CS57300 Homework — Homework 4 — Last update: April 3, 2016

https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/index.html

Due: April 8th, 2016

Instructions and Policy: Each student should write up their own solutions independently. You need 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

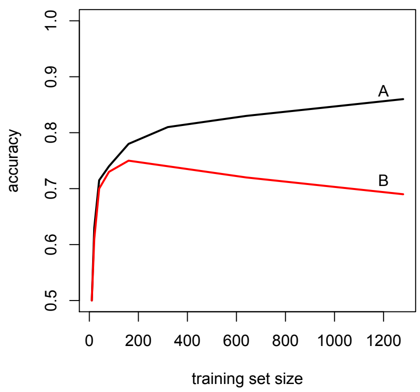
You need to submit your TYPED answer in PDF via Blackboard. LaTeX is typesetting is encouraged but not required. Unless directed you are not required to submit your R and Python codes. Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.

Q1 (1 pts):
Describe the difference between BIC and AIC for model selection. When should you choose BIC over AIC? And vice-versa?

Q2 (2 pts):

(i) In the figure to the right, the plot shows two learning curves for a model M as the size of the training set is increased. Which line reports the results of the evaluation on the:

Training set: __________ Test set: __________



(ii) Is the model M overfitting? Why or why not? Explain how the data in the figure supports your conclusion.

Q3 (4 pts):
You need to decide if the bank should or should not lend money to new costumers. The bank should not lend money for costumers that do not repay the full amount they borrowed. Download the training and validation sets at
`https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/hw/hw4/new_bank_data.zip`
The file `Bank_Data_Train.csv` is the same training data used in the midterm.
The file `Kaggle_Public_Validation.csv` is the test data used by Kaggle.com for its public leaderboard, which we will now use as validation data. These files have the following data format separated by commas:

1. Record Number
2. Amount Requested
3. Interest Rate Percentage
4. Loan Length in Months
5. Loan Title
6. Loan Purpose
7. Monthly Payment
8. Total Amount Funded
9. Debt-To-Income Ratio Percentage
10. FICO Range (`https://en.wikipedia.org/wiki/FICO`)
11. Status (1 = Paid or 0 = Not Paid)

Download the python3 code at
`https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/hw/hw4/LR.TrainingSz.py`
Read and run the code (using python3) and answer the following questions related to its output:

(i) (Theory) Note that we use L_2 regularization. What are the mean and variance of the Normal priors over the Logistic parameters w? (relate these to the value of $\sigma^2/\sigma^2_{\mathbf{w}}$)

(ii) (Theory) Explain the behavior of the red curve as the values in the x-axis increase. Why does it increase with more training data?

(iii) (Theory) Explain the behavior of the blue curve as the values in the x-axis increase. Why does it decrease with more training data?

(iv) (Theory) Explain why the two curves are converging.

(v) (Theory/Experiment) Replace `sigma_sq = 1e-10` in the code with `sigma_sq = 1`. What did you do? Are the new results expected? Explain why.
Hint: This is a question about model selection using priors.

(vi) **(extra 1pt)** (Experiment) Describe how to use the validation data to select the best prior ratio $\sigma^2/\sigma^2_{\mathbf{w}}$. Modify the code to select a better prior ratio $\sigma^2/\sigma^2_{\mathbf{w}}$ and report the best value you found (you just need to test a few values).

Q4 (3 pts):
Using the data from the previous question download the python3 code at
`https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/hw/hw4/LR.ROC.py`

Read and run the code (using python3) and answer the following questions related to its output:

(i) (Theory) Describe the three curves. What do they represent and how are they different?
Hint: Pay attention to the validation data we are using in the code.

(ii) (Theory) Justify using the log-likelihood of the model why this result is expected. That is, explain using the theory why we know that the curve $\sigma^2/\sigma_{\mathbf{w}} = 1000$ sits between that of $\sigma^2/\sigma_{\mathbf{w}} = 100$ and $\sigma^2/\sigma_{\mathbf{w}} = 10000$.

(iii) (Theory/Experiment) Replace `validation_file = "Bank_Data_Train.csv"` with `validation_file = "Kaggle_Public_Validation.csv"` in the code. Explain why the curve for $\sigma^2/\sigma_{\mathbf{w}} = 100$ is not consistently worse than the one for $\sigma^2/\sigma_{\mathbf{w}} = 10^{-10}$ anymore.