Reti neurali nel Data Mining, altre tecniche utilizzate nel DM e valutazione dei modelli.

Vincenzo Antonio Manganaro

vincenzomang@virgilio.it, www.statistica.too.it

Indice

Nella prima parte del presente articolo cercheremo di completare la descrizione di tecniche utilizzate nel DM. Tuttavia l'intero articolo presenta un certo grado di generalità, sia perchè lo studio approfondito di determinati strumenti non rientra negli obiettivi che ci siamo prefissi (vedi ad esempio uno studio approfondito di strumenti quali reti neurali di cui daremo solo i concetti essenziali nel primo paragrafo), sia per il motivo opposto che alcune tecniche, anch'esse usate nel DM, come la regressione lineare semplice e multipla , la regressione logistica, i modelli lineari generalizzati (GLM) e i modelli additivi generalizzati (GAM) risultano maggiormente familiari agli statistici . Riteniamo dunque utile citare queste tecniche senza entrare troppo in dettagli che sicuramente saranno noti anche ai meno attenti.
Nella seconda parte del capitolo vedremo infine come valutare i modelli ottenuti attraverso le varie fasi del DM.

1 Utilizzo di reti neurali nel DM.

Le reti neurali sono tecniche di analisi statistica che permettono di costruire dei modelli di comportamento a partire da un insieme di *esempi* (definiti attraverso una serie di variabili numeriche e categoriali) di questo comportamento. Una rete neurale, "ignorante" in una fase iniziale, attraverso un processo di "training" (apprendimento), si trasforma in un modello di dipendenze tra variabili descrittive così da prevederne il comportamento e come vedremo un pò meno a spiegarne il meccanismo sottostante. Per esempio attraverso i dati che descrivono individui che richiedono un prestito (età, reddito, sposati o no, ecc.) associati ai dati relativi ai ritardi nei pagamenti o a problemi insorti nei pagamenti, può essere costruito un modello che contiene le relazioni tra ognuna di queste variabili e il risultato del prestito (cattivo creditore o buon creditore) in modo tale da utilizzare il modello, rappresentato in tal caso dalla rete dopo il processo di training, come un classificatore per nuovi clienti che vogliono accedere ai prestiti.
Le reti neurali sono tipicamente organizzate in strati e gli strati sono costituiti da un numero di "nodi" interconnessi ciascuno dei quali contiene una "funzione di attivazione". Gli esempi, attraverso i quali il meccanismo della rete apprende i comportamenti, sono presentati attraverso i nodi dello strato di input che comunicano con uno o più stati interni nei quali viene eseguito il processo attraverso un sistema di connessioni pesate. Successivamente gli strati interni si legano ad uno strato di output dove vi è la risposta del processo effettuato dalla rete.
Molte reti neurali svolgono il loro processo di apprendimento da esempi attraverso dei meccanismi che modificano il sistema dei pesi delle connessioni degli strati interni in relazione ai patterns di input che vengono presentati alla rete. In tal modo il processo di apprendimento delle reti neurali artificiali avviene con le stesse modalità delle loro controparti biologiche e cioè alla stessa maniera di come un bambino riesce a riconoscere i cani attraverso esempi di cani.

Analizzeremo ora, per quanto ci è consentito in questa sede, la struttura di una semplice rete neurale nonchè il più diffuso meccanismo di training della rete, noto in letteratura come *backpropagation*. La figura 6.1 mostra una semplice rete neurale costituita da uno strato di input composto da due nodi, da un solo strato interno

