1. Record Nr. UNINA9910409993103321

Autore	Masters Timothy
Titolo	Modern Data Mining Algorithms in C++ and CUDA C : Recent Developments in Feature Extraction and Selection Algorithms for Data Science / / by Timothy Masters
Pubbl/distr/stampa	Berkeley, CA : , : Apress : , : Imprint : Apress, , 2020
ISBN	9781484259887 1484259882
Edizione	[1st ed. 2020.]
Descrizione fisica	1 online resource (ix, 228 pages)
Disciplina	006.312
Soggetti	Data mining Computer software Statistics Programming languages (Electronic computers) Data Mining and Knowledge Discovery Professional Computing Statistics, general Programming Languages, Compilers, Interpreters
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Includes index.
Nota di contenuto	1. Introduction -- 2. Forward Selection Component Analysis -- 3. Local Feature Selection -- 4. Memory in Time Series Features -- 5. Stepwise Selection on Steroids -- 6. Nominal-to-Ordinal Conversion.
Sommario/riassunto	As a serious data miner you will often be faced with thousands of candidate features for your prediction or classification application, with most of the features being of little or no value. You'll know that 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. This book helps you solve this problem by presenting modern feature selection techniques and the code to implement them. Some of these techniques are: Forward

selection component analysis Local feature selection Linking features and a target with a hidden Markov model Improvements on traditional stepwise selection Nominal-to-ordinal conversion All algorithms are intuitively justified and supported by the relevant equations and explanatory material. The author also presents and explains complete, highly commented source code. The example code is in C++ and CUDA C but Python or other code can be substituted; the algorithm is important, not the code that's used to write it. You will: Combine principal component analysis with forward and backward stepwise selection to identify a compact subset of a large collection of variables that captures the maximum possible variation within the entire set. Identify features that may have predictive power over only a small subset of the feature domain. Such features can be profitably used by modern predictive models but may be missed by other feature selection methods. Find an underlying hidden Markov model that controls the distributions of feature variables and the target simultaneously. The memory inherent in this method is especially valuable in high-noise applications such as prediction of financial markets. Improve traditional stepwise selection in three ways: examine a collection of 'best-so-far' feature sets; test candidate features for inclusion with cross validation to automatically and effectively limit model complexity; and at each step estimate the probability that our results so far could be just the product of random good luck. We also estimate the probability that the improvement obtained by adding a new variable could have been just good luck. 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.