Capitolo 1

Dall'Inferenza all'Induzione

1.1 Introduzione: cosè la scoperta di conoscenza?

Consideriamo lo schema relazionale di fig. 1.1. L'utilizzo tipico di un siffatto database consiste in un vasto numero di operazioni di aggiornamento. Una struttura simile, ad esempio, la si trova in tutte le università. L'elevato numero di aggiornamenti/inserimenti giustifica la struttura normalizzata. Incidentalmente, il database può essere utilizzato anche per effettuare queries abbastanza semplici:

1. Che voti hanno avuto negli esami di Informatica gli studenti di Cosenza?

2. Che media hanno gli studenti del corso di Data Mining?

3. Qual'è lo studente che ha avuto il voto più alto?

L'algebra relazionale, e nel concreto il linguaggio SQL, hanno permesso di esprimere *inferenza*: ovvero, di estrarre informazione dettagliata da un insieme di dati. Formalmente, l'inferenza si può esprimere come $\vdash$ tale che $\mathbf{DB} \vdash \tau$, dove τ è una tupla di uno schema relazionale e **DB** è un database (ovvero un insieme di tabelle). Le espressioni relazionali (le queries SQL) sono lo strumento linguistico che ci permette di concentrarci su sottoinsiemi del database di interesse che si possono inferire: in pratica, se q è una query, $\mathbf{DB}(q) = \{\tau | \mathbf{DB} \vdash_q \tau\}$.

Nel nostro caso, esempi di espressioni che soddisfano le richieste esemplificate sono:

1.
```
SELECT Voto
FROM Esame INNER JOIN Studente
        ON (Studente = ID)
WHERE Residenza = 'Cosenza'
```

Studente

ID	*Nome*	*Eta′*	*Residenza*	*Voto di laurea*
1	Enrico	26	Rende	105
2	Alfredo	27	Cosenza	100
3	Domenico	30	Napoli	96
4	Luisa	24	Lecce	108
5	Giovanna	27	Taranto	104
6	Diego	32	Caserta	94

Esame

Studente	*Corso*	*Data*	*Voto*
1	1	13-02-2002	25
1	2	14-02-2002	26
2	1	21-02-2002	23
2	2	15-02-2002	24
3	4	14-02-2002	18
3	3	14-02-2002	20
4	5	15-02-2002	30
4	3	16-02-2002	29
4	4	18-02-2002	28
5	3	18-02-2004	27
5	4	20-02-2004	28
6	1	21-02-2002	18
6	6	13-02-2002	18

Corso

ID	*Nome*	*Settore*
1	Sistemi Distribuiti	Telematica
2	Programmazione ad Oggetti	Informatica
3	Statistica	Matematica
4	Data Mining	Informatica
5	Intelligenza Artificiale	Informatica
6	Comunicazioni Elettriche	Telematica

$$\forall s, e, d, v \quad (\quad Esame(s, e, d, v) \rightarrow \exists i_1, i_2, x, y, z, c(Studente(i_1, s, x, y, z) \wedge Corso(i_2, e, c)) \ )$$

Figura 1.1: Le tre tabelle rappresentano il database di studenti ed esami sostenuti dagli stessi

2.
```
SELECT AVG(Voto)
FROM Esame INNER JOIN Corso
        ON (Corso = ID)
WHERE Nome = 'Data Mining'
```

3.
```
SELECT Nome
FROM Studente INNER JOIN Corso
        ON (Corso = ID)
WHERE Voto =
        (SELECT MAX(Voto
         FROM Esame)
```

Una trattazione dettagliata dell'inferenza si può trovare in [1]: qui ci interessa invece analizzarne le relazioni e le differenze con il concetto fondamentale di questa dispensa: l'*induzione*. Per comprendere meglio tale concetto, consideriamo il seguente problema. Supponiamo di voler capire, dalla nostra base di dati, come sono fatti *di solito* gli studenti che hanno voto di laurea alto. Si tratta, in pratica, di caratterizzare gli studenti profittevoli in base a tutte le caratteristiche che si possono desumere dalla base di dati. Si intuisce immediatamente che la rappresentazione normalizzata esemplificata in fig. 1.1 non aiuta di certo. L'informazione è troppo sparsa, e non è possibile (almeno non visivamente cercare di identificare eventuali relazioni ta i dati. La prima cosa da fare quindi è una *trasformazione* dei dati in un formato più facile da interpretare.

