Machine, Data and Learning (CS7.301)

Spring 2022, IIIT Hyderabad
21 Mar, Monday (Lecture 15)

Taught by Prof. Vikram Pudi

Overview of Data Analytics (contd.)

Classification (contd.)

Classification systems can be evaluated on one of several metrics – accuracy, running time, training time, memory usage, or model size.

Accuracy can be measured by taking the ratio of the number of correctly classified points to the total number of points (the holdout method). Further processes are stratification (ensuring that all classes are represented in the partitions), random subsampling (repeating holdout k times), and k-fold cross-validation.

Classifiers are sometimes combined, since different classifiers perform well on different datasets. They can be combined by *bagging* (taking a majority vote) or, more generally, *boosting* (giving a weight to each classifier's vote proportional to its accuracy).

There are many classification algorithms – k nearest neighbours (taking the most frequent class among the k nearest records), decision trees, bayesian methods (like naive bayes, which uses the classification of independent attributes of data points).