Predicting gas usage as a function of driving behavior

Saurabh Suryavanshi, Manikanta Kotaru

Abstract

The driving behavior, road, and traffic conditions greatly affect the gasoline consumption of an automobile. We believe that accurate prediction of gasoline usage as a function of the above can be used to obtain optimal driving behavior, for given road and traffic conditions, to reduce gasoline consumption. This results in great dollar savings and also reduces the harmful effects of gasoline usage on environment. In this project, we propose machine learning framework to predict gasoline usage as a function of driving behavior with high accuracy. Analysis of the results shows that we are able to predict the rate of gasoline consumption (Miles/hour) much better than fuel economy (Miles/gallon), as a function of driving behavior. We then propose an application by constructing a machine learning framework to minimize the total gasoline consumption over a period of time.

1 Introduction

It is known that driving behavior and traffic conditions play a significant role in the fuel consumption of automobiles [1]. For example, traffic congestion results in high gasoline usage due to frequent gear shifts, braking, and accelerating. Also, on highways, by traveling at different speeds (below the speed limits), the vehicles consume different levels of gasoline. If we are able to accurately predict the values of gasoline usage as a function of driving behavior and road conditions, we would be able to suggest a driving profile to minimize the gasoline consumption along a given path. If we are given different paths from source to destination, we may choose the path that consumes least amount of gasoline (as opposed to time considerations alone in $Google^{TM}$ maps).

Previously, various regression models have been proposed for modeling the vehicle fuel consumption [1]. The key input variables for these regression models are instantaneous speed and magnitude of acceleration/deceleration. For example, [1] models the fuel consumption of light-duty vehicles as the interaction between linear, quadratic, and cubic speed and acceleration terms. However, the experiments in [1] were conducted in a controlled setting, where the fuel consumption is measured for different instantaneous speed and acceleration values (within the operating limits of vehicle). In section 2, we explore the possibility of using velocity and acceleration as input variables to predict fuel economy for a set of data samples obtained in a normal driving setting. In section 3, we introduce a new target variable and quantitatively demonstrate the improvement in the prediction accuracy. In section 4, we apply Support Vector Regression (SVR) to the data to obtain very good prediction accuracy. In section 5, we propose an application and obtain a solution using Reinforcement learning (RL) framework.

1.1 Data

This project is inspired from the *Identify $ 100 Billion in Yearly Gas Savings from Driver Behavior* project floated on CS 229 project suggestions. The dataset provided by MetroMile, Inc includes high frequency accelerometer readings (normalized with respect to g, acceleration due to gravity), speed (in miles per hour), heading angle (in degrees from north), and fuel economy (in miles per gallon) measurements of seventeen different automobiles. Each dataset contains varying number of data instances, varying from 40,000 to 11,00,000 instances. The dataset for each automobile consists of numerous trips of different time lengths. The sensor measurements are taken each second for the entire duration of each trip.

Prediction of gasoline usage

In section 1, we argued that the gasoline consumption of an automobile is a function of driving behavior and road conditions. The driving behavior is characterized by the speed and acceleration profile of automobile, which we have access to. However, we do not have access to the road conditions from the given data. So, we make a reasonable assumption that the road conditions are implicitly incorporated into driver behavior (see figure 1). For example, traffic congestion (which is a component of road conditions) results in low average speed and frequent change in acceleration (which are components of driver behavior). Hence, the input variables into our regression algorithms, to predict gasoline usage, are speed and acceleration components.

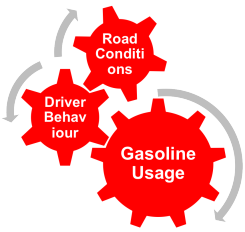

Figure 1: Driving behavior implicitly accounts for road conditions

2 Simple regression models

For each automobile data, linear regression is performed and one-fold cross validation error is obtained using randomly selected 30 percent of the data as hold-out cross validation set. The Correlation Coefficient (CC) between actual and predicted values for six different automobile datasets are presented in the *unpolished data* section of figure 2. We can observe that the CC values are not satisfactory with CC value being close to 0.5.

