Ἄ γὰρ δεῖ μαθόντας ποιεῖν, ταῦτα ποιοῦντες μανθάνομεν.
Ciò che dobbiamo imparare a fare, lo impariamo facendolo.
Aristotele, Etica Nicomachea

Susi Dulli, Sara Furini, Edmondo Peron

Data mining

Metodi e strategie

Springer

SUSI DULLI
Dipartimento di Matematica Pura ed Applicata
Università di Padova, Padova
dulli@math.unipd.it

SARA FURINI
Miriade
Noventa Padovana (PD)
s.furini@miriade.it

EDMONDO PERON
Miriade
Noventa Padovana (PD)
e.peron@miriade.it

In copertina: “Il Mondo” (2005) di Filippo Maconi. 120×100. Particolare
Tecnica mista

ISBN 978-88-470-1162-5 Springer Milan Berlin Heidelberg New York
ISBN 978-88-470-1176-2 (eBook) Springer Milan Berlin Heidelberg New York

Springer-Verlag fa parte di Springer Science+Business Media
springer.com
© Springer-Verlag Italia, Milano 2009

Quest'opera è protetta dalla legge sul diritto d'autore e la sua riproduzione è ammessa solo ed esclusivamente nei limiti stabiliti dalla stessa. Le fotocopie per uso personale possono essere effettuate nei limiti del 15% di ciascun volume dietro pagamento alla SIAE del compenso previsto dall'art. 68, commi 4 e 5, della legge 22 aprile 1941 n. 633. Le riproduzioni per uso non personale e/o oltre il limite del 15% potranno avvenire solo a seguito di specifica autorizzazione rilasciata da AIDRO, Via Corso di Porta Romana n. 108, Milano 20122, e-mail segreteria@aidro.org e sito web www.aidro.org. Tutti i diritti, in particolare quelli relativi alla traduzione, alla ristampa, all'utilizzo di illustrazioni e tabelle, alla citazione orale, alla trasmissione radiofonica o televisiva, alla registrazione su microfilm o in database, o alla riproduzione in qualsiasi altra forma (stampata o elettronica) rimangono riservati anche nel caso di utilizzo parziale. La violazione delle norme comporta le sanzioni previste dalla legge.

9 8 7 6 5 4 3 2 1

Impianti: PTP-Berlin, Protago $\text{\TeX}$ -Production GmbH, Germany (www.ptp-berlin.eu)
Progetto grafico della copertina: Francesca Tonon, Milano
Stampa: Signum Srl, Bollate (MI)
Stampato in Italia

Springer-Verlag Italia srl – Via Decembrio 28 – 20137 Milano

Prefazione

In passato lo sviluppo dell'informatica e della telematica ha dato risalto alla tecnologia e non all'applicazione della tecnologia ai processi che appartengono ai molteplici aspetti della vita umana, quali i processi produttivi, amministrativi e di marketing. Oggi, la complessità dei fenomeni (processi) presi in esame, fa sì che ci sia una tendenza a considerare importante l'aspetto applicativo e questo sia per il gran numero di indicatori presenti che per la massa di informazioni disponibili, oltre al fatto che le variabili che si trattano sono sì numerose ma soprattutto legate fra loro. Quest'ultimo un aspetto comune ed intrinseco ai fenomeni reali, e al quale ci si riferisce col termine di multidimensionalità, e che sta nella natura stessa della realtà. L'emergere di questa tendenza a focalizzarsi giustamente sull'applicazione della tecnologia più che sulla tecnologia stessa, è avvalorato dagli sviluppi recenti della tecnologia dell'informazione, che con la messa a disposizione di strumenti hardware e software sempre più potenti e performanti, permette alle imprese di raccogliere ed organizzare grandi quantità di dati al fine di ottenere informazioni utili a supporto delle decisioni. A ciò si aggiunge per le aziende, l'alto livello di competizione, costantemente messo a rischio dalla crescente globalizzazione oggi presente sul mercato, che obbliga chi è impegnato in attività direzionali ad una sempre maggior attenzione all'utilizzo di strumenti informatici e a metodi di analisi statistiche che possono influenzare decisioni determinanti al raggiungimento di risultati strategici per l'impresa. Le metodologie di analisi di grandi moli di dati al fine di individuare tendenze nei comportamenti dei clienti, note come sistemi di data mining, costituiscono oggi un approccio consolidato che ha trovato la sua più significativa affermazione grazie a strumenti software di mercato ad alte prestazioni ed affidabilità.

Essendo il data mining un argomento molto vasto, lo scopo di questo volume è quello di trasmettere le idee presentando gli argomenti principali, privilegiando gli aspetti procedurali e di calcolo. Quando abbiamo ritenuto opportuno, abbiamo reso l'esposizione più concreta introducendo pseudo codice ed esempi. Il libro è rivolto a un corso universitario ma può essere usato anche per chi vuole utilizzare questi concetti nella realtà aziendale. Si partirà con la definizione di KDD distinta da quella di data mining e con alcuni termini essenziali e usati in letteratura.

Verranno accennati i principali passi necessari per la scoperta di conoscenza e sarà posto l'accento più alla parte algoritmica di alcuni dei principali strumenti di data mining. Si vedranno, spiegati in termini algoritmici, il clustering, gli alberi decisionali, le regole di associazione, le serie temporali, i metodi di classificazione e le reti neurali. Nessuna meraviglia, quindi, che il termine data mining diventi di uso familiare in molte organizzazioni aziendali, anche se resta il fatto che il suo impiego presuppone la disponibilità di conoscenze e strumenti specifici che ne condizionano ovviamente il costo. D'altro canto il profitto che ne consegue, come testimoniano molti casi aziendali, giustifica abbondantemente la scelta di intraprendere questa attività.

