CS 229: Machine Learning Final Report
Identifying Driving Behavior from Data

Robert F. Karol
Project Suggester: Danny Goodman from MetroMile

December 13th 2013

Problem Description

For my project, I am looking into how the driving behaviors influence the fuel efficiency of different vehicles. Danny Goodman, with MetroMile, has provided me with a set of data consisting of accelerometer, heading, speed, and gas usage information which I am using in order to determine how efficient a given driving style is. The goal of this project is to identify techniques which people could use in order to save fuel which has become a significant cost in the lives of many Americans.

Data Parsing

Initial work needed to be done in order to parse the data into a usable format in MATLAB. Since each set of data came in a different file, each file had to be parsed, imported into MATLAB, and manipulated into an easy to access form. In particular, each file contains months worth of trips, and each trip needs to be separated out in order to determine when a car stopped and restarted. In order to do this I looked through the timestamps of data points, and whenever there was a gap greater than 6 seconds between data entries which should be occurring at a rate of 1 per second, I recorded it as a new trip. Then, any trip which lasted less than a minute was removed as short trips could introduce too much variability into the gas consumption since gas consumption is given as an instantaneous quantity so by coasting with the car an artificially high efficiency can be created.

All of the data came from small units which have been attached to the on-board car computer which is where measurements of gas consumption come from. Since every car is different the measurements coming from each vehicle have slightly different noise characteristics. The unit has a built in accelerometer and magnetometer, so while the data might be noisy, or be off by a scale factor with bias, the data from each vehicle should be consistent. In total, there is data from 18 different cars which collected data for months on a number of different trips. The data collected does have a few major problems which need to be addressed.

One of the biggest problems with the provided data is that the accelerometer is misaligned in ever vehicle. This means that the data must be transformed from the reference frame of the accelerometer to the reference frame of the car before it can be used. Additionally, the data was collected at a rate of 1 Hz. While this is not a problem for heading data, it makes the accelerometer data extremely noisy. This is a problem which might skew the final values calculated for some quantities of interest, however, it shouldn't be a problem for spotting the important fuel consumption trends. A second problem for which there is no correction is the presence of multiple drivers in a single vehicle. While we are able to separate out the data obtained from each vehicle, there is no information available who is driving, or what the road conditions happen to be. This could become quite a serious issue if different fuels are being used. One example would be the different energy densities contained in the summer and winter

fuel mixtures that exist in California.

I wrote a series of MATLAB scripts which import the data and put it in a more useful form. First, I wrote a script which parsed through the csv file, and saved all of the data into a .mat file as given in the dataset. This way, the data is much more accessible since it doesn't have to be imported for every use. A second script was written in order to take the instantaneous data, and integrate it over time. This way I was able to get estimates of the total gas consumption and distance traveled for each trip the car took. Using this data I was able to work out the average fuel consumption for a given trip without as much error as is achieved by averaging the raw data. A second function was written capable of plotting the path the car by combining information gained from both the heading information, and instantaneous speed. This helps to determine how the car was moving for a given trip adding to the intuition of what is going on for a given trip.

The most difficult data parsing task has been working with the accelerometer data. This is some of the most important data collected since acceleration is a huge part of fuel consumption, however, the misalignment problem is very difficult to deal with. In order to determine the correct orientation for the car I determined the times when the car was stopped, and used the average accelerometer values to determine what the gravity vector was. Using this information I was able to reconstruct the rotation matrices necessary in roll and pitch to move the gravitational vector to point straight down. While this orients the gravity vector and determines 2 directional components, it is unable to determine which direction the car is facing relative to the accelerometer. In other words, we are unable to determine yaw using this information. In order to determine a value for yaw, I had to use my knowledge of what the x and y accelerations should be. By finding the parts of the data where the car was moving relatively straight, and using the fact that accelerations should be located in the x and z directions where positive z points down, x out the front of the vehicle, and y out the right side. By choosing the yaw value so that the acceleration is concentrated in the x direction while the heading is not changing, I was able to estimate the overall vehicle orientation.

Sample Trip

Figure 1 shows some sample data from a typical trip. The speed, distance traveled, gas consumed, and an estimated path along which the car traveled. This gives a decent idea as to how useful the data might be, as well as which features might be worth looking into in order to save gas. Over the course of my projects I plotted a number of other features, some of which I thought were highly relevant to fuel efficiency, others which should have no affect.