2.1 Polishing of the data

The accelerometer readings contain instantaneous acceleration experienced by the accelerometer device, which includes the effect of acceleration due to gravity. We assume that over a large number of observations, the average accelerometer reading will point towards the earth's gravitation. The affect of acceleration due to gravity can now be removed by subtracting the gravity value from all the measurements.

Following [1], we add additional features like linear, quadratic, and, cubic speed and acceleration terms. The CC values for the data after polishing and adding the additional features are presented in the *Additional features* section of figure 2. We can observe that not much improvement in the performance is obtained by adding additional features. Also, locally weighted regression is performed and the hold-out cross validation errors are obtained by using randomly selected 5 percent of the data as hold-out cross validation set. The Root Mean Square (RMS) values obtained are smaller than but comparable to those obtained using simple linear regression, and are presented in figure 3.

2.2 Analysis

The RMS values and CC values obtained for different datasets are not satisfactory. The failure of locally weighted regression to predict the gasoline usage as a function of input variables indicates that we are not able to estimate the value of the target variable as a locally linear function of features in nearby space. There are two approaches to obtain better algorithms. First, we can create a higher dimensional feature vector from the available feature vector (like we did in SVM). We then hope that, in this high dimensional space, the value of the target variable at a point can be obtained, with reasonable accuracy, as a linear (or locally linear) function of input features. Second, may be, the features considered are not able to capture the gasoline usage. Either we need a different/extended feature set or change the target variable. We observe that a combination of the two approaches provides the best results for our problem.

Key Insight: The rate at which gasoline is consumed is proportional to the energy spent. The energy from combustion of gasoline is used to increase the kinetic energy of the automobile (depends on the difference of squares of successive speeds), and to work against the friction forces like air drag (depends on speed) and friction against ground. Hence, it may be appropriate to model the rate of gasoline consumption as a function of speed and acceleration rather than trying to model fuel economy (miles per gallon measurements) of the automobile. Hence, we designed a new target variable $gasoline\ rate = \frac{1}{fuel\ economy} * speed$, measured in gallons per hour.

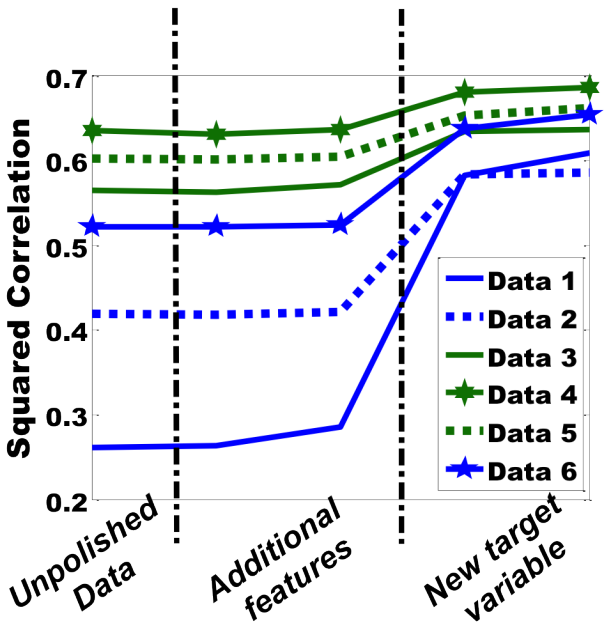

Figure 2: Correlation coefficient for different experiments for different automobile Datasets. Different curves correspond to different automobile datasets. The dataset is divided into 3 sections corresponding to different experiments.

