Homework Assignment # 4

Due: Friday, April 7, 2023, 11:59 p.m.

Total marks: 100

Question 1. [50 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) [30 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. Completing this question spans multiple parts of the Julia file, namely Q1a, c, d, e and f.

(b) [20 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. [30 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) [20 MARKS] Now let's run the paired t-test. (Note: we should actually check for violated assumptions. In the julia code, we do visualize the errors so you can see for yourself if it is acceptable to use the t-test). To run this test, you need to compute the p-value. Implement the `tDistPValue` method, which returns the p-value for the one-tailed paired t-test.

(c) [5 MARKS] Report the p-value. Would you be able to reject the null hypothesis with a significance threshold 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

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

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 to estimate the GE(f) using a sample average on $\mathcal{D}_{\text{test}}$.

(b) [10 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] 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 $\mathbf{x}$.

Homework policies:

Your assignment should be submitted as two pdf documents and a .jl notebook, on eClass. **Do not** submit a zip file with all three. One pdf is for the written work, the other pdf is generated from .jl notebook. The first pdf containing your answers of the write-up questions must be written legibly and scanned or must be typed (e.g., Latex). This .pdf should be named Firstname_LastName_Sol.pdf, For your code, we want you to submit it both as .pdf and .jl. To generate the .pdf format of a Pluto notebook, you can easily click on the circle-triangle icon on the right top corner of the screen, called Export, and then generate the .pdf file of your notebook. The .pdf of your Pluto notebook as Firstname_LastName_Code.pdf while the .jl of your Pluto notebook as Firstname_LastName.jl. All code should be turned in when you submit your assignment.

Because assignments are more for learning, and less for evaluation, grading will be based on coarse bins. **The grading is atypical**. For grades between (1) 80-100, we round-up to 100; (2) 60-80, we round-up to 80; (3) 40-60, we round-up to 60; and (4) **0-40, we round down to 0**. The last bin is to discourage quickly throwing together some answers to get some marks. The goal for the assignments is to help you learn the material, and completing less than 50% of the assignment is ineffective for learning.

We will not accept late assignments. Plan for this and aim to submit at least a day early. If you know you will have a problem submitting by the deadline, due to a personal issue that arises, please contact the instructor as early as possible to make a plan. If you have an emergency that prevents submission near the deadline, please contact the instructor right away. Retroactive reasons for delays are much harder to deal with in a fair way. If you submit a few minutes late (even up to an hour late), then this counts as being on time, to account for any small issues with uploading in eClass.

All assignments are individual. All the sources used for the 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 the University of Alberta Code of Student Behaviour.

Good luck!