ID	*Nome*	*Eta'*	*Residenza*	*Telematica*	*Informatica*	*Matematica*	*Voto di laurea*
1	Enrico	26	Rende	Si	Si	No	105
2	Alfredo	27	Cosenza	Si	Si	No	100
3	Domenico	30	Milano	No	Si	Si	96
4	Luisa	24	Lecce	No	Si	Si	108
5	Giovanna	27	Taranto	No	Si	No	104
6	Diego	32	Bergamo	Si	No	No	94

La tabella di cui sopra è stata ottenuta agendo sulla tabella *Studente* di figura 1.1. Ogni tupla rappresenta uno studente, che è l'oggetto principale di cui vogliamo *misurare* alcune caratteristiche. In più, le caratteristiche sono state ottenute incrociando i dati della tabella *Studente* con quelli della tabella *Corso* (utilizzando la tabella *Esame*): per ogni settore scientifico-disciplinare, è stata aggiunta una colonna che descrive la relazione tra uno studente e tale settore. In pratica, ci interessa capire se lo studente ha auto a che fare con tale settore scientifico-disciplinare (sostenendo un qualche esame in quel settore). Così, lo studenteEnrico ha avuto a che fare con esami di *Informatica* e *telematica*, mentre lo studente domenico si è cimentato essenzialmente in esami di *Informatica* e *Matematica*.

Possiamo adesso provare a riformulare il nostro intento originario: guardando alla tabella sopra definita, ci interessa capire quali combinazioni di fattori portano gli studenti ad avere voti alti. Questa volta abbiamo tutti i dati che ci potrebbero interessare, ma c'è un ulteriore problema che non abbiamo ancora risolto: che significa "voto alto"? Quand'è che possiamo dire che uno studente ha un voto alto, e quando invece ha un voto basso? È chiaro che quello che ci manca questa volta è la caratterizzazione del *Concetto* di interesse: nella tabella sopra definita il concetto non è esplicitato, e quindi non può essere opportunamente indagato. Proviamo a formulare il concetto in maniera rigorosa: diciamo

che un voto è *Alto* quando è superiore a 104; è *Medio* quando è inferiore a 105 ma superiore a 97; è infine *Basso* quando è inferiore a 98. Nella nostra tabella:

ID	*Nome*	*Eta'*	*Residenza*	*Telematica*	*Informatica*	*Matematica*	*Voto di laurea*
1	Enrico	26	Rende	Si	Si	No	Alto
2	Alfredo	27	Cosenza	Si	Si	No	Medio
3	Domenico	30	Milano	No	Si	Si	Basso
4	Luisa	24	Lecce	No	Si	Si	Alto
5	Giovanna	27	Taranto	No	Si	No	Alto
6	Diego	32	Bergamo	Si	No	No	Basso

