6.867: Project Proposal

Tony Zhang, Menghua Wu

November 17, 2016

1 Introduction

Each year, the United States Bureau of Transportation Statistics compiles comprehensive datasets regarding the nation's transportation systems, including aviation, maritime, highway, and rail. In this paper, we are particularly interested in the airline dataset, since college students often travel to and from by air. This dataset reports on a wide range of variables, a few of which include airline, carrier, origin, and destination, and flight lateness.

We have downloaded exhaustive airline data from June 2015 to May 2016, inclusive. Each month's data provides approximately 500,000 samples of individual flight data, some of which may be incomplete. We will investigate the contributions of different variables to the lateness of airline flights. We will also construct a model using a variety of methods to predict the potential lateness of a flight, given its flight information.

2 Methods

2.1 Dataset partitions

We make the assumption that month of year may affect the expected lateness of flights, so we select an equal number of samples from each month. We divide this remaining dataset into 60% training, 20% validation, and 20% testing.

2.2 Feature selection

One difficulty lies in the wide variety of data available. Some variables, such as carrier (type of plane), are inherently discrete, while others, such as departure time, can be considered continuous or discrete. For example, we can consider departure time as time since the start of the day, or partition the day into several discrete periods (e.g. early morning, late morning, early afternoon, etc.). We initially plan to consider the following features in the specified ways.

	date	time	airline	carrier	origin	destination
discrete	month	period of day	airline	carrier	origin	destination
continuous	days into year	minutes into day	—	—	—	—

Discrete variables will be encoded as one-hot vectors, and continuous variables will report their values. For single features that may be either discrete or continuous, we will take both versions as two individual features. These features will be concatenated into a single vector that represents a given flight. We will then perform feature selection using the LASSO method, in which L_1 regularization forces some feature weights to 0.

2.3 Binary classification

We will initially attempt to predict whether a given flight will be late (we consider any non-zero minutes late as a "late" flight). For this, we will compare the results of 1) logistic regression, 2) a linear SVM classifier, 3) a SVM classifier with Gaussian RBF kernel, 4) a SVM classifier with polynomial kernel, 5) a traditional neural network, and 6) random forest. We will apply L_2 regularization as necessary, to prevent overfitting, depending on the model.

2.4 Regression

Afterwards, we will proceed to predict the exact number of minutes that a flight will be delayed by. We will compare the results of 1) a linear model and 2) a traditional neural network to see how accurately we can predict lateness. We will determine the better model by mean squared 1-0 loss.

2.5 Subsequent investigations

There are several possible extensions we could make, if all else succeeds in this project. First, we may also normalize the sample numbers to account for the number of flights that occur at different times of the day. For realistic reasons, there are naturally more flights during the day than at night, especially since our dataset is primarily limited to the United States. Second, this dataset also includes reasons for delays, including weather, carrier issues, and security. These reasons might be interesting to investigate, beside variables such as time of year, carrier type, and location.

3 Risks

Due to the large amount of data, it may be hard to extract meaningful features from samples. It may also be difficult to determine the optimal representation for some variables. However, we can mitigate this risk by decreasing the number of variables we consider in the end.

This study will also be computationally intensive, so we might not be able to leverage the full size of our data as we hope. However, even a fraction of this data is substantial, and if necessary, we can restrict our analysis to certain months of the year.

4 Distribution of work

Both members of the team will work jointly and equally on writing the paper and its underlying code. We plan to spend the last quarter of our time writing, and half of the remaining time on binary classification and regression, each. For a concrete timeline, we plan to finish classification by Nov. 27 and regression by Dec. 9. Naturally, the paper will be completed by the deadline. Feature selection and dataset partitioning will be incorporated as intermediate steps.