3 New target variable

For each automobile data, linear regression is performed and results are obtained as described in section 2, and are presented in the *New Target Variable* section of figure 2. We cannot use RMS values to compare the performance of the algorithm for the two targets because the ranges of gasoline rate and mileage are widely different. Hence, we used correlation coefficient (CC) values to compare the improvement in prediction accuracy. We observe significant improvement in the prediction accuracy (taking CC between predicted and actual measurements of target variables as measure of accuracy) when we change the target variable to gasoline rate.

Hence, we can conclude that we can *follow* gasoline rate much better than fuel economy as a function of speed and acceleration. Also, adding difference between current speed and previous speed value as an additional feature slightly improved the CC values. We believe that this is because the increase in the kinetic energy (which effects the gasoline rate) depends on the difference in speeds. This result further supports our argument to change the target variable to gasoline rate.

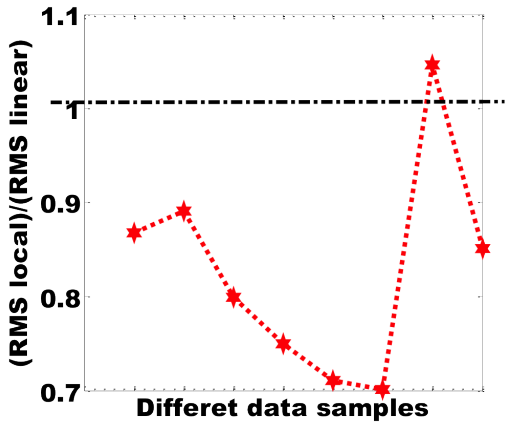

Figure 3: Comparision of RMS errors for linear and locally weighted regression. The black line corresponds to RMS error for linear regression. The RMS errors are normalized with respect to the RMS error of linear regression for the corresponding dataset.

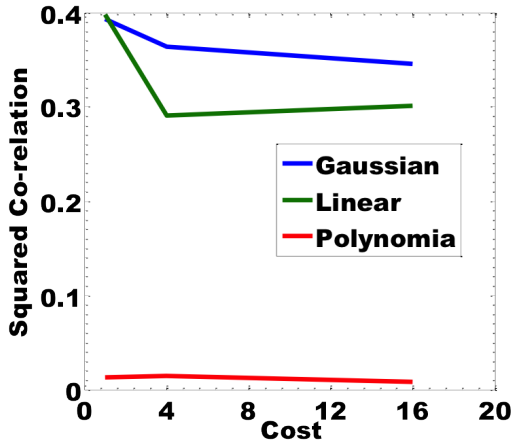

Figure 4: Correlation coefficient for different kernels

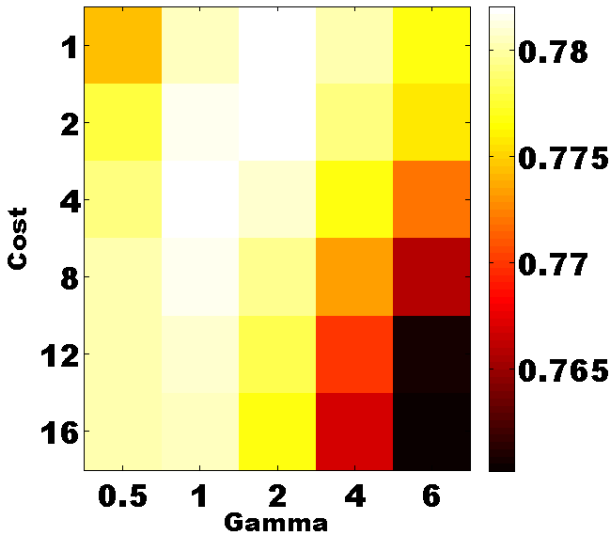

Figure 5: Grid search to get optimal SVR parameters. Cost is C and Gamma is γ

4 Support Vector Regression

In this section, we describe the implementation of the other approach, to map the feature vectors into higher dimensions with gasoline rate as the target variable.