Siamo quindi pronti a caratterizzare il concetto di nostro interesse. Riformulando il problema, vogliamo sapere in quale coincidenza di fattori appare in una tupla il valore *Alto* in corrispondenza dell'attributo *Voto di Laurea*. In particolare, vorremmo analizzare i nostri dati ed essere in grado di affermare con sicurezza che, in corrispondenza di un valore specifico per gli altri attributi, il concetto di nostro interesse sia caratterizzato: ad esempio, ci interessano regole generali del tipo: "chi sostiene esami in *Informatica* ha un voto alto". Se tuttavia analizziamo la nostra tabella, scopriamo che una tale caratterizzazione non si evince. L'attributo *Matematica* non caratterizza il voto di laurea, poiché al valore *No* corrispondono i concetti *Alto* , *Medio* e *Basso*, e al valore *Si* corrispondono i concetti *Alto* e *Basso*. La stessa cosa accade per l'attributo *Telematica*. Questi due valori hanno una caratteristica che li accomuna: i valori che possono assumere sono equi-distribuiti, e tuttavia non riescono a caratterizzare il concetto di interesse. Una situazione diversa avviene invece con l'attributo *Informatica*: quest'ultimo, infatti, presenta una estremamente bassa variabilità, e nel contempo non caratterizza i dati. Più precisamente, quasi tutti gli studenti descritti nel nostro esempio hanno sostenuto esami di informatica, per cui l'attributo esibisce nella colonna corrispondente un (quasi) unico valore *Si*. Si può affermare che l'attributo non *discrimina*, e quindi può essere eliminato. Se pensiamo invece agli altri attributi, la situazione è esattamente l'opposto: *ID* è unico per ogni studente (è imposto dalla normalizzazione); di solito non si trovano molti nomi simili; l'età e la provenienza possono essere estremamente frammentate. Questo ci porta a pensare che anche i primi due attributi possono essere eliminati dalla nostra tabella.

Eta'	*Residenza*	*Telematica*	*Matematica*	*Voto di laurea*
26	Rende	Si	No	Alto
27	Cosenza	Si	No	Medio
30	Milano	No	Si	Basso
24	Lecce	No	Si	Alto
27	Taranto	No	No	Alto
32	Bergamo	Si	No	Basso

Abbiamo quindi eliminato tre attributi: *Informatica*, *ID* e *Nome*. È importante notare che l'eliminazione è dovuta a caratteristiche diverse che portano alla stessa conclusione: *Informatica* non può caratterizzare perché è troppo generica (lo stesso valore è condiviso pressoché da tutti), mentre *ID* e *Nome* sono troppo

frammentati e quindi non potrebbero caratterizzare (nessun valore è condiviso da due studenti).

Perché non abbiamo eliminato anche *Età* e *Residenza*? Perché possiamo provare a trasformare tali attributi in modo tale da poter ridurre il fenomeno della frammentazione. Sappiamo infatti che i comuni possono essere accorpati per gruppi geografici: ad esempio per regione. Se applichiamo tale accorpamento sui nostri dati,

Eta′	*Residenza*	*Telematica*	*Matematica*	*Voto di laurea*
26	Calabria	Si	No	Alto
27	Calabria	Si	No	Medio
30	Lombardia	No	Si	Basso
24	Puglia	No	Si	Alto
27	Puglia	No	No	Alto
32	Lombardia	Si	No	Basso

arriviamo finalmente a notare qualcosa. In corrispondenza infatti del valore *Lombardia* compare il valore *Basso*: questa è una caratterizzazione di un concetto, che vale nella totalità dei casi che stiamo osservando. Siamo riusciti cioè ad *indurre* un concetto dall'osservazione di altre caratteristiche osservabili del concetto stesso. Potrò applicare questo concetto in situazioni future, riuscendo così a prendere decisioni che permettano di essere più incisivo con i miei dati.

Non riesco ancora a caratterizzare in maniera soddifacente il concetto che mi interessava: il voto *Alto*. Infatti, le regole che lo caratterizzano dovrebbero essere: "lo studente abita in *Puglia* e non ha sostenuto esami di *Telematica*", oppure "abita in *Calabria* e non ha sostenuto esami di *Matematica*" (ma in quest'ultimo caso sarei ancora troppo generico). Una situazione migliore invece si ottiene ancora una volta trasformando l'attributo *Età*: si può infatti osservare che studenti possono essere suddivisi in fascie d'età: giovanile (fino a 26 anni), tipica (da 27 a 29 anni), alta (dai 30 anni in su). Applicando questa osservazione ai nostri dati, otteniamo:

Eta′	*Residenza*	*Telematica*	*Matematica*	*Voto di laurea*
$(-\infty, 26]$	Calabria	Si	No	Alto
(26,30]	Calabria	Si	No	Medio
$(29,+\infty)$	Lombardia	No	Si	Basso
$(-\infty, 26]$	Puglia	No	Si	Alto
(26,30]	Puglia	No	No	Alto
$(29,+\infty)$	Lombardia	Si	No	Basso

Osservando i dati alla luce di quest'ultima trasformazione, siamo in grado con un buon margine di sicurezza di esprimere una caratterizzazione del concetto che ci interessa: uno studente ha un voto alto quando è giovane (ovvero, la sua età risiede nell'intervallo $(-\infty, 26]$) oppure non ha sostenuto esami di telematica (o, equivalentemente, risiede in *Puglia*).

