Stanford CS229 Machine Learning Final Project

Prediction of Airline Ticket Price

Team Member: Ruixuan Ren, Yunzhe Yang, Shenli Yuan

Mentor: Bryan McCann

Motivation

As International students, we inevitably need to travel frequently and have to deal with all the expenses associated with it, among which airfare is one of the most significant expenses. Therefore, we become really interested in a model that is able to predict the airfare.

Data Source

The data, provided by Professor Maria Gini [1], were originally collected using daily price quotes from a major travel search web site over the period February 22, 2011 to June 23, 2011.

Data Features

- Departure week begin
- Weekday of departure
- Price quote date
- Weekday of the price quote
- #days between quote and departure
- Number of stops in the itinerary

Models

Linear regression

Both unweighted and weighted linear regressions were attempted. For weighted linear regression, three bandwidth values (0.8, 2, 10) were used for comparison.

Models

Naïve Bayes

Multinomial event model of Naïve Bayes with Laplace smoothing was applied. The target variable is discretized relative price (each price over overall minimum price). We used equal interval for discretization.

Softmax regression

Softmax regression was applied with the same method of discretization as that used in Naïve Bayes.

Support Vector Machine (SVM)

SVM was also used with the same discretization method, producing a similar value of accuracy. However, when the data is discretized into more than two bins, the error increases significantly.

Training Error

Model		Error
Linear regression	Unweighted	0.3733
	$\tau = 0.8$	0.3989
	$\tau = 2$	0.3774
	$\tau = 10$	0.3733
Naïve Bayes		0.2694
Softmax regression		0.2316
SVM (two bins)		0.1939

SVM regression and Logistic regression have also been used; the results, however, are not satisfying enough and therefore discarded.

Diagnostics

Learning curve was utilized to investigate the price prediction problem. We picked Naïve Bayes and SVM to investigate further.

