Where's my Ferry?

Support Vector Machines for Modeling Ferry Tardiness

March 23, 2013

Contents

1 Introduction

Ferries (as in **Figure 1** present an interesting opportunity to examine a complex traffic system filled with data and affecting the lives of many individuals. This paper will discuss a new approach to predicting the timeliness of ferries using a support vector machine. An overview of how support vector machines work and their complications in real world use will introduce the discussion of how this powerful construct presents a powerful and intuitive model for tackling the massive data associated with ferries. A functioning model is developed to examine the complexities of this approach, the practicality, and as a means of exploring some of the data.

Figure 1: A prototypical WSDOT ferry on its route.

We will attempt to predict the tardiness of ferries, where tardy is defined as no less than three minutes later than the initial (at day's beginning) scheduled arrival or departure time. This definition of tardiness was chosen to allow for a non-trivial number of tardy events (more than 10restroom in three minutes time or otherwise make use of that information. The Washington State Department of Transportation's ferry system in the Pacific Northwest served as the source for all data. As will be discussed later on, this ferry system is notoriously on time with over 95% of trips being on schedule (within one minute). This system is of particular interest for its sheer volume of passengers: 22 million per year. The volume lends itself well to this project's goals of using data mining.

2 Preliminaries

Since the goal of the project was to use a large data set for machine learning and to develop a model for possibly predict future ferry tardiness, two approaches popular in

available implementations and the literature were initially evaluated: artificial neural networks (ANN) and support vector machines (SVM). These two approaches fit into the category of **supervised learning**: labeled (known) data is fed into an algorithm that learns to make better predictions through comparing its own predictions to the labeled data set. ANN and SVM are often considered to be very similar in the problems they attempt to tackle, especially since each can be used as a linear classifier: a function taking in some object and determining a class to which it belongs. Rather than considering tardiness as a spectrum of degrees, we approach tardiness as a binary feature: late or not. This conceptually simplifies our problem. If artificial neural networks and support vector machines can both solve classification problems, however, which are we supposed to choose?

Byvatov et al examine the differences of the two approaches for classifying drugs [1]. While their model certainly isn't for a ferry system, their discussion suggests data sets with many features (an aspect of the ferry model we discuss later) are better suited for SVM, and they also find little difference in the predictions of ANN and SVM. Most importantly, their conclusion, and that of their references, is the two approaches are complementary. Each tool catches different cases, but the results are comparable in each approach, and SVM may even require less "tweaking".

Support vector machines are also more intuitively motivated. Consider mapping the features of a ferry trip (time of departure, weather, boat name, etc.) into a coordinate plane. Now, if we do this for all the points in our training set, we can also attach labels to the points (late or not, as in **Figure 2**). In two dimensions, we could try drawing a line through the late and not-late points to separate them. This line is simply a hyperplane in higher dimensions, and SVM has its theoretical underpinnings based on the idea of drawing this hyperplane optimally. In **Figure 3**, we show a sample training of a simple support vector machine.

SVM are used as the tool for finding a linear classifier in this project mostly for ease of understanding and the possibility for use in regression analysis (a topic discussed in [2]).

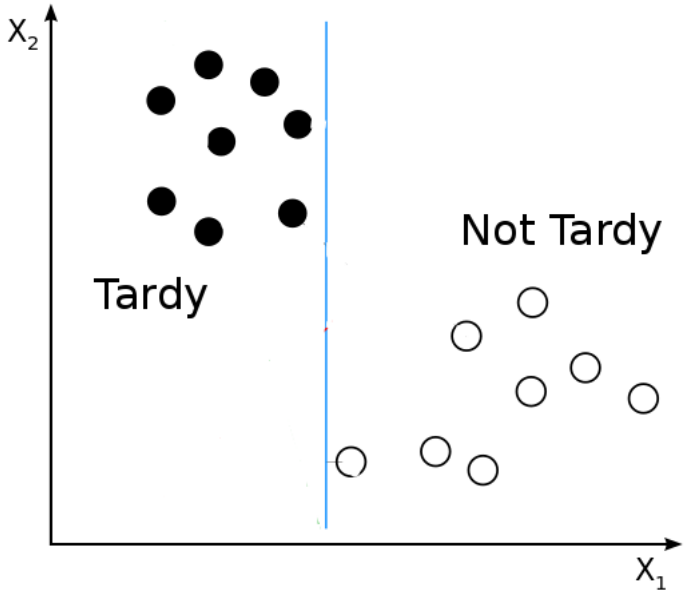

Figure 2: A simple example of a separating line for tardiness (i.e. hyperplane). Each data point (or instance) has features X_1 and X_2.

2.1 SVMs for tardiness

2.2 Linear classifiers

2.2.1 Feature vectors

2.2.2 Training

2.2.3 Kernels

2.3 LibSVM

3 The problem with ferries

3.1 Washington State's ferries

3.2 Where the data comes from

3.3 The first pass

3.4 General results

3.5 The problem with weather

4 Closing thoughts

4.1 Remaining questions

4.2 Extensions

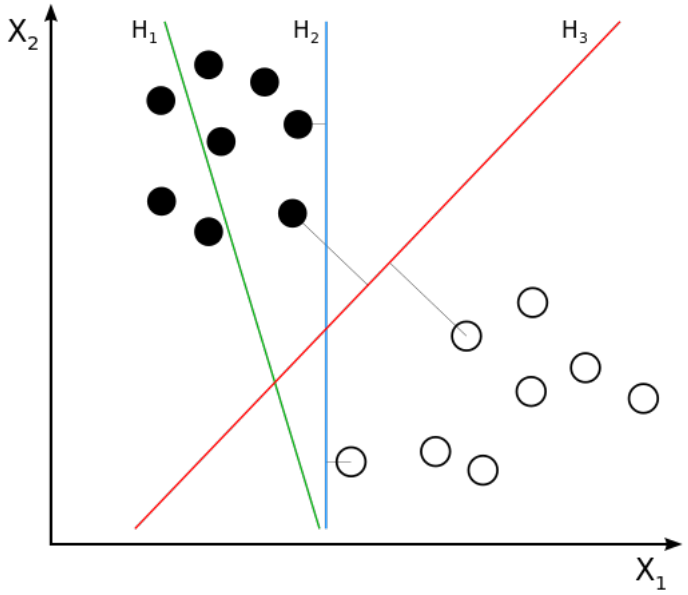

Figure 3: As in **Figure 2**, we are trying to separate labeled points with a line. A SVM does this iteratively, meaning line H_1 would be the first pass, H_2 the result of refitting, and H_3 the final classifier produced by the SVM.

References

[1] Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider, *Comparison of support vector machine and artificial neural network systems for drug/nondrug classification*, Journal of Chemical Information and Computer Sciences **43** (2003), no. 6, 1882–1889.

[2] Chih-Chung Chang and Chih-Jen Lin, *Libsvm: a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011), no. 3, 27.