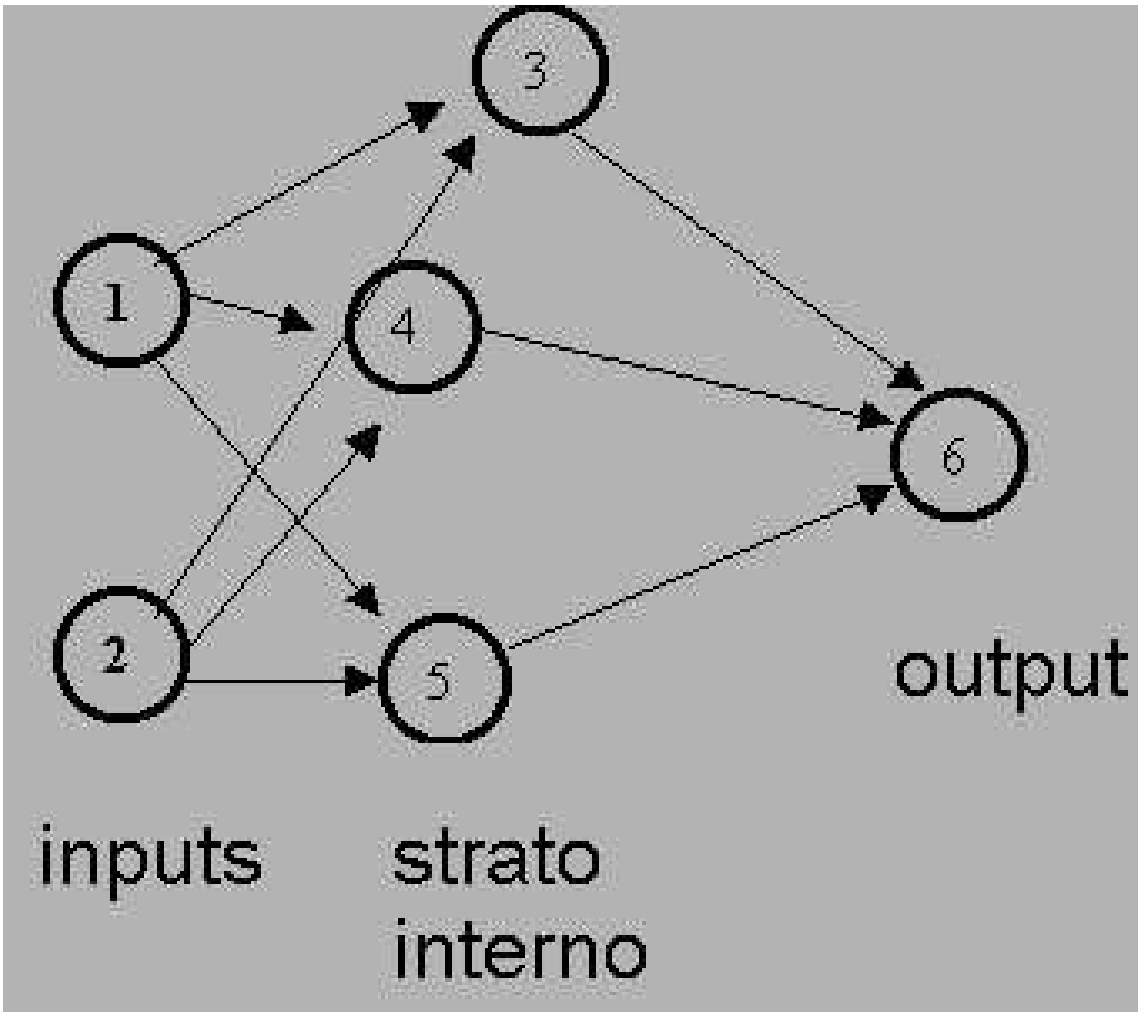

Figura 1: Una rete neurale con un solo strato interno.

composto da tre nodi e dallo strato di output composto da un singolo nodo[29]. Escludendo lo strato di input, ogni nodo contiene un set di input; questi inputs vengono moltiplicati con dei pesi di connessione W_{XY} (ad esempio il peso dal nodo 1 al nodo 3 è W_{13}) e, sommati i valori ottenuti, viene applicata a questi valori la funzione di attivazione associata a quel nodo e trasferito l'output ottenuto al nodo o ai nodi nel successivo strato. Per esempio il valore trasferito dal nodo 4 al nodo 6 è:
funzione di attivazione applicata a ([W_{14} * valore del nodo 1] + [W_{24} * valore del nodo 2]). La rete neurale con i pesi di connessione è mostrata nella figura 6.2.

