Analisi dei dati ed estrazione della conoscenza

Esercizi

Gianluca Amato

18 maggio 2005

1. Perché è oppurtuno *potare* un albero di classificazione?
2. Dato un albero di classificazione, ci sono due possibilità: (a) convertire l'albero in un insieme di regole di classificazione e potare queste ultime, (b) potare direttamente l'albero di decisione. Qual è il vantaggio della soluzione (a) rispetto alla (b) ?
3. Si considerino i seguenti dati aggregati

department	**status**	**age**	**salary**	**count**
sales	junior	31 ... 35	46K ... 50K	30
sales	junior	26 ... 30	26K ... 30K	40
sales	junior	31 ... 35	31K ... 35K	40
systems	junior	21 ... 25	46K ... 50K	20
systems	senior	31 ... 35	66K ... 70K	5
systems	junior	26 ... 30	46K ... 50K	3
systems	senior	41 ... 45	66K ... 70K	3
marketing	senior	36 ... 40	46K ... 50K	10
marketing	junior	31 ... 35	41K ... 45K	4
secretary	senior	46 ... 50	36K ... 40K	4
secretary	junior	26 ... 30	26K ... 20K	6

   dove, per ogni riga, **count** è il numero di tuple dati aventi quelle caratteristiche. Si supponga che **salary** sia l'attributo classe.

   Come si modifica l'algoritmo ID3 in modo da tener conto del campo **count** per ogni tupla generalizzata? Usa la versione modificata per costruire un albero di classificazione per questi dati.
4. Data una istanza con i valori "systems", "junior", "26 ... 30" rispettivamente per gli attributi **department**, **status** e **age**, quale sarebbe la classe predetta da un classificatore Bayesiano naive, assumendo lo stesso insieme di addestramento dell'esercizio precedente?
5. Si considerino i seguenti dati:

x	**y**
1	0.37588
2	1.00450
3	0.83363
4	1.14230
5	0.69800
6	1.85084
7	0.96593
8	1.01001

x	**y**
9	1.81403
10	1.55026
11	1.29721
12	1.62657
13	1.37217
14	1.78812
15	1.79291
16	1.68171

Calcolare i coefficienti β_0 e β_1 per l'equazione di regressione $\log y = \beta_0 + \beta_1 \log x$.

6. Si consideri il seguente insieme di dati:

name	**gender**	**trait-1**	**trait-2**	**trait-3**	**trait-4**
Kevan	M	N	P	P	N
Caroline	F	N	P	P	N
Erik	M	P	N	N	P

Ogni riga rappresenta una persona, di cui sono noti nome (la chiave primaria), il sesso e vari *trait* (tratti, caratteristiche). Ad esempio, un *trait* potrebbe essere "ama i giochi di ruolo". L'attributo **gender** è simmetrico, gli altri sono asimmetrici. I dati sono utilizzati da un servizio che tenta di trovare coppie di "amici di penna" compatibili.

Si supponga che la distanza tra istanze venga calcolata solo sulla base degli attributi asimmetrici:

- mostra la *matrice di contingenza* per ogni coppia di persone
- calcola il *simple matching coefficient* per ogni coppia
- calcola il *coefficiente di Jaccard* per ogni coppia
- qual è la migliore coppia di "amici di penna" e quale la peggiore?
- supponete di inserire anche la variabile simmetrica **gender** nel calcolare la distanza. Sulla base dell'indice di Jaccard, qual è la coppia più compatibile?

7. Cos'è una analisi di raggruppamento? Descrivi brevemente i seguenti approcci alla analisi di raggruppamento: metodi basati sulle partizioni, metodi geratchici, metodi basati sulle densità.

8. Supponete di voler usare l'algortimo delle k-medie per raggruppare i seguenti punti (calcolando la distanza come distanza euclidea)

$$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9) \ .$$

Supponete di scegliere inizialmente A_1, B_1 e C_1 come centri dei cluster. Usate l'algoritmo delle k-medie per mostrare

- i tre cluster dopo il primo ciclo di esecuzione;
- i tre cluster finali.