Airline Departure Delay Prediction

Brett Naul
bnaul78@stanford

December 12, 2008

1 Introduction

As any frequent flier is no doubt aware, flight delays and cancellations are a largely inevitable part of commercial air travel. In the past ten years, only twice have more than 80% of commercial flights arrived on-time or ahead of schedule. Punctuality is an issue for all major carriers, with some struggling more than others: through September 2008, American Airlines flights were on time just 66.9% of the time, bottoming out at 58.8% on-time in the month of June.

In many cases, the causes for delays are unpredictable. For example, as of September 30th, 2008, 24.6% of flights were delayed; of these delays, 43% can be traced back to inclement weather. But in many other cases, historical data would suggest that some flights are far more likely to be delayed than others, even without taking present or future weather conditions into consideration. The airline itself is an obvious predictor of the chance that a flight is delayed; as mentioned previously, American Airlines often has one of the highest percentages of delayed flights among all carriers, while Southwest Airlines consistently beats the national averages for punctuality.

Despite the extreme complexity of flight patterns in the United States, there are many factors that do allow us to gauge the likelihood of a flight being delayed. The aim of this project is to use large historical datasets to make predictions about the punctuality of future flights far in advance, e.g. for a customer in the process of purchasing tickets for a future flight.

2 Data

Searchable flight data from the last 20+ years is available for download from the Bureau of Transportation Statistics, part of the Research and Inovation Technology Administration and the Department of Transportation[1]. This database includes basic information such as carrier, origin/destination, and expected departure, as well as detailed delay information. For the purposes of the this paper, only departure delays are studied. However, preliminary results suggest that the same analyses are valid for arrival information, as well.

2.1 Data Collection

Given that there are up to $30,000$ commercial flights per day in the United States, this is a truly massive dataset. I limited my analysis to the seven major carriers (American, Delta, Continental, Southwest, US Airways, United, and Northwest) and fifty highest-traffic airports. The same techniques can be applied to additional airports and carriers outside the very busiest, but including these greatly increases the amount of data that must be analyzed, since proportionally more total data points are needed to extract a significant sampling of less-traveled routes.

2.2 Features

The data from the BTS is very comprehensive, so I was able to include a wide variety of features in my analysis, including: Day of Week, Month of Year, Day of Year, Holiday Proximity, Airline, Origin, Destination, Scheduled Departure Time, Actual Departure Time, and Distance (only used in arrival time prediction). It may not be obvious that these features are adequate for our prediction problem; Figure 1 provides some motivation by showing that these features are indeed strongly correlated with delay.

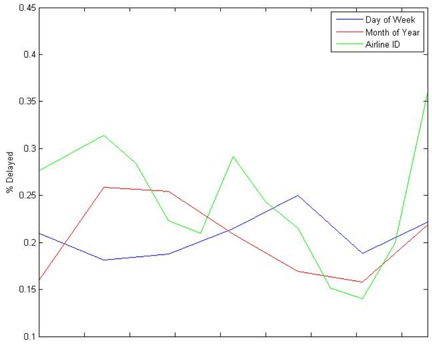

Figure 1

One type of feature that is conspicuously absent from the list above is weather data. Without a doubt, weather is the most important factor in airline delays. However, the primary motivation for this project is to be able to provide information long in advance of the travel date (e.g. when a passenger is first ticketing his or her flight, perhaps many months in advance). Weather models are reliable up to at most a week into the future, and so in most cases would not apply to the problem at hand. Long-term climate models can give seasonal information, but I claim that this information is largely redundant: in particular, the existing model already takes into account geography and season, so inserting vague climate data would only unnecessarily complicate the setup.

3 Methods

Using the data and features described above, we wish to be able to calculate some measure of the expected delay or likelihood of delay for an arbitrary future flight. There are a number of possible such measures that could be useful; this section lists the various delay estimators considered, as well as methods for computing them.

3.1 Binary Classification

The FAA officially designates flights that miss their declared departure or arrival time by more than 15 minutes as "delayed." Using this criteria (or even with a different cutoff), we can attempt to predict whether a future flight will fall into the category "on-time" or "delayed." I attempted to solve this problem using a number of classifiers, including logistic regression (both weighted and unweighted), Naive Bayes classifiers, and SVM classifiers (with linear, polynomial, and Gaussian kernels). For the first two, I used my own MATLAB routines; for the last, I used SVM^{light}[2]. Ultimately, none of these attempts produced satisfactory results, but I attribute this to the nature of the problem itself, rather than the method (see Results).