4.1 Experiments

We used the *libsvm* software package [2] to implement ϵ-SVR, with 10,000 randomly selected samples for each automobile as training data. We then obtain the SVR parameters by performing ten-fold cross-validation. SVR performed better than linear regression over all the datasets (see Figure 6).

Choice of kernel: It can be observed from figure 4 that Gaussian kernel performs better than linear or polynomial kernels. We believe that Gaussian kernel performs better because it introduces larger number of features. Also, scaling of the data, so that each of the feature vectors lies in the same range, improved the performance of the algorithm for all the kernels. Scaling of the feature attributes avoids attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation [3].

Choice of parameters: We choose Gaussian kernel for SVR. It has two parameters, C and γ (Inner product between x_i and x_j, for Gaussian kernel is, $exp(-\gamma||x_i - x_j||^2)$ and C is the cost parameter similar to cost C in SVM). We initially perform a grid search to obtain a smaller range for optimal C and γ values. We then perform a finer grid search (see figure 5) in this smaller range, to obtain optimal C and γ values as 4 and 1 respectively.

Results: With this choice of optimal C, γ, and Gaussian kernel, we obtain very good value of 0.8 for correlation coefficient between predicted and actual gasoline rate. Figure 6 presents a comparision of CC values for linear regression and SVR for different automobiles.

5 An application

We explore the following problem in this section "Can we come up with an optimal velocity and acceleration profile to minimize the total gasoline consumption over a given finite time period?" We can take a greedy approach to solve the above problem. Given the present velocity and acceleration of the vehicle, velocity of vehicle in the next time instant is automatically determined using equations of motion. We then greedily choose the next acceleration as the acceleration that minimizes the gasoline consumption for the fixed next velocity. We then proceed greedily in this fashion.

Proceeding in greedy manner may result in a non-optimal solution. Moreover, gasoline usage is not a simple function of velocity and acceleration. Hence, in order to find the optimal acceleration for a given velocity, we may resort to gradient descent schemes, which again is non-optimal. It is also necessary to use numerical approximations to obtain partial derivative, with respect to acceleration components, of the func-

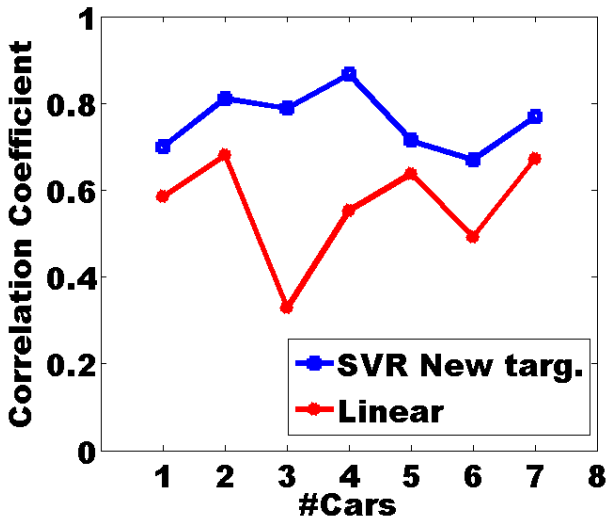

Figure 6: Correlation coefficients for different Datasets. SVR New targ referes to SVR. Linear refers to linear regression. The target variable is gasoline rate.

tion that describes gasoline rate as a function of input variables. We propose an alternative and very natural reinforcement learning framework to solve this problem in a *non-greedy fashion*.

5.1 RL framework

We divide the continuous space of speed and acceleration into a finite number of discrete states. For this setting, we assume gasoline usage is a function of speed and acceleration alone (in line with assumptions in [1]). Reward in each state is negative of gasoline rate. The action space is {accelerate, decelerate, neither accelerate nor decelerate}. The movement of the automobile from one state to another can now be seen as MDP (see figure 7). The uncertainty comes both due to the choice of exact value of next acceleration and due to the discretization of space. We note that the next speed is fixed by the present acceleration. The data collected can now be seen as a series of experiments where the driver takes random action in the present state, moves to the new state, and observes the reward in the new state (see figure 7). As mentioned in section 1.1, we the data for each car consists of multiple trips. We can use these trips as experiments to create and update a model of MDP.

