CS251 Spring 2015 Lecture 15

Stephanie R Taylor

March 9, 2015

1 Fundamentals of Data Mining and Machine Learning

The goal for the course is to learn how to find meaningful patterns in data. We do that through judicious use of pre-processing, visualization and analysis. Before we go further, let's get some context. There are two, overlapping, subfields of computer science in which people are trying to find patterns in data. They are machine learning and data mining. Both use methods that are also used in the field of statistics.

Machine learning. Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders. (grabbed from the Wikipedia page on machine learning March 29, 2013)

Data Mining. The goal of data mining is finding and describing structural patterns in data (from the textbook).

A rule of thumb for the difference between machine learning and data mining is that machine learning focuses on prediction, based on known properties learned from the training data. Data mining focuses on the discovery of (previously) unknown properties on the data. (grabbed from the Wikipedia page on machine learning March 28, 2013).

We will be studying four categories of algorithms in this course.

1. Numerical prediction: the output to be predicted is not a discrete class but a numeric quantity. Regression is a popular form of numerical prediction.

2. Dimensionality reduction: Principal component analysis allows us to concisely capture structure in the data by forming a new (smaller) set of features. To do so, it takes advantage of the linear relationships among the original features.

3. Clustering: We seek to automatically group together observations with similar values.

4. Classification learning: the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. The is also called *supervised* learning. We think of the majority of features as input and the classification as output. Since the training data we use in this sort of learning has known classes, it has known output and the learning is called supervised. The algorithms we will learn include Naïve Bayes method, decision trees, neural networks.

2 Data mining and machine learning must be performed intelligently

There are two “theorems” that together indicate there is no sliver bullet. The first indicates there is no silver bullet algorithm. The second indicates there is no silver bullet distance metric.

2.1 No Free Lunch

The 'No Free Lunch Theorem' is one explanation for why machine learning methods require tuning, tweaking, and intelligent selection based on the data and the problem.

The NFL theorem states that all search and optimization algorithms have the same average performance over all problems. In other words, there is no machine learning algorithm–which all fall into the category of search and/or optimization–that will consistently outperform all other algorithms on all problems. Since not all problems are the same, and different ML methods have different internal models, it stands to reason that some ML method is going to be optimal for a problem, but that must be balanced by the same ML method having relatively poorer performance on a different problem.

Note that the NFL theorem makes the assumption that all problems are equally likely. It may be the case that there is a bias to the type of problems that exist in the world, in which case some ML methods may dominate other methods, on average. Factors other than performance–training time, or run time, for example–are not part of the NFL theorem considerations.

It's also very important to keep in mind that factors other than the differences between general learning methods can play a much more significant role in overall performance. Decisions about how to use prior information, the distribution and number of training examples, and the particulars of the cost, error or reward functions can overwhelm any inherent positive or negative impact on performance due to the selected machine learning method.

2.2 Ugly Duckling

The ugly duckling theorem has to do with distance metrics (termed predicates), which are methods used to determine if two feature vectors represent the same category or different categories. The theorem considers all possible ways of measuring similarity and claims that, for any two patterns, the number of possible predicates claiming the patterns are similar is constant. As with the NFL theorem, the ugly duckling theorem implies that careful selection of the predicates for a particular set of categories is critical to the success of a pattern recognition system.

> **Ugly Duckling Theorem**
> Given that we use a finite set of predicates that enables us to distinguish any two patterns under consideration, the number of predicates shared by any two such patterns is constant and independent of the choice of those patterns. Furthermore, if pattern similarity is based on the total number of predicates shared by two patterns, then any two patterns are ”equally similar.”
> (Duda, Hart, and Stork, Pattern Classification, 2001)

Another way to look at the ugly duckling theorem is to think of the patterns as being highly dependent on the distance metric. You can get your data to follow almost any pattern if you choose the right (or wrong) distance metric. So you need to choose the distance metric that makes the most sense for your data.

Take-home message: There is no catch-all method that will find a reasonable pattern in all data sets. You must choose your algorithms carefully and pay attention to the details (i.e. choose an appropriate distance metric).