linear regression for this binary classification dataset.

(a) [15 MARKS] Implement naive Bayes, assuming a Gaussian distribution on each of the features. Try including the columns of ones and not including the column of ones in the predictor. What happens? Explain why.

(b) [15 MARKS] Implement logistic regression.

(c) [20 MARKS] Implement a neural network with a single hidden layer, with the sigmoid transfer.

(d) [10 MARKS] Briefly describe the behavior of these classification algorithms you have implemented. You do not need to make claims about statistically significant behavior, but report average error and standard error. You do not need to run on the whole dataset.

Question 2. [20 MARKS]

Consider a classification problem where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{0, 1\}$. Based on your understanding of the maximum likelihood estimation of weights in logistic regression, develop a linear classifier that models the posterior probability of the positive class as

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{2}\left(1 + \frac{\mathbf{w}^\top \mathbf{x}}{\sqrt{1 + (\mathbf{w}^\top \mathbf{x})^2}}\right)$$

Implement the iterative weight update rule and compare the performance on the physics dataset to logistic regression. As before, you do not need to check for statistically significant behavior, but in a few sentences describe what you notice.

Question 3. [20 MARKS]

In this question, you will add regularization to logistic regression, and check the behavior again on the expanded physics dataset, which has 18 features (called `susy_complete` in the code). Note that using regularization on the base physics dataset, which only has 9 features, would not have as strong of an effect; with more features, the regularization choice is likely to have more impact. For all the three settings below, implement the iterative update rule and in a few sentences briefly describe the behavior, versus vanilla logistic regression and versus the other regularizers.

(a) [5 MARKS] Explain how you would add an ℓ_2 regularizer on $\mathbf{w}$ and an ℓ_1 regularizer on $\mathbf{w}$. Implement both of these.

(b) [10 MARKS] Pick a third regularizer of your choosing, and explain how you would learn with

this regularizer in logistic regression (i.e., provide an iterative update rule and/or an algorithm). Implement this regularization.

(c) [5 MARKS] In a few sentences, briefly describe the behavior of the regularizers.

Homework policies:

Your assignment will be submitted as a single pdf document and a zip file with code, on canvas. The questions must be typed; for example, in Latex, Microsoft Word, Lyx, etc. or must be written legibly and scanned. Images may be scanned and inserted into the document if it is too complicated to draw them properly. All code (if applicable) should be turned in when you submit your assignment. Use Matlab, Python, R, Java or C.

Policy for late submission assignments: Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rule:

on time: your score 1

1 day late: your score 0.9

2 days late: your score 0.7

3 days late: your score 0.5

4 days late: your score 0.3

5 days late: your score 0.1

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments are individual, except when collaboration is explicitly allowed. All the sources used for problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

Good luck!