La descrizione del processo che abbiamo visto in quest'esempio ci ha permesso di cogliere con mano due aspetti essenziali:

- il primo aspetto riguarda la *scoperta di conoscenza*, che è tipicamente un processo complesso. Tale processo si compone di numerose attività di

trasformazione e preparazione dei dati, per arrivare alla fase di estrazione *non banale, utile e significativa* dell'informazione che può essere valutata e applicata ai casi di interesse.

- il secondo aspetto riguarda più da vicino il task che abbiamo considerato. Abbiamo fatto *modellazione predittiva*, ovvero caratterizzazione di un insieme di concetti con l'obiettivo di poter sfruttare tale caratterizzazione a casi in cui i concetti non sono noti.

La caratterizzazione della scoperta di conoscenza in termini di processo complesso caratterizza anche altri task tipici, che possono mirare anche solo alla descrizione sommaria dei dati. Possiamo denotare tali tecniche con il nome di *modellazione descrittiva*. Vediamone due esempi.

Il primo esempio riguarda il seguente problema: vogliamo sapere quali correlazioni esistono tra gli esami, osservando il comportamento degli studenti. Ci interessa in pratica capire quali sono le preferenze degli studenti. Se proviamo a ristrutturare l'informazione originaria di figura 1.1, possiamo ottenere la seguente tabella.

ID	*SistemiDistribuiti*	*Programmazione a Oggetti*	*Statistica*	*Data Mining*	*Intelligenza Artificiale*	*Comunicazioni Elettriche*
1	Si	Si	No	No	No	No
2	Si	Si	No	No	No	No
3	No	No	Si	Si	No	No
4	No	No	Si	Si	Si	No
5	No	No	Si	Si	No	No
6	Si	No	No	No	No	Si

Si noti la differenza rispetto alla tabella precedentemente esaminata: qui le informazioni relative agli studenti sono ridotte al minimo, mentre sono riportate con un maggior livello di dettaglio le informazioni relative agli esami. Riformulato in termini di tale tabella, il task che ci interessa è il seguente: esistono gruppi di esami che (di norma) vengono sostenuti insieme? Più in dettaglio: quali sono le colonne che hanno le stesse corrispondenze di valori *Si* e *No*? Nella tabella si evince chiaramente che un tale insieme è composto dagli esami *Data Mining* e *Statistica*. Tali esami sono *correlati*: chi ha sostenuto l'esame relativo al primo, ha sostenuto l'esame anche per il secondo. Una correlazione vi è anche tra *Sistemi Distribuiti* e *Programmazione a Oggetti*, anche se con una misura di certezza minore: in quest'ultimo caso, infatti, non vi è una corrispondenza piena tra i valori delle tuple (i valori dell'ultima tupla non corrispondono).

Il secondo esempio riguarda una caratterizzazione comportamentale degli studenti: ci sono delle affinità di comportamento degne di nota? Più nello specifico: Quali sono i gruppi di studenti che si comportano allo stesso modo, e qual'è il comportamento che esibiscono? Proviamo a ricostruire i dati del primo esempio in maniera più arricchita:

ID	*Eta'*	*Telematica*	*Informatica*	*Matematica*
1	26	25	26	-
2	27	23	24	-
3	30	-	18	20
4	24	-	28.5	30
5	27	-	27	28
6	32	18	-	18

In questa tabella, abbiamo riportato la media dei voti ottenuti dagli studenti nei vari settori. Ad esempio, lo studente di *ID* 1 ha riportato i voti 25 a *Sistemi Distribuiti* e 26 *Programmazione a Oggetti*. Lo studente 4 ha riportato i voti 30 e 28 agli esami di *Data Mining* e *Intelligenza Artificiale* (con una media quindi di 28.5), e 29 all'esame di *Statistica*.