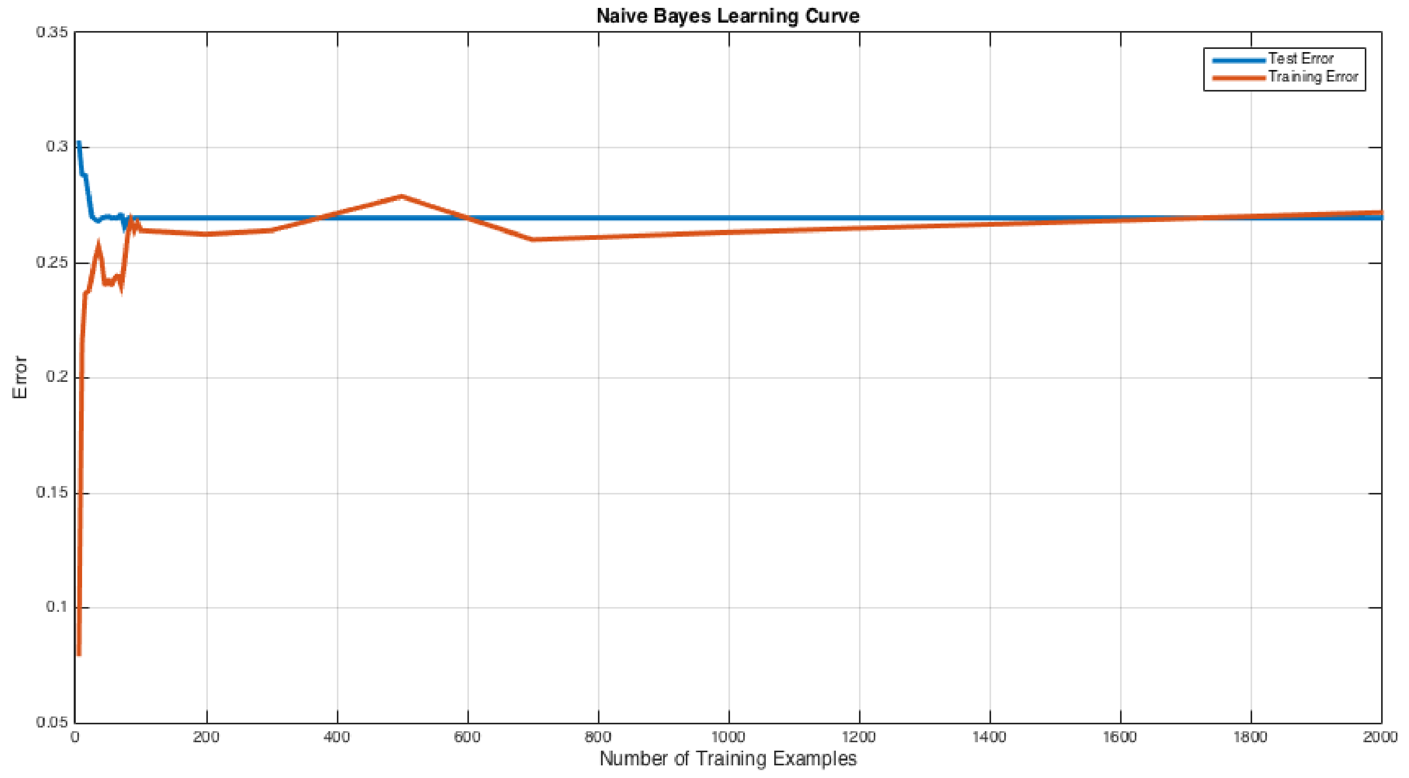

Figure 1: Naive Bayes Learning Curve

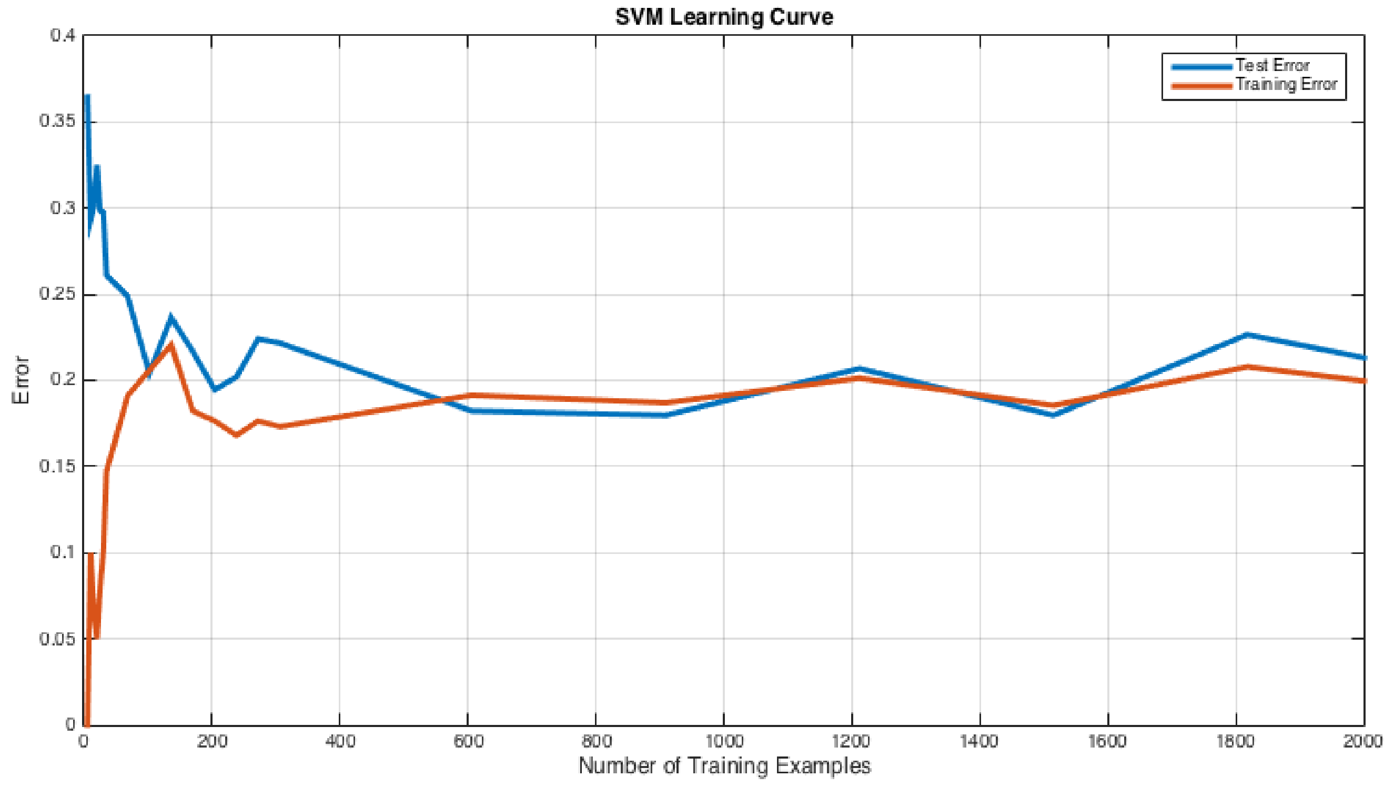

Figure 2: SVM Learning Curve

From the plots above, it is obvious that for both model, we have a high bias. Therefore, we tried adding features to our models, which resulted in smaller training errors. In the future, other features, such as the available seat and departure time of a day, need to be considered.

Reference

[1] A Regression Model For Predicting Optimal Purchase Timing For Airline Tickets, Groves and Gini, 2011