In these lectures we will discuss some methods from the machine learning (ML) literature that we think will be useful for economists. There has been a fast growing literature in computer science and related fields developing new, and modifying existing, methods for analyzing large data sets. This literature builds heavily on traditional statistical methods, though often with new terminology and new questions. Many of these methods are focused on finding patterns in data sets that are useful for prediction and classification. The focus in this literature is typically on methods that “work,” more than on deriving asymptotic (large sample) results of the type that are common in the econometrics and mathematical statistics literature. There is also less emphasis on confidence intervals and standard errors. On the other hand there is a heavy emphasis on out-of-sample comparisons, in particular cross-validation. There is generally also less emphasis on causal effects as opposed to prediction.

Here we discuss some of the most prominent methods. These include regularized regression, where least squares parameter estimates are shrunk towards zero, with the most popular method LASSO. We also look at regression trees and their extensions, including random forests. We will also discuss unsupervised learning methods such as clustering algorithms. We discuss some of the principles, and in particular how these methods relate to methods that are traditionally more familiar to economists, and how they can be implemented in practice.

We also discuss some of the more recent literature where the focus is on modifying these methods to estimate causal effects.

Breiman (2001a) is a very good easy to read paper on the differences in culture between the traditional statistics literature (which is very much like the econometrics literature) and the machine learning literature. If there is one paper to read before the lectures, it is that one. Wu et al. (2008) is a good example of the difference in the cultures: the goal is often to derive algorithms rather than statistical properties. Even if some of the algorithms (like the EM algorithm and nearest-neighbor methods) are familiar to economists, the philosophy

**References**


Tan, P., M. Steinbach, and V. Kumar, (2006), Introduction to Data Mining, Addison Wesley.


**Schedule**

1:30-1.35 Welcome

1:35-2.30 Introduction to Supervised ML Concepts and Algorithms (Guido Imbens)

2:30-2.40 Break

2:40-3.35 Introduction to Unsupervised ML Concepts and Algorithms (Guido Imbens)

3:35-3.45 Break

3:45-4.40 Machine Learning and Causal Inference (Susan Athey)

4:40-4.50 Break

4:50-5.45 Unsupervised Learning: Applications to Networks and Text Mining (Susan Athey)