Riformulando il nostro problema, ci interessa capire se le righe della tabella di cui sopra possono raggruppate in gruppi in cui i valori riportati in ogni colonna siano *simili*. Qui, il concetto di similitudine è ambiguo e va espresso in maniera più rigorosa: sappiamo che uno studente che abbia 24 anni è più simile a un'altro che ne ha 27 che a uno che ne ha 32. Tuttavia, uno che ne ha 24 e ha effettuato solo esami informatici è più o meno vicino a uno che ne ha 27 e ha effettuato solo esami telematici, rispetto a uno di 32 che ha effettuato esami sia telematici che informatici? Ancora: uno studente che ha voti alti su esami informatici è più o meno vicino ad uno studente che ha voti alti su esami telematici, rispetto ad uno che ha voti bassi sia in esami telematici che in esami informatici? Se gli esami telematici hanno tutti voti bassi, uno studente che ha conseguito voti bassi in tali esami e voti alti negli esami informatici con chi va raggruppato?

Possiamo risolvere questa ambiguità cercando di *normalizzare* i dati: di riportarli, ovvero, in un unico range che permetta di confrontare valori di colonne distinte. Se ad esempio normalizziamo l'attributo *Età*, otteniamo

ID	Eta'^*	$Telematica^*$	$Informatica^*$	$Matematica^*$
1	0.5	25	26	-
2	0.625	23	24	-
3	0.75	-	18	20
4	0	-	28.5	30
5	0.625	-	27	28
6	1	18	-	18

Se dovessimo ignorare tutti gli altri attributi e raggruppare solo per età, otterremmo il seguente ordine nei dati

ID	Eta'^*	$Telematica^*$	$Informatica^*$	$Matematica^*$
4	0	-	28.5	30
1	0.5	25	26	-
2	0.625	23	24	-
5	0.625	-	27	28
3	0.75	-	18	20
6	1	18	-	18

In base a tale ordine, risulterebbe che gli studenti 1,2,5,3 formano un unico gruppo, mentre gli studenti 4 e 6 formano dei gruppi a parte. Normalizziamo anche tutti gli altri attributi:

ID	Eta'^*	$Telematica^*$	*Informatica*	*Matematica*
4	0	-	1	1
1	0.5	1	0.76	-
2	0.625	0.7	0.57	-
5	0.625	-	0.85	0.83
3	0.75	-	0	0.16
6	1	0	-	0

Notiamo adesso che l'ordine imposto dal solo attributo *Età* non va più bene. In particolare, lo studente 4 assomiglia moltissimo allo studente 6, mentre differisce da tutti gli altri. Analogamente, lo studente 1 assomiglia molto allo studente 2 ed entrambi differiscono dagli altri, ovvero esibiscono valori che non sono simili ai valori esibiti dagli altri. Infine, gli studenti 3 e 6 sono molto atipici rispetto agli altri: gli esami che hanno sostenuto sono eterogenei, e comunque hanno caratteristiche diverse. Riassumento, possiamo trovare 4 tendenze principali nei dati, riassunte dai raggruppamenti evidenziati nella seguente figura:

ID	*Eta'**	*Telematica**	*Informatica*	*Matematica*
1	0.5	1	0.76	-
2	0.625	0.7	0.57	-
4	0	-	1	1
5	0.625	-	0.85	0.83
3	0.75	-	0	0.16
6	1	0	-	0

Il primo gruppo descrive studenti interessati principalmente in *Telematica*: infatti, la caratteristica principale è il forte peso attribuito a *Telematica*, mentre un peso leggermente inferiore appare per *Informatica*. Per contro, il secondo gruppo è fortemente orientato a esami informatici e matematici. I restanti gruppi (che possono considerasi *outliers*) non hanno similarità forti con nessuno dei gruppi sopra considerati.

1.2 Data Mining e Knowledge Discovery

1.2.1 Il processo

1.2.2 Le applicazioni

1.2.3 Il ruolo del Data Warehousing