Machine Learning

Machine Learning Introductory Overview

Machine Learning includes a number of advanced statistical methods for handling regression and classification tasks with multiple dependent and independent variables. These methods include Support Vector Machines (SVM) for regression and classification, Naive Bayes for classification, and *k*-Nearest Neighbours (KNN) for regression and classification. Detailed discussions of these techniques can be found in Hastie, Tibshirani, & Freedman (2001); a specialized comprehensive introduction to support vector machines can also be found in Cristianini and Shawe-Taylor (2000).

Support Vector Machines (SVM)

This method performs regression and classification tasks by constructing nonlinear decision boundaries. Because of the nature of the feature space in which these boundaries are found, Support Vector Machines can exhibit a large degree of flexibility in handling classification and regression tasks of varied complexities. There are several types of Support Vector models including linear, polynomial, RBF, and sigmoid.

See the Support Vector Machines topic for more detailed information.

Naive Bayes

This is a well established Bayesian method primarily formulated for performing classification tasks. Given its simplicity, i.e., the assumption that the independent variables are statistically independent, Naive Bayes models are effective classification tools that are easy to use and interpret. Naive Bayes is particularly appropriate when the dimensionality of the independent space (i.e., number of input variables) is high (a problem known as the curse of dimensionality). For the reasons given above, Naive Bayes can often outperform other more sophisticated classification methods. A variety of methods exist for modeling the conditional distributions of the inputs including normal, lognormal, gamma, and Poisson.

Variance Components
Statistical Advisor
Distribution Tables
References Cited
Send Comments
Business Solutions
Free Resources
About Textbook

See the Naive Bayes Classifier topic for more detailed information.

k-Nearest Neighbors

k-Nearest Neighbors is a memory-based method that, in contrast to other statistical methods, requires no training (i.e., no model to fit). It falls into the category of Prototype Methods. It functions on the intuitive idea that close objects are more likely to be in the same category. Thus, in KNN, predictions are based on a set of prototype examples that are used to predict new (i.e., unseen) data based on the majority vote (for classification tasks) and averaging (for regression) over a set of *k*-nearest prototypes (hence the name *k*-nearest neighbors).

See the *k*-Nearest Neighbors topic for more detailed information.

To index