Ogni nodo può essere visto come una variabile predittrice o come una combinazione di variabili predittrici. Inoltre una rete neurale è in qualche modo un meccanismo che nasce da una generalizzazione molto complessa della semplice regressione lineare che in taluni casi può ridursi proprio a quest'ultima. Si pensi ad esempio alla rete neurale in figura 6.3 ottenuta dalla precedente eliminando lo strato interno e considerando nel nodo 6 una funzione di attivazione lineare. In tale rete la variabile risposta costituita dal nodo 6 è una combinazione lineare dei valori nei nodi 1 e 2 per mezzo dei pesi W_{16} e W_{26} che si possono stimare attraverso la nota tecnica di regressione con un processo di minimazzione dell'errore abbastanza noto. Infatti i pesi di connessione in una generica rete neurale sono parametri sconosciuti che vengono stimati nel suddetto meccanismo di apprendimento.

Consideriamo ora il funzionamento dell'algoritmo di training della rete cui accennavamo e che consiste sostanzialmente nella individuazione di una buona stima per i pesi W_{XY}. Esso comprende le due fasi seguenti:

- *Feed forward*: nella quale il valore del nodo o dei nodi di output sono calcolati sulla base dei nodi di inputs e di un set iniziale di pesi. I valori dei nodi di inputs sono combinati nei nodi interni ed i valori di questi nodi sono combinati per calcolare il valore o i valori di outputs.

- *Backpropagation*: è la vera e propria fase di stima dei parametri[30]. In essa l'errore nell'output è calcolato trovando la differenza tra l'output ottenuto dalla rete e l'output desiderato[31](ovvero i valori effettivi del training set)[32]

[29]Si possono anche avere strati di output con due o più variabili risposta.
[30]L'algoritmo è appunto noto in letteratura con il nome di una sua fase.
[31]In statistica non è per nulla inusuale un tale procedimento di stima dei parametri.
[32]Vedremo successivamente come un modo per testare la validità di un modello consista nel

