Samuel Wolfe

MSBA 206 Study Guide

KNN

Advantages – Simplicity and lack of parametric assumptions. With larger training sets these methods perform very well. This method especially performs well when each class is characterized by multiple combinations of predictors. Example is real-estate databases being able to predict fast sell time vs not.

Disadvantages – No time to required to estimate parameters from training data means it causes all KNN models to take longer when being applied to validation sets. Use of PCA is advised to assist with this. Use alternative versions like almost nearest neighbor. The second issue is number of records required is exponentially increasing with every new predictor. Curse of Dimensionality. KNN is a lazy learner, time consuming computation. Not ideal for real time.

Naïve Bayes

Advantages – Simplicity, computational efficiency, good classification performance, and ability to handle categorical variables directly. Often outperforms more sophisticated methods of classifiers. This effect is even more pronounced at extreme data set sizes.

Disadvantages – First is it requires a very large number of records to obtain good results. Second if a predictor category is not present in the training data it assumes that a new record with that category of the predictor has zero probability. This can be an issue. Can be corrected or countered marginally with smoothing. Third good performance is obtained when the goal is classification or ranking or records according to their probability of belonging to a certain class. If the goal is to estimate the probability of class membership (propensity) this method provides very biased results.

Classification and Regression Trees

Advantages – Good off-the-shelf classifiers and predictors. Useful for variable selection, with the most important predictors usually showing up at the top of the tree. Little effort from users in the following senses: First no need for transformation of variables. Second variable subset selection is automatic since it is a part of split selection. Intrinsically robust to outliers since the choice of a split depends on the ordering of values and not on the absolute magnitudes of these values. Very sensitive to changes in test data. Great for handling missing data or creating very explicit rules.

Disadvantages – Likely to miss relationships between predictors. Require large data sets to create good classifiers. Trees are computationally expensive. Pruning or cross validation adds to the computation time. Trees favor predictors with larger potential splits.