3.2 Estimated Delay

One very useful quantity to produce would be an estimate (in minutes) of the amount by which a flight will be delayed. In the Binary Classification problem, most of my models exhibited severe overfitting, so for the regression problem I attempted to use simpler methods. In particular, I computed most of my regression estimates using my own MATLAB implementation of weighted linear regression, which solves the weighted normal equations in a space-efficient way in order to compensate for the extremely large number of training examples.

3.3 Probability Estimation

An “easier,” more tractable problem than that of binary classification is estimating the probability that a flight is delayed. I experimented with multiple models for this problem, but ultimately found that the speed and simplicity of Naive Bayes best complemented the structure of the problem. In particular, I used my own MATLAB implementation of the Naive Bayes classifier, and adapted it so that features are modeled as multinomial rather than Bernoulli random variables. The features themselves are easily modeled by multinomial distributions, since they all take on bounded integer values. Because of the algorithm’s extreme simplicity, I was able to train it on full months, or even years, of flights (upwards of 10,000,000 at once).

3.4 Comparison with Historical Averages

Since the context of this project is predicting delays at purchase time, it felt natural to compare my methods with those used by many travel websites (e.g. kayak.com, orbitz.com). In particular, these sites provide consumers with delay warnings for flights that are historically prone to delay (a flight in this case is identified only by its route and flight number, so that the same flight might occur daily, weekly, etc.). Using the information on flights from the last several years, I created an algorithm that searches for past instances of a given flight, and uses the sample mean of these past flights as a sort of naive historical predictor. One of my primary goals was to improve upon this method of prediction, as it is the one most commonly used in practice.

4 Results

4.1 Binary Classification

My attempts at binary classification of future flights were unsuccessful to say the least. Figures 2 and 3 demonstrate the learning curves for test and training error for the Naive Bayes and Support Vector Machine classifiers. The Naive Bayes classifier’s error is shockingly low at first glance, but upon further examination the results are less surprising. Since the algorithm constructs probability estimates for the on-time departure of each flight, it only predicts that a flight will depart late if its estimated probability is greater than 50%. Intuitively, it is clear that very few flights would actually have a greater than 50% chance of being late; although the long-run fraction of delayed flights is as high as 25%, the variance is not high enough that many flights would be more likely to be late than on time.

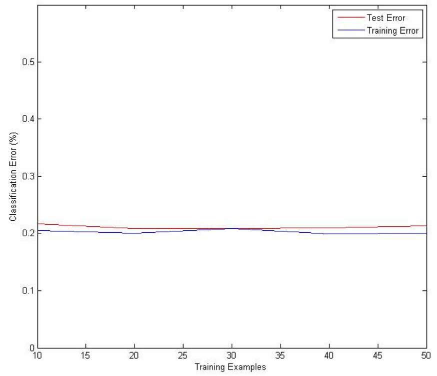

Figure 2

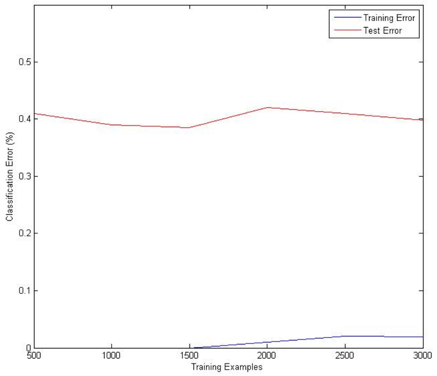

Figure 3

The Naive Bayes classifier’s outputs were consistent with the idea that flights are inherently more likely to be on-time; its classifications consisted of anywhere from 90 to 95% negative (i.e. on-time) predictions. Since up to 80% of all flights are on-time, this results in a seemingly low error rate for such a difficult problem; in reality, the error is almost the same as simply predicting that every flight will be on time. The error on the training set is nearly as high as the test error, indicating that this classifier suffers from high bias, as we would expect.

The Support Vector Machine classifier, on the other hand, performs extremely well on the training set, achieving perfect accuracy for sets of up to 1500 training examples. But these results do not translate to impressive performance on the test set; indeed, even given several thousand training examples, the test set error remains much worse than that of the Naive Bayes classifier. The data in Figure 3 comes from a Gaussian Kernel with $\gamma = .1, C = 1$, but all variations on these parameters produced similarly poor results. It is clear that the problem is one of high variance, but I saw no way to resolve this issue. The number and variety of flights in the United States is so high that any prediction algorithm must be able to digest hundreds of thousands, if not millions, of training examles; even with my relatively small feature space, the quadratic programming problem for the optimal margin classifier becomes unreasonably complicated as the number of training examples approaches this point. For the remaining problems, I tried to make use of simple, fast algorithms that would be able to handle the millions of training examples needed to describe flight patterns.