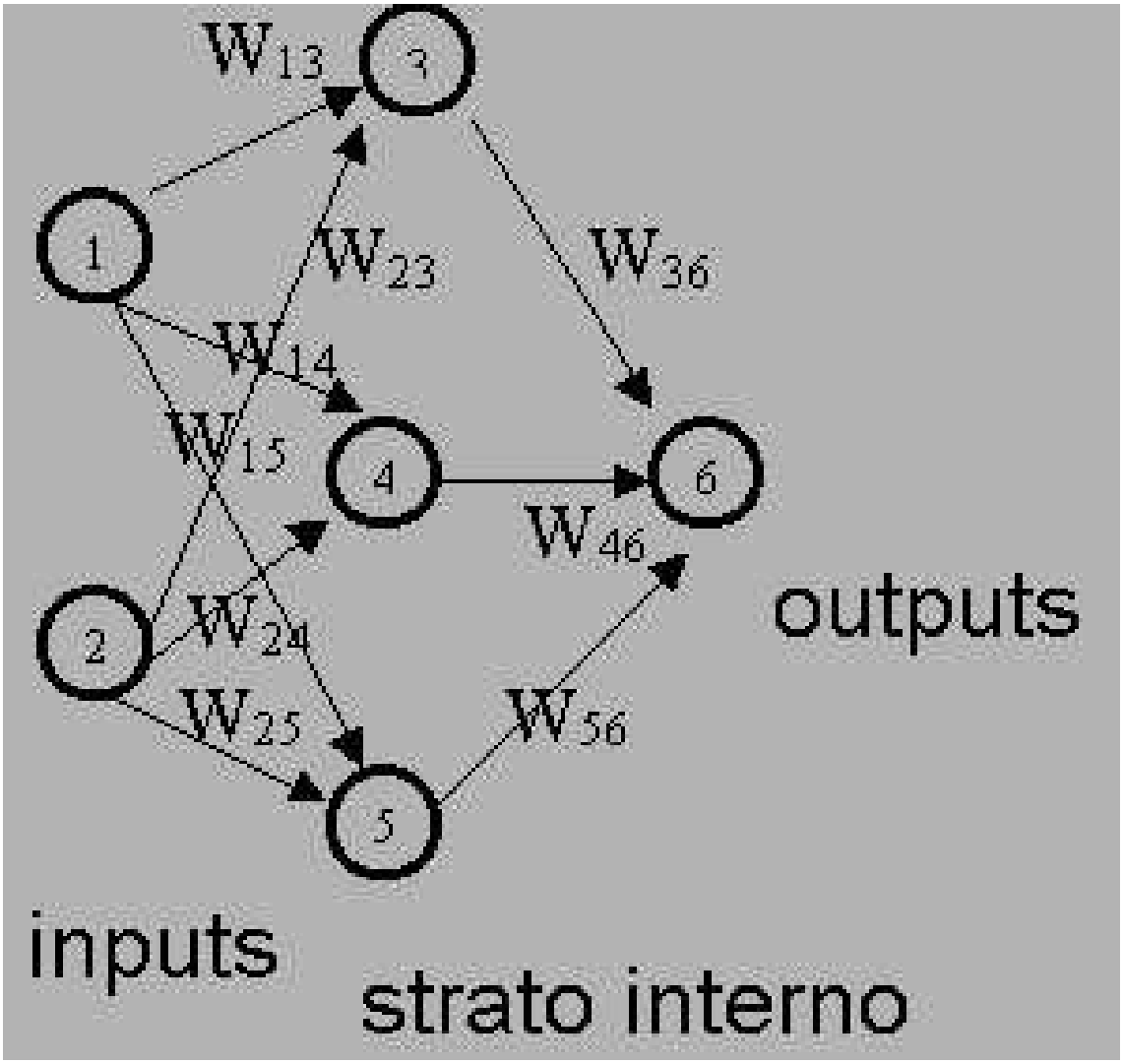

Figura 2: W_{XY} rappresenta il peso dal nodo X al nodo Y.

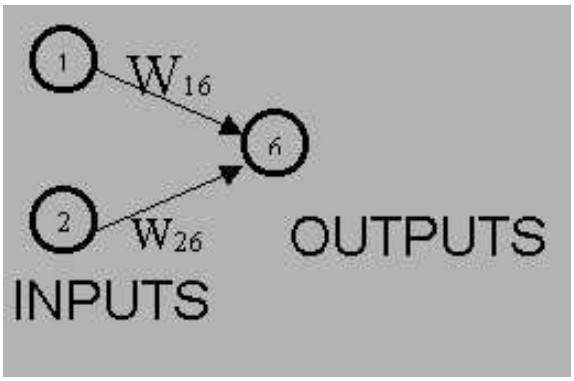

Figura 3: Rete neurale ottenuta dalla precedente con l'eliminazione dello strato interno.

Successivamente l'errore ottenuto nell'output è assegnato ad ogni nodo proporzionalmente ai loro pesi. Ciò consente di calcolare un errore per ogni nodo di output e per ogni nodo interno in modo tale che l'errore in ognuno dei nodi di output ed interni venga usato dall'algoritmo per aggiustare il peso in quel nodo allo scopo di ridurre l'errore totale.

Questo processo di apprendimento consistente in una progressiva diminuzione dell'errore totale della rete è ripetuto per ogni riga del training set. Il passo dell'algoritmo consistente nell'utilizzo di tutte le righe del training set è chiamato *epoca* e l'algoritmo effettuerà sul training set varie epoche ripetutamente finché l'errore della rete non diminuirà più.

Tuttavia l'eccessivo numero di nodi presente negli strati interni di una generica rete neurale con il conseguente elevato numero di parametri consente quasi sempre, e con un numero sufficientemente alto di epoche, un adattamento pressocchè completo della rete ai dati del training set. Non è però affatto vero che una rete che adatti perfettamente i dati del training set si comporti bene anche per altri datasets in generale e per il test set in particolare. Spesso è dunque opportuno avere una rete che non si comporti eccessivamente bene con i dati del training set purchè abbia un valore accettabile dell'errore nel test set. Il grafico in figura 6.4 è costruito valutando periodicamente l'andamento dell'errore sul test set (test set error) durante la successive epoche nella fase

suddividere i dati a disposizione in un training set (attraverso il quale costruire il modello) ed in un test set (attraverso il quale testare il modello costruito).