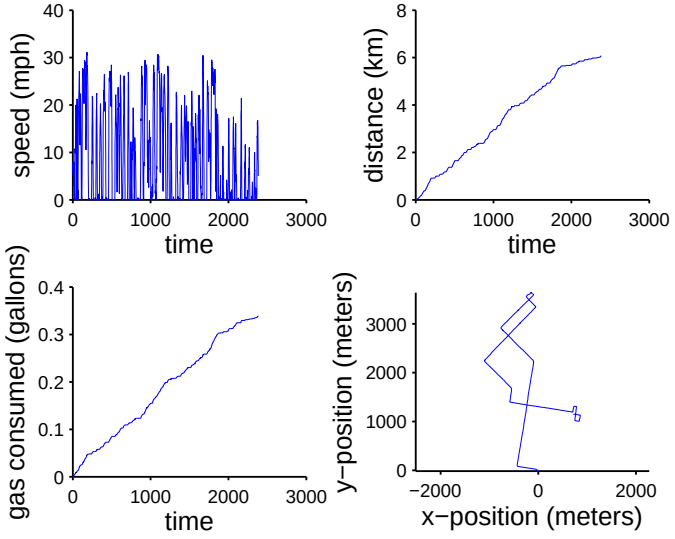

Figure 1: Example of a typical trip.

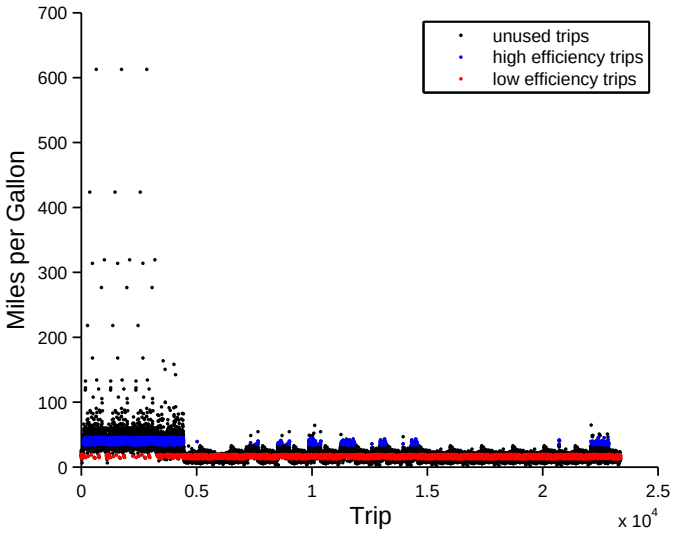

Figure 2: Dividing trips by efficiency.

Defining Efficiency

The first step in identifying features of high or low efficiency driving is to determine what constitute a high or low efficiency trip. First, I defined the efficiency of a trip by the total distance traveled in a given trip divided by the amount of gas consumed. In order to do this I used k-means clustering to find two separate natural groupings. These two groups nicely divide the trips into lower and higher efficiency categories. By using k-means rather than sorting the trip efficiencies and dividing it at the median, or using a parameter like half the maximum efficincy, we can get more natural groupings which are less biased by outliers. Additionally, there is no reason for a fixed percentage of the trips to be considered efficient.

Due to the nature of the data collection, some trips may have been cut off after a long period of coasting, or slowing down. Another possibility is the trip only includes time when the car was accelerating if the data dropped out before cruising or slowing down. As can be seen in figure 2 there are certain trials resulting in unrealistic efficiencies. In particular those with efficiencies of hundreds of miles per gallon, or near 0 are probably not the trips to focus on. In order to remove some of these anomolies, only the second and third quartiles of data are used removing the outliers in each of the two categories. In figure 2 the red points indicate trips categorized as high efficiency trips while the blue points indicate low efficiency trips. The remaining black points represent trips considered outliers in either of the two categories.

Features

Since the goal of this project is to determine relevant features which a user can keep track of in order to help maximize the fuel efficiency, I used the data provided to come up with 29 potential features which a driver could keep track of with relative ease if these features ended up being relevant to gas consumption. Using the data provided, I came up with a list of 29 features to look into for determining their relative importance. The following features were chosen for this analysis.

In particular I looked at the maximum and mean values of various quantities as these would give a decent estimate of how the magnitude of these quantities directly affects fuel efficiency. By looking at the standard deviation of various quantities, I am able to get an idea of the affect that more stable quantities have as opposed to changing quantities. Finally by looking for peas in the Fourier transformation of a variable I can try to see if any natural cycling occurs in a variable and determine how important such cycles could be.

- Total Acceleration ($\sqrt{\text{x-acceleration}^2 + \text{y-acceleration}^2 + \text{z-acceleration}^2}$)
  - maximum, mean, standard deviation, peak in Fourier transform
- Planar Acceleration ($\sqrt{\text{x-acceleration}^2 + \text{y-acceleration}^2}$)
  - maximum, mean, standard deviation
- x, y, and z accelerations
  - maximum, mean, standard deviation, peak in Fourier transform
- Speed
  - maximum, mean, standard deviation
- Heading
  - maximum, mean
- Total distance.
  - maximum
- Gas used
  - maximum, mean, standard deviation, peak in Fourier transform

PCA

One of the first tests I ran was a principal component analysis, in order to determine what combinations of features contributed to the largest variation across trials. While a linear combination of features may not be the best for giving drivers a concise and easy to remember set of instructions for reducing their fuel consumption, I thought that it could provide some useful insights into the data itself, as well as the particular features I choose to use. By examining the results of the principal component analysis, the following results appear.

