Homework Assignment # 4
Due: Friday, April 8, 2022, 11:59 p.m.
Total marks: 100

Question 1. [40 MARKS]

In this question, you will implement logistic regression for binary classification. Initial code has been given to you, in `A4.jl`. You will be running on a physics data set, with 8 features and 100,000 samples (called susysubset). The features are augmented to have a column of ones (to create the bias term) in the code (not in the data file itself). We should be able to outperform random predictions, provided by a random classifier.

(a) [25 MARKS] Implement a mini-batch stochastic gradient descent approach to logistic regression, using RMSProp. Report the error, using the number of epochs given in the script.

(b) [15 MARKS] Implement a mini-batch stochastic gradient descent approach to logistic polynomial regression. As a hint, consider calling the `LogisticRegression` algorithm you wrote in (a), within `PolynomialLogisticRegression`, to avoid code duplication and to re-use an already debugged algorithm. Report the error, using the same parameters and step size as in (a).

Question 2. [40 MARKS]

In this question, you will use the paired t-test to compare the performance of two models. You will compare the above `LogisticRegression` model and `PolynomialLogisticRegression` model, both using RMSProp. You will run this comparison using `A4.jl`. You hypothesize that `PolynomialLogisticRegression` is better than `LogisticRegression`, and so want to run a one-tailed test to see if that is true.

(a) [5 MARKS] Define the null hypothesis and the alternative hypothesis. Use μ_1 to be the true expected squared error for `LogisticRegression` and μ_2 the true expected squared error for `PolynomialLogisticRegression`.

(b) [15 MARKS] Before running the paired t-test, you should check if the assumptions are not violated. One way to satisfy the assumption for the paired t-test is to check if the errors are (approximately) normally distributed with (approximately) equal variances. To do this, you need to implement the `checkforPrerequisites` method in `A4.jl`. For each model, you can plot a histogram of its errors on the test set. You can do so using the two vectors of errors and the function `plotTwoHistograms` function to visualize the error distributions simultaneously. Discuss why it is ok or not ok to use the paired t-test to get statistically sound conclusions about these two models, based on your histograms.

(c) [15 MARKS] Regardless of the outcome of Part b, let's run the paired t-test. (Note, we are not advocating that you check for violated assumptions and then ignore the outcome of that step. The goal of this question is simply to give you experience actually running a statistical significance test. Presumably, in practice, you would pick an appropriate one after verifying assumptions). To run this test, you need to compute the p-value. Implement the `getPValue` method, which returns the p-value for the one-tailed paired t-test.

(d) [5 MARKS] Report the p-value. Would you be able to reject the null hypothesis with a significance threshold of 0.05? How about of 0.01?

Question 3. [20 MARKS]

In your implementation you measured the accuracy of your classifier using the 0-1 cost. We can write this cost as $1(\hat{y} \neq y)$ for prediction $\hat{y}$ and observed target y. The generalization error (the expected cost) for your binary classifier $f(\mathbf{x})$, across all pairs $(\mathbf{x}, y)$ is

$$\text{GE}(f) \doteq \mathbb{E}[1(f(\boldsymbol{X}) \neq Y)] \tag{1}$$

where $\boldsymbol{X}, Y$ are the random variables with instances $\mathbf{x}, y$ drawn from joint distribution $p(\mathbf{x}, y)$.

(a) [5 MARKS] Assume you are given $\mathcal{D}_{\text{test}} = \{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}_{i=1}^m$, where we use the tilde notation above these variables to distinguish them from the pairs used in the training set. Write the formula for estimate the $\text{GE}(f)$ using a sample average on $\mathcal{D}_{\text{test}}$.

(b) [5 MARKS] When we talked about squared costs and GE, we found that the GE decomposed into reducible error and irreducible error. We have a similar decomposition for the 0-1 cost for classification, though instead of equality we only have an upper bound

$$\mathbb{E}[1(f(\boldsymbol{X}) \neq Y)] \leq \underbrace{\mathbb{E}[1(f(\boldsymbol{X}) \neq f^*(\boldsymbol{X}))]}_{\text{reducible error}} + \underbrace{\mathbb{E}[1(f^*(\boldsymbol{X}) \neq Y)]}_{\text{irreducible error}} \tag{2}$$

where $f^*(\mathbf{x}) = \arg\max_{y \in \{0,1\}} p(y|\mathbf{x})$ is the optimal predictor that uses the true probabilities $p(y|\mathbf{x})$ (not estimated ones). Imagine you have a huge dataset of billions of samples, and you learn f_1 with logistic regression and f_4 with polynomial logistic regression with $p = 4$. Do you think f_1 or f_4 will have lower reducible error? Explain your answer in a few sentences.

(c) [5 MARKS] Now imagine that you learn f_4 on a smaller dataset and notice that the weights $\mathbf{w}$ are high magnitude, potentially indicating overfitting. Explain one strategy to address this, in a few sentences.

(d) [5 MARKS] Give an example in classification to explain the irreducible error. Make sure your example highlights why the irreducible error is non-zero. Be specific in your example, with a concrete example of targets and the features in $\mathbf{x}$.