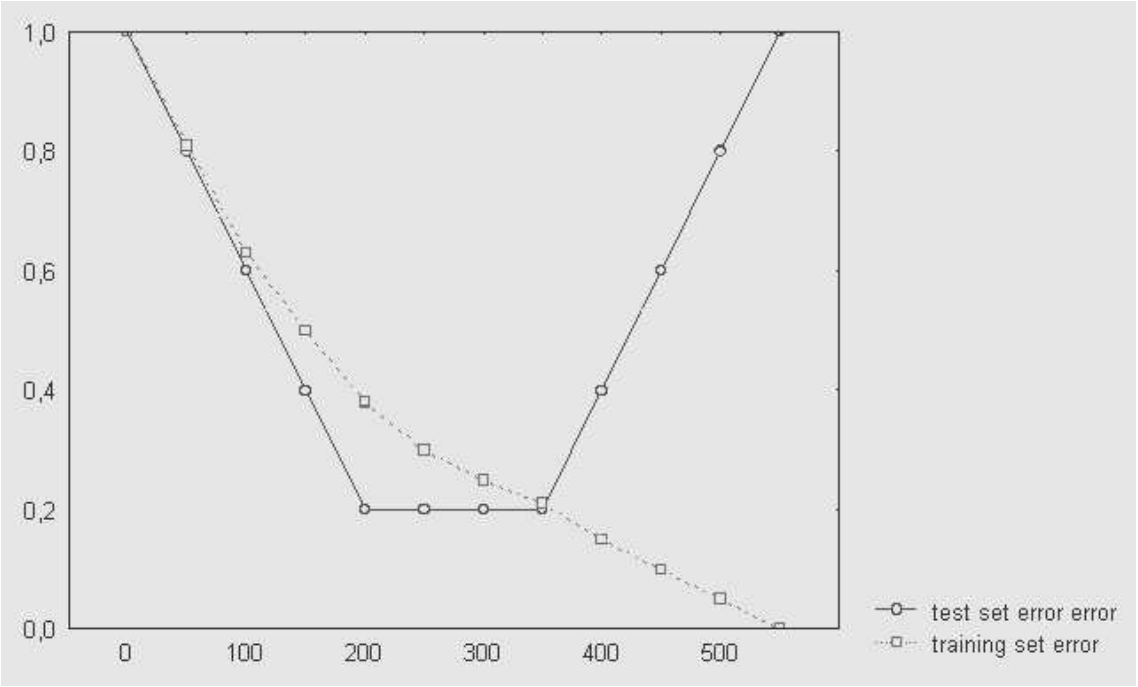

Figura 4: Andamento del training set error e del test set error in funzione del numero di epoche in una generica rete neurale.

di training. Come si vede dalla figura a fronte di una continua diminuzione dell'errore nel training set (training set error), errore che potremmo ipotizzare tendente a 0 con un numero sufficientemente elevato di epoche, l'errore sul test set prima decresce e poi, dopo un periodo di costanza, continua a crescere. Essendo l'obiettivo del DM quello di effettuare previsioni su altri sets di dati che non siano il training set può essere opportuno minimizzare l'errore sul test set che non quello sul training set.

Se si vuole quindi utilizzare nel DM una rete neurale per risolvere una qualsiasi tipologia di problema (vedi classificazione, ricerca di regole di associazione, ecc . . .) è opportuno in una prima fase progettare una specifica architettura di rete (cioè una rete costituita da un determinato numero di strati ciascuno dei quali caratterizzati da un determinato numero di nodi). Evidentemente non è in questa sede che si spiegherà come costruire una rete neurale; diremo solo che costruire una rete la cui dimensione e struttura sia adatta per lo studio di un determinato fenomeno richiede conoscenze su quel fenomeno che evidentemente non sempre si hanno in uno