88 % of the variation in by the data is primarily due to the maximum and mean of the heading information. While this makes intuitive sense as heading is completely independent of all fuel efficiency since the same trip can be done with any orientation with respect to north.

By including the maximum and mean speed, as well as the total distance. 98% of all variation is explained. Once again though, total distance traveled is completely irrelevant towards overall efficiency.

From these results I determined that while there might be better separation along a reoriented coordinate system, in order to perform feature selection it would be better to choose between the initial features rather than a linear combination to maximize variance.

Feature Detection

The most important part of this project is the feature selection algorithm to determine which features contribute the most to predicting high or low efficiency driving. In order to accomplish this, I decided to use a linear kernel for a support vector machine in order to determine which features, when removed. Would affect how the data was separated the least. I developed a cross validation scheme which would take a matrix containing all the trials, train a support vector machine using a linear kernel on 90% of the data, and test it against the remaining 10% of the data to determine the number of classification errors. This is repeated 10 times, each time holding out a different 10%.

Using the cross validation scheme described above, I used the following feature selection algorithm. First, I ran the cross validation scheme on the matrix of trials and features 29 times. Once each using the matrix of trials and features having removed a single feature. Be comparing the cross validation error in each of these cases and selecting the one with minimal error, we could find which feature was least important in data separation since the support vector machine was capable of perfectly separating the data using all 29 features. This procedure was run repeatedly to determine the next least important feature which could be removed, resulting in a ranked list of least important to most important features for separating the data.

In the end this method indicated the following ranking for the most important features which would contribute the most to determining high and low efficiency driving.

1. Average speed
2. Average gas used
3. Standard deviation of gas used.
4. Standard deviation of the planar acceleration
5. Average of the planar acceleration
6. Standard deviation of the y-acceleration
7. Maximum of the planar acceleration

While this feature selection algorithm is not perfect, it does give us a good idea of what driving techniques might cause higher gas usage. What this method does not do however, is distinguish which of these features come from driving habits, and which come from natural road features. For example, highway driving vs stop and go traffic are very different road conditions which are completely out of the control of the driver.

SVM

I have included a plot from the final run of the support vector machine cross validation to see how well separated the data is when only using the two most important features identified by the feature selection algorithm.

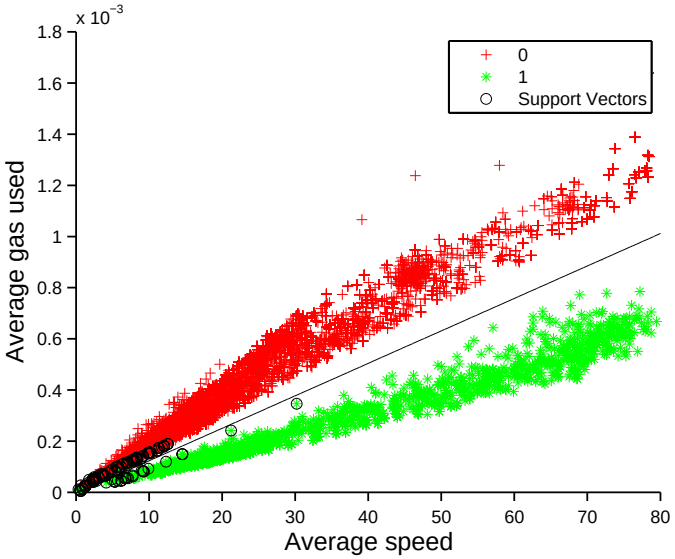

Figure 3: Separation using only the 2 most important features.

Using these two features alone, prediction errors are significantly lower than 1%.

Conclusion

In the end, the goal of this algorithm was to determine a list of features, all easy to compute in real time, which do a good job at predicting how efficient a driving experience will be. However, more importantly than that, this algorithm has determined some ways that drivers might be able to reduce their fuel consumption. Reducing average speed, the mean, and standard deviations of the planar acceleration are two of the most significant which the driver has control over. In particular a driver can save fuel by reducing the their average speed. While this is not practical in all situations, it suggests that following the speed limit rather than exceeding it by 10-15 miles per hour, is not only safer, but could help you save on gas. Additionally, in situations with stop and go traffic it is best to accelerate slowly so the average speed is reduced, and smoothly so the standard deviation of acceleration is reduced. This makes intuitive sense which gives some credibility to the algorithm itself. While this provided for a good method of predicting efficiency as well, there is less use in that as an individual driver may not have control over the variables at all times.