Using Bayesian Method on
Microsoft Malware Classification Challenge

Chuan He & Ziang Zhu

I. BACKGROUND AND JUSTIFICATION

The Kaggle competition has the following link:
https://www.kaggle.com/c/malware-classification
It is a very well defined classical machine learning project.
It has a deadline of Fri, 17 Apr, right before the final presentation, and is perfect for the class project.

II. OBJECTIVES

A. Development objectives

The task is, of course, to get as high rank as possible using Bayesian method. A reasonable result would be top 20%~30%. It might requires tons of work to get higher score than that. Matlab would be used for the project. Some external library like Gaussian process library would also be expected to be used for better results.

B. Immediate objectives

The first immediate objective is to pre-process the raw data. Since the size of raw data is huge (nearly 35 GB), we might not be able to process all the data in the beginning. Our first attempt is using N-gram/Tf-idf to extract features from the raw data and try to improve the final result by applying feature selection method if we have enough time.

The plan is working on this project majorly using the Bayesian method taught in class. We will work on a simple prior distribution and linear classification first and record its result. After the result, the parameters of my models will be tuned a little bit for better result and low oversitting. This would serve as the baseline for this project.

Then we would apply more advanced technics in Bayesian method, i.e. Gaussian process, and farther technics that we will study. These results would be compared to the baseline result, and a discussion would be made for why it is better/worse than the baseline result.

III. PROJECT MONITORING AND EVALUATION

By the status report on Friday, 3 April. The project should have good result on linear classification using the some simple prior.