Contents

1

Introduction

Serious data miners are often faced with thousands of candidate features for their prediction or classification application, with most of the features being of little or no value. Worse still, many of these features may be useful only in combination with certain other features while being practically worthless alone or in combination with most others. Some features may have enormous predictive power, but only within a small, specialized area of the feature space. The problems that plague modern data miners are endless.

My own work over the last 20+ years in the financial market domain has involved examination of decades of price data from hundreds or thousands of markets, searching for exploitable patterns of price movement. I could not have done it without having a large and up-to-date collection of data mining tools.

Over my career I have accumulated many such tools, and I still keep up to date with scientific journals, exploring the latest algorithms as they appear. In my recent book "Data Mining Algorithms in C++", published by the Apress division of Springer, I presented many of my favorite algorithms. I now continue this presentation, divulging details and source code of key modules in my *VarScreen* variable screening program that has been made available for free download and frequently updated over the last few years. The topics included in this text are:

Hidden Markov models are chosen and optimized according to their multivariate correlation with a target. The idea is that observed variables are used to deduce the current state of a hidden Markov model, and then this state information is used to estimate the value of an unobservable target variable. This use of memory in a time series discourages whipsawing of decisions and enhances information usage.

Forward Selection Component Analysis uses forward and optional backward refinement of maximum-variance-capture components from a subset of a large group of variables. This hybrid combination of principal components analysis with stepwise selection lets us whittle down enormous feature sets, retaining only those variables that are most important.

Local Feature Selection identifies predictors that are optimal in localized areas of the feature space but may not be globally optimal. Such predictors can be effectively used by nonlinear models but are neglected by many other feature selection algorithms that require global predictive power. Thus, this algorithm can detect vital features that are missed by other feature selection algorithms.

Stepwise selection of predictive features is enhanced in three important ways. First, instead of keeping a single optimal subset of candidates at each step, this algorithm keeps a large collection of high-quality subsets and performs a more exhaustive search of combinations of predictors that have joint but not individual power. Second, cross validation is used to select features, rather than using the traditional in-sample performance. This provides an excellent means of complexity control, resulting in greatly improved out-of-sample performance. Third, a Monte-Carlo permutation test is applied at each addition step, assessing the probability that a good-looking feature set may be not good at all, but rather just lucky in its attainment of a lofty performance criterion.

Nominal-to-ordinal conversion lets us take a potentially valuable nominal variable (a category or class membership) that is unsuitable for input to a prediction model, and assign to each category a sensible numeric value that can be used as a model input.

As is usual in my books, all of these algorithm presentations begin with an intuitive overview, continue with essential mathematics, and culminate in complete, heavily commented source code. In most cases, one or more sample applications are shown to illustrate the technique.

The *VarScreen* program, available as a free download from my website TimothyMasters.info, implements all of these algorithms and many more. This program can serve as an effective variable screening program for data mining applications.