I also experimented with reducing the number of airports and carriers considered in order to give the SVM classifier another chance. Ultimately I did not pursue this strategy as it felt like I would have been cheapening the problem, or at least reducing it to a narrower and less interesting one. It also cuts down on the diversity of the already small feature space, which did not seem like a prudent strategy. Furthermore, I decided that the classification problem is not really the best one to consider; it is too much to expect our algorithm to predict exactly which flights will be on time. Instead, I attempted to develop statistics that would distinguish more fluidly between varying degrees of punctuality.

4.2 Estimated Delay

The regression problem associated with flight delays is predicting the exact number of minutes by which a flight will be delayed (recall that the FAA only classifies a flight as "delayed" if it misses its scheduled departure by at least 15 minutes). Initially, I expected this aspect of the problem to be infeasible; as with the classification problem, expecting an algorithm to give an exact delay prediction seems in some sense to be expecting too much. However, I was surprised by the success of simple regression methods. Figure 5 shows the test error of locally-weighted linear regression using Gaussian weights with bandwidth parameter $\tau = .15$. Small bandwidth parameters performed better in general; intuitively, we would like our prediction to depend primarily on very similar examples, as the relationships between parameters are highly nonlinear (e.g. Delta, is not "closer" to American than other airlines in any meaningful sense, although they are labeled 1 and 2, respectively).

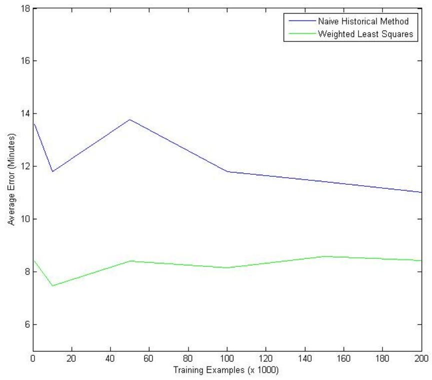

Figure 4

The predictions were far from perfect, but nevertheless improve upon the naive historical predictions generated according to section 3.4. The mean error of the least squares predictor was 8.22 minutes (generated using 200,000 training examples, after which there was no discernable improvement), compared with 10.9 minutes for the historical estimator. This result is perhaps not terribly meaningful, especially since travel sites do not present passengers with this type of information. Nevertheless, it is worth nothing that this very simple approach was able to generate surprisingly good predictions.

4.3 Probability Estimation

The most useful and plausible method of predicting of flight delay turned out to be probability estimation. These probabilities allow travel agencies to warn buyers before they commit to a delay-prone route, and similarly they allow passengers to make smarter and more confident purchasing decisions. Fortunately, this was also the aspect of the problem in which machine learning techniques enjoyed the most success. As seen in Figure 6, the multinomial Naive Bayes predictor definitively outperformed the historically-based estimates used by many online travel sites. In the classification problem, the error was seemingly quite good, but in reality the predictions were largely uniform and uninformative. In contrast, the probability estimates themselves are varied and really illuminate which flights are comparatively prone to delay.

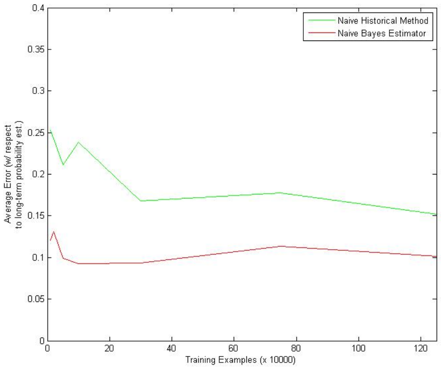

Figure 5

It is not obvious what the best way to test these probability estimates should be, and it is even less obvious how the “error” should be interpreted. Thankfully, the sheer amount of data provided one way to diagnose our results: the same combinations of features appear numerous times throughout the BTS database, so it was possible to compute one set of “past” estimates to use as the historical estimators, and another of “results” to which we could compare our predictions.

5 Conclusion

In all aspects of this problem, simple, uncomplicated algorithms outperformed more sophisticated methods. The number of training examples required to encapsulate the patterns that characterize the aviation system is truly staggering, and complicated techniques were unable to digest this quantity of information in a reasonable quantity of time. Furthermore, the simple algorithms actually performed rather impressively: in particular, the Naive Bayes probability estimates proved to be more accurate than the historical estimates used by many travel sites. The success of the Naive Bayes predictor is reason to believe that such a method, or perhaps a similar machine learning technique, could be effectively employed by travel agencies to provide improved delay estimates to its customers.

6 References

1. Bureau of Transportation Statistics. TranStats Database. http://www.transtats.bts.gov/

2. T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.