CS57300 Homework — Homework 5 — Last update: April 25, 2016

https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/index.html Due: April 29th, 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 (5 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/hw5/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/hw5/m_dt.py`
Read and run the code (using python3) and answer the following questions related to its output:

(i) (Theory) Describe which classifier is implemented in `class MisteryClassifier`. Describe the role of parameter `nWL` (e.g. what `nWL=5000` means).

(ii) (Experiment/Theory) The output plot seems deceiving. Explain which parameter combination should be the best and why theory backs up your answer. **Hint:** Replacing `train_sizes = [300, 500,1000]` with other training data sizes such as `train_sizes = [300, 500,1000,2000,3000]` or `train_sizes = [300, 500,1000,2400,3400]` might help you decide.

(iii) (Theory) If we were to replace the decision tree classifier with a logistic regression classifier, should we use strong (zero mean, low variance) or weak (zero mean, high variance) priors for the logistic parameters (i.e., strong or weak regularization)? Why?

Q2 (5 pts):
In this question, you will learn about the CP/PARAFAC tensor decomposition, and how to use to spot interesting structures in data. Suppose that you have a traffic trace of Internet, consisting of a list of triplets

```
src-IP dst-IP port-number
```

This can be envisioned as a 3-mode tensor. For example the triplet

```
128.1.1.1 128.2.1.115 5432
```

means that the machine in IP address 128.1.1.1 got connected to the second machine, over port 5432 (which is the default port for postgres). You will use the CP/PARAFAC decomposition of such a tensor to discover (3-mode) communities, where a set of source addresses connects to a set of destination addresses, using a set of destination-ports. Such a "community" may indicate an attack: for example, some members of a botnet, may be probing, say, your company servers, looking for open ports. For your convenience, we re-numbered the IP addresses and ports, to make them integers. Thus, the dataset `https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/hw/hw5/traffic.dat` is in tab separated format, with one quadruplet per line, of the form

```
i j k value
```

where the indices start from 1, and value is always 1, meaning there was 1 packet from source i to destination j over port k.

(i) (Experiment) Run the CP/PARAFAC decomposition in `https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2016/hw/hw5/tensor.py` on the tensor you created, for rank K = 2. Provide 6 plots, one for each of the components $(\vec{a}_1, \vec{a}_2, \vec{b}_1, \vec{b}_2, \vec{c}_1, \vec{c}_2)$. That is, for say, component $\vec{a}_1$, plot the score $\vec{a}_{1,i}$, versus the index i. A high value for $\vec{a}_{1,i}$ means that source-IP i participates in the first "concept". Explain how components $\vec{a}_j$,$\vec{b}_j$, $\vec{c}_j$ relate to src-IP, dst-IP, and port numbers.

(ii) (Experiment) Consider the first component (with the three vectors $\vec{a}_1, \vec{b}_1, \vec{c}_1$). Entries with non-zero values of the rank-one component, "belong" to that 3-mode community. Give

(a) the count of src-IPs in this community

(b) the src-IPs of the community

(c) the count of dst-IPs in this community

(d) the dst-IPs of those users

(e) the ports used in this community

(iii) (Theory) Suppose we add a time mode to our tensor using the timestamp of the connections. Describe how the PARAFAC tensor decomposition can help better identify botnets (botnet attacks tend to be synchronized).