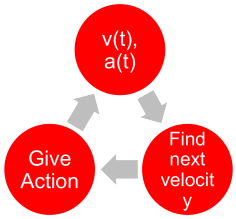

Figure 7: RL framework showing transition from one state to next

5.2 Problem formulation

The problem we are trying to solve can be posed as

$$\{a_i\} = \arg\min_{a_i} \sum_{i=1}^{N} g(a_i, v_i) * T \tag{1}$$

Here, $\{a_i\}$ is the set of optimal accelerations (decisions we need to make). We make the decision at every time-instant. N is the total time-period. g is a function which describes the function which takes speed and acceleration as inputs and gives the gasoline rate as output. T is the length of each time period.

5.3 Obtaining the MDP model

Gasoline rate is nothing but gasoline usage (measured in gallons) in one second. Hence the cumulative gasoline usage of an automobile over a time period is nothing but the sum of gasoline rates over the time period (assuming that the gasoline rate is constant over each time period which here is one second). In reinforcement learning framework, we minimize the expected payoff, which in this case is total gasoline consumption with gasoline consumption in time period t is discounted by a factor γ^t. We can use this function of total expected payoff as an approximation to the gasoline usage over a finite time period. Hence, the modified problem is

$$\{a_i\} = \arg\min_{a_i} E[\sum_{i=1}^{\infty} \gamma^i g(a_i, v_i)] \tag{2}$$

We may argue that this is a reasonable proxy for our original optimization problem if $\gamma < 1$ and positive. This is because the gasoline consumption at far away time instants is discounted by a very large factor.

5.4 Results

We learn the model of MDP and obtain the values, which here are negative of the total expected gasoline consumption starting in the particular state, using value iteration. The reward for each state obtained after applying the reinforcement learning framework is presented in figure 8. It can be observed that lower speeds are associated with lower gasoline consumption. Also, acceleration is associated with higher gasoline consumption (lower reward) than corresponding deceleration of same magnitude. This observation complies with similar observation made in [1].

We argued that reinforcement learning framework provides a better acceleration profile that achieves lower gasoline usage than the greedy approach. To support our argument, we present the cumulative gasoline usage, the value of the objective function in equation 1, achieved by greedy approach and RL approach in figure 9. In RL approach, if action is to accelerate, we choose a random acceleration in some range depending on the distribution of acceleration values observed in data. Similarly, we decide the exact value of deceleration if the action is to decelerate. We observe that

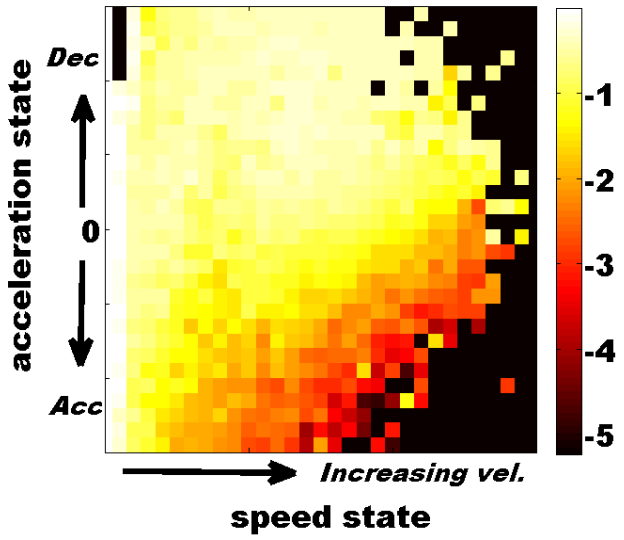

Figure 8: Reward values for different states. Vel. refers to speed. Acc refers to acceleration with increasing magnitude. Dec refers to deceleration with increasing magnitude. Black boxes represent states which are never reached.

although greedy approach may result in short-term gains in gasoline consumption, gasoline consumption achieved by RL approach is smaller than the greedy approach over a longer duration.