Gli autori ringraziano i collaboratori del mondo delle imprese con i quali hanno condiviso progetti di stage per gli studenti in tematiche di Business Intelligence. Gli autori desiderano inoltre ringraziare la casa editrice Springer Italia per avere accettato la proposta di realizzazione di questo testo; in particolare ringraziano la dottoressa Francesca Bonadei per la preziosa collaborazione nella redazione e revisione dell'opera. Siamo grati infine a quanti vorranno segnalare errori e suggerimenti presenti in questa versione in modo da permettere la pubblicazione di nuove edizioni aggiornate e corrette.

Padova, gennaio 2009

Susi Dulli
Sara Furini
Edmondo Peron

Indice

1	Introduzione	1
1.1	Definizione di data mining e KDD	1
1.2	Fasi dell'attività di KDD e data mining	3
1.3	Tecniche di data mining	5
1.4	Criteri per scegliere gli strumenti per il data mining	6
1.5	Applicazioni delle tecniche	7
1.6	Organizzazione del testo	8
1.7	Esercizi di riepilogo	10
2	Trattamento preliminare dei dati	13
2.1	Campionamento	13
2.1.1	Premessa	14
2.1.2	Training and Test o Holdout	14
2.1.3	<i>K</i> -fold Cross Validation	15
2.1.4	Leave one out	16
2.1.5	Bagging	16
2.1.6	Boosting	16
2.2	Inferenza	16
2.2.1	La distribuzione normale	17
2.3	Pre-elaborazione dei dati	18
2.3.1	Data Cleaning	18
2.3.2	Data Missing	18
2.3.3	Dati inaccurati	19
2.3.4	Discretizzazione	20
2.4	Analisi esplorativa	21
2.4.1	Tipi di attributi	21
2.4.2	Analisi univariata	23
2.4.3	Indicatori di tendenza centrale	24
2.4.4	Misure di dispersione	25
2.4.5	Misure di eterogeneità e similarità	26
2.4.6	Riduzione della dimensionalità	27

2.5	Rappresentazioni grafiche per distribuzioni univariate	28
2.5.1	Istogrammi	29
2.5.2	Distribuzioni di frequenza per i dati qualitativi	30
2.6	Esercizi di riepilogo	32
3	Misure di distanza e di similarità	37
3.1	Concetto di distanza	37
3.2	Distanza Euclidea	39
3.3	Distanza di Minkowski	39
3.4	Distanza di Lagrange-Tchebychev	40
3.5	Distanza di Mahalanobis	40
3.6	Similarità fra vettori binari (SMC)	41
3.7	Correlazione	41
3.8	Esercizi di riepilogo	42
4	Cluster Analysis	45
4.1	Distinzione fra Classificazione e Cluster Analysis	45
4.2	Cluster Analysis	46
4.3	Gli algoritmi di clustering	47
4.4	Algoritmi partizionativi	48
4.4.1	Algoritmo <i>K</i> -Means	48
4.5	I metodi gerarchici agglomerativi	52
4.5.1	AGNES (AGglomerative NESTing)	54
4.5.2	DIANA (DIVisive ANALysis)	55
4.6	Clustering basato sulla densità	56
4.6.1	DBSCAN	56
4.6.2	Complessità dell'algoritmo DBSCAN	61
4.6.3	I parametri dell'algoritmo	61
4.7	Esercizi di riepilogo	61
5	Metodi di classificazione	63
5.1	Alberi di decisione	63
5.1.1	Algoritmo ID3	65
5.1.2	Tipi di attributi	67
5.1.3	Algoritmo C4.5	69
5.1.4	Potatura di un albero di decisione	70
5.2	Classificatori bayesiani	70
5.3	Nearest Neighbor clustering	75
5.4	Reti neurali artificiali	77
5.4.1	Il neurone biologico	77
5.4.2	Il modello matematico del neurone	78
5.4.3	Algoritmo di backpropagation	80
5.5	Valutazione dei metodi di classificazione	84
5.5.1	Matrice di confusione per problemi a due classi	84
5.5.2	Curva ROC	86

5.5.3	Curva lift	90
5.6	Esercizi di riepilogo	91
6	Serie Temporal	95
6.1	Criteri di similarità	95
6.2	Dynamic Time Warping	99
6.2.1	Definizione del problema	99
6.2.2	Formalizzazione dell'algoritmo	101
6.3	Il filtro di Kalman	104
6.3.1	Le origini del filtro	105
6.3.2	Le origini probabilistiche del filtro	106
6.3.3	L'algoritmo del filtro di Kalman discreto	107
6.3.4	Parametri del filtro e regolazione	108
6.3.5	Stima di una variabile casuale con il filtro di Kalman	109
6.4	Analisi di regressione	113
6.4.1	Retta di regressione di Y rispetto a X	114
6.4.2	Retta di regressione di X rispetto a Y	114
6.4.3	Relazione fra i coefficienti angolari b_1 e b_2	115
6.5	Esercizi di riepilogo	116
7	Analisi delle associazioni	119
7.1	Formalizzazione del problema	120
7.2	Algoritmo Apriori	123
7.2.1	Generazione degli itemset frequenti	125
7.2.2	Generazione delle regole	128
7.3	Esercizi di riepilogo	133
8	Analisi dei link	137
8.1	Prestigio	137
8.2	Matrice dominante degli autovettori	138
8.3	Pagerank	140
8.4	Autorità e connessioni	142
8.5	Esercizi di riepilogo	143
	Soluzioni degli esercizi	145
	Glossario dei termini di data mining	149
	Bibliografia	163
	Indice analitico	175