

---

# Indice

<b>1</b>	<b>Introduzione</b>	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
<b>2</b>	<b>Trattamento preliminare dei dati</b>	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
<b>3</b>	<b>Misure di distanza e di similarità</b> .....	<b>37</b>
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
<b>4</b>	<b>Cluster Analysis</b> .....	<b>45</b>
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
<b>5</b>	<b>Metodi di classificazione</b> .....	<b>63</b>
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
<b>6</b>	<b>Serie Temporali .....</b>	<b>95</b>
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
<b>7</b>	<b>Analisi delle associazioni .....</b>	<b>119</b>
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
<b>8</b>	<b>Analisi dei link .....</b>	<b>137</b>
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
	<b>Soluzioni degli esercizi .....</b>	<b>145</b>
	<b>Glossario dei termini di data mining .....</b>	<b>149</b>
	<b>Bibliografia .....</b>	<b>163</b>
	<b>Indice analitico .....</b>	<b>175</b>