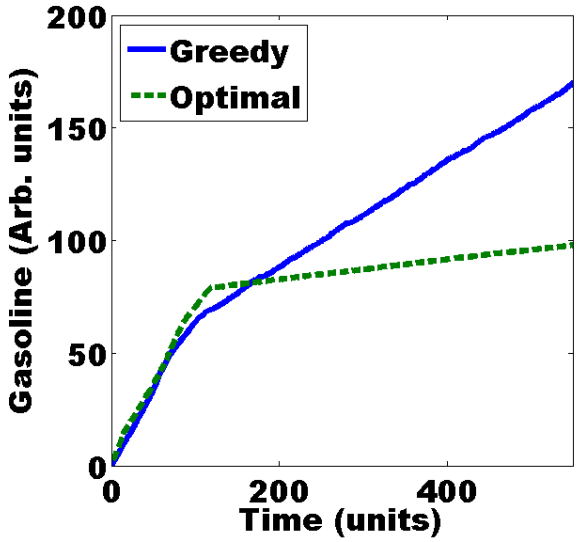

Figure 9: Cumulative gasoline consumption. The experiment is performed for time period 10 minutes. The gasoline consumption is scaled by 3600 on y-axis.

6 Conclusions

We have shown that changing the target variable from fuel economy to gasoline rate improves prediction accuracy both for linear and SVR methods. We have shown that using SVR on scaled training data provides very good prediction accuracy of around 0.8 measured in terms of correlation coefficient between predicted and actual gasoline rate values. We have also developed a reinforcement learning framework to solve the problem of minimizing cumulative gasoline consumption over a given time period.

Future work: A set of controlled measurements, where we have some measurements where we know that car is moving in a plane perpendicular to gravity direction, would help us to come up with better approaches for data cleaning. Especially, it helps in taking into account directly the effects of road conditions on driving behavior. Specifically, clean data will help us to understand the transformation from accelerometer frame to car frame to ground frame (see figure 10). This is essential in city like San Francisco to understand the terrain (like slopes) and include effect of gravity. For example, when moving down a slope, even though we are travelling at a high speed, gasoline consumption may be low because gravity is helping us 'push' forward. Hence, by working in ground frame, we can take into account gravity and come up with better prediction of gasoline consumption.

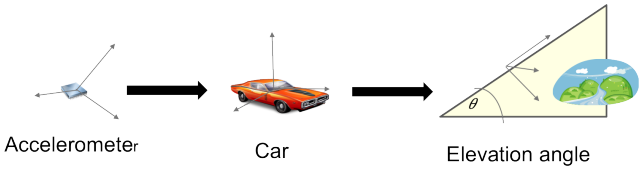

Figure 10: Transformation of axes to ground frame

A more interesting application would be to minimize the gasoline usage over a particular path (which may be a path a user commutes daily), or at least for a particular distance travelled. Problem formulation for the later problem is

$$\{a_i\} = \arg\min_{a_i} \sum_{i=1}^{\infty} g(a_i, v_i) * T$$

$$s.t. \quad \sum_{i=1}^{\infty} v_i * T = D$$

Here, D is the distance to be covered. We have not yet come up with tractable solutions or approximate solutions for this problem. But, we believe that it is an interesting direction to proceed.

References

[1] Ahn K., Rakha H., Trani A., and Van Aerde M., *Estimating Vehicle Fuel Consumption and Emissions Based on Instantaneous Speed and Acceleration Levels*, Journal of Transportation Engineering, Vol. 128, No. 2, March/April 2002, pp. 182-190

[2] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[3] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, *A Practical Guide to Support Vector Classification*, http://www.csie.ntu.edu.tw/ cjlin