1. Record Nr. UNISA996198490203316

Autore	Larose Daniel T.
Titolo	Discovering knowledge in data : an introduction to data mining / / Daniel T. Larose, Chantal D. Larose
Pubbl/distr/stampa	Hoboken, New Jersey : , : IEEE, , 2014 ©2014
ISBN	1-118-87357-2 1-118-87405-6 1-118-87358-0
Edizione	[2nd ed.]
Descrizione fisica	1 online resource (336 p.)
Collana	Wiley Series on Methods and Applications in Data Mining
Classificazione	COM021040COM021030
Disciplina	006.3/12
Soggetti	Data mining
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Includes index.
Nota di bibliografia	Includes bibliographical references at the end of each chapters and index.
Nota di contenuto	DISCOVERING KNOWLEDGE IN DATA; Contents; Preface; 1 An Introduction to Data Mining; 1.1 What is Data Mining?; 1.2 Wanted: Data Miners; 1.3 The Need for Human Direction of Data Mining; 1.4 The Cross-Industry Standard Practice for Data Mining; 1.4.1 Crisp-DM: The Six Phases; 1.5 Fallacies of Data Mining; 1.6 What Tasks Can Data Mining Accomplish?; 1.6.1 Description; 1.6.2 Estimation; 1.6.3 Prediction; 1.6.4 Classification; 1.6.5 Clustering; 1.6.6 Association; References; Exercises; 2 Data Preprocessing; 2.1 Why do We Need to Preprocess the Data?; 2.2 Data Cleaning; 2.3 Handling Missing Data 2.4 Identifying Misclassifications2.5 Graphical Methods for Identifying Outliers; 2.6 Measures of Center and Spread; 2.7 Data Transformation; 2.8 Min-Max Normalization; 2.9 Z-Score Standardization; 2.10 Decimal Scaling; 2.11 Transformations to Achieve Normality; 2.12 Numerical Methods for Identifying Outliers; 2.13 Flag Variables; 2.14 Transforming Categorical Variables into Numerical Variables; 2.15 Binning Numerical Variables; 2.16 Reclassifying Categorical Variables; 2.17 Adding an Index Field; 2.18 Removing Variables that are Not Useful; 2.19 Variables that Should Probably Not Be Removed 2.20 Removal of Duplicate Records2.21 A Word About Id Fields; THE R ZONE; References; Exercises; Hands-On Analysis; 3 Exploratory Data

Analysis; 3.1 Hypothesis Testing Versus Exploratory Data Analysis; 3.2 Getting to Know the Data Set; 3.3 Exploring Categorical Variables; 3.4 Exploring Numeric Variables; 3.5 Exploring Multivariate Relationships; 3.6 Selecting Interesting Subsets of the Data for Further Investigation; 3.7 Using EDA to Uncover Anomalous Fields; 3.8 Binning Based on Predictive Value; 3.9 Deriving New Variables: Flag Variables; 3.10 Deriving New Variables: Numerical Variables
3.11 Using EDA to Investigate Correlated Predictor Variables3.12 Summary; THE R ZONE; Reference; Exercises; Hands-On Analysis; 4 Univariate Statistical Analysis; 4.1 Data Mining Tasks in Discovering Knowledge in Data; 4.2 Statistical Approaches to Estimation and Prediction; 4.3 Statistical Inference; 4.4 How Confident are We in Our Estimates?; 4.5 Confidence Interval Estimation of the Mean; 4.6 How to Reduce the Margin of Error; 4.7 Confidence Interval Estimation of the Proportion; 4.8 Hypothesis Testing for the Mean; 4.9 Assessing the Strength of Evidence Against the Null Hypothesis
4.10 Using Confidence Intervals to Perform Hypothesis Tests4.11 Hypothesis Testing for the Proportion; THE R ZONE; Reference; Exercises; 5 Multivariate Statistics; 5.1 Two-Sample t-Test for Difference in Means; 5.2 Two-Sample Z-Test for Difference in Proportions; 5.3 Test for Homogeneity of Proportions; 5.4 Chi-Square Test for Goodness of Fit of Multinomial Data; 5.5 Analysis of Variance; 5.6 Regression Analysis; 5.7 Hypothesis Testing in Regression; 5.8 Measuring the Quality of a Regression Model; 5.9 Dangers of Extrapolation; 5.10 Confidence Intervals for the Mean Value of Given 5.11 Prediction Intervals for a Randomly Chosen Value of Given

Sommario/riassunto

"This is a new edition of a highly praised, successful reference on data mining, now more important than ever due to the growth of the field and wide range of applications. This edition features new chapters on multivariate statistical analysis, covering analysis of variance and chi-square procedures; cost-benefit analyses; and time-series data analysis. There is also extensive coverage of the R statistical programming language. Graduate and advanced undergraduate students of computer science and statistics, managers/CEOs/CFOs, marketing executives, market researchers and analysts, sales analysts, and medical professionals will want this comprehensive reference"--