

Data Mining Techniques

Christopher Loudon, MS

The increase of computing power has given doctors, statisticians and others access to datasets on a scale not imagined when classical statistics was first developed. To handle such large datasets, a group of techniques called “data mining” was developed. In the past twenty years, these techniques have become a ubiquitous part of our lives through our use of technologies such as Google Search and spam filters. In medicine, data mining techniques are indispensable in the study of the human genome, proteomics and other fields where there are more variables of interest than observations. This creates a needle-in-the-haystack environment that renders the usual statistical methods useless. Nevertheless, data mining techniques have had extraordinary success in diverse areas such as prognosis and identifying heretofore unknown disease subtypes.

This talk will cover data mining techniques in general with a focus on their use in prediction of events. Various measures of accuracy will be discussed as will the problem of bias in the reporting of predictive ability.