CS 247 : Advanced Data Mining Learning **(Due: 11:59 pm 04/16/19)**

Homework Assignment #1

Instructor: Yizhou Sun *TA:* Zeyu Li

Homework Policy:

- Read the ***Homework Submission Guidance*** carefully before you start working on the assignment, and before you make a submission.
- This is an individual homework. Please **DO NOT** collaborate with others.
- Write your answers in the notebook `cs247_hw1.ipynb`. Only submit a **pdf** converted from `cs247_hw1.ipynb` to Gradescope.

Problem 1: Multinomial Naïve Bayes

For multinomial naïve Bayes model, prove the MLE estimator β for is what as stated in Slide 21.

Problem 2: Iterative Optimization

When implementing gradient ascent/descent or Newton-Raphson algorithms in some programming languages, such as Python and Matlab, it is important to write down the gradient vector and Hessian matrix using matrix operation instead of using for-loop to compute each entry.

1. Write down the matrix form operation for gradient vector and Hessian matrix for logistic regression, according to Slide 35 and Slide 36;
2. Implement both versions (for-loop version and matrix version) of the gradient ascent algorithm for logistic regression. Compare the running time using the *tic-tac-toe* dataset from UCI machine learning repository. The notebook (`cs247_hw1.ipynb`) contains some skeleton code as a starter. Please implement on top of the existing code between the `TODO` comment blocks. The dataset is released together with this handout (`tic-tac-toe.data`).

Problem 3: Poisson Regression

Poisson regression is another example of generalized linear model, where $y|x, \beta \sim \text{Poisson}(f(x^T\beta))$.

1. Prove Poisson distribution belongs to exponential family;
2. Follow the recipe of GLM, write down Poisson regression model;
3. Give an application example of Poisson regression, and discuss why Poisson regression is superior to other GLMs in this case.

Problem 4: Text Classification

In this task, we are going to play with a basic dataset for text classification – the 20 Newsgroups. The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Please refer to this link for other helpful information regarding the metadata of the dataset.

Please implement Multinomial Naïve Bayes and Logistic Regression for the given text dataset, compare the performance of the two algorithms and explain why it is the case. For comparison, the metrics you can take advantage of are confusion matrix and accuracy. You can also compare on other metrics and show the difference by plots. You will find a few hints from the Ipython notebook as well.