stadio iniziale. Tuttavia è anche vero che esistono in commercio dei software molto complessi basati su tecniche di Intelligenza Artificiale che possono aiutare l'utente a costruire una buona architettura di rete per lo studio di un determinato problema. Una volta che una rete viene costruita, dopo la fase di training, è pronta per essere usata nelle previsioni. Essa rappresenta l'equivalente di un modello statistico ma esalta soprattutto gli aspetti pratici di un tale modello. Difatti, a differenza dei modelli statistici che spesso oltre all'importante e fondamentale aspetto pratico intendono descrivere la tipologia di relazioni tra variabili e il sottostante fenomeno , una rete neurale si dice che rappresenta un approccio di ricerca "non teorico", una sorta di scatola nera che può fornire si delle ottime previsioni con notevole importanza dal punto di vista applicatico ma che nulla può fornire circa la sottostante natura del fenomeno in questione.
Quest'ultima limitazione può però essere secondaria in molte applicazioni di DM: si pensi ad esempio ad una banca interessata a riconoscere i clienti ai quali possono essere concessi dei prestiti. Alla banca interessa avere delle buone previsioni sui potenziali nuovi clienti che intendono accedere ai prestiti e non certo il meccanismo che porta quel cliente ad essere un potenziale buon creditore o cattivo creditore. In ogni caso le reti neurali, nonostante le loro limitazioni quali ad esempio la non facile interpretabilità, il tempo richiesto nella fase di training, la possibilità di utilizzo per sole variabili di tipo numerico[33] e la preparazione dei dati che il loro utilizzo richiede, non minore rispetto agli altri metodi, si adattano bene agli obiettivi del DM e quindi agli obiettivi di ricerca di informazione per scopi previsivi.

2 Altre tecniche usate nel DM: una panoramica.

- **K-nearest neighbor and memory-based reasoning (MBR):**

Si tratta di una tecnica di classificazione che decide di assegnare un nuovo caso ad una determinata classe esaminando l'appartenenza alle classi dei k casi più vicini o più simili al caso da classificare. Il metodo conta il numero dei casi per ogni classe ed assegna il nuovo caso alla stessa classe cui appartiene la maggior parte dei vicini. La prima cosa da fare per applicare tale tecnica consiste nel trovare una misura

[33]Nel caso di variabili categoriali queste uiltime vengono trasformate in un numero opportuno di variabili numeriche dicotomiche avvenendo di fatto una *esplosione categoriale*.

della distanza tra casi che permette di trattare anche con variabili categoriali (vedi ad esempio la distanza nell'equazione 5.3). Una volta individuato il metodo per definire la distanza è possibile utilizzare i casi già classificati per classificare nuovi casi, decidere quanto grande deve essere il "vicinato" ovvero il set di casi vicini al caso da esaminare e decidere come contare gli stessi vicini (potremmo pensare ad esempio di dare maggior peso ai vicini più vicini rispetto a quelli più lontani). Tale tecnica (K-NN) richiede un grosso carico computazionale; a differenza del tempo richiesto per applicare un nuovo caso ad un albero di decisione o ad una rete neurale, tale metodo richiede che per ogni nuovo caso vengano rieseguiti i calcoli. I metodi di MBR (memory - based reasoning) si riferiscono all'utilizzo dei metodi K-NN dove ad esempio i risultati di un classificatore K-NN vengono tenuti in memoria allo scopo di velocizzare l'algoritmo.

- **Boosting:**

É una tecnica sperimentata da Freund e Schapire nel 1996 anch'essa nell'ambito dei problemi di classificazione. Essa consiste nello scegliere vari campioni casuali dai dati e costruire un modello di classificazione per ogni campione. L'idea che sta alla base delle tecnica è la seguente: se viene costruito un modello usando un campione di dati e successivamente usando lo stesso algoritmo viene costruito un altro modello ma su un altro campione, i modelli possono originare risultati contrastanti. Tuttavia dopo aver saggiato la validità dei due modelli sarà possibile scegliere quello che tra i due meglio rispecchia gli obiettivi prefissati. Risultati ancora migliori possono essere ottenuti se si costruiscono molti modelli su diversi campioni in modo tale che i risultati della classificazione per un oggetto siano costituiti dalla classe assegnata più spesso a quell'oggetto dai vari modelli.

- **Rule induction:**

Si tratta di un metodo per derivare un insieme di regole per la classificazione di casi. É diverso dai metodi di classificazione che prevedono l'uso di alberi di decisione poichè esso genera un set di regole indipendenti che non necessariamente generano un albero. Inoltre le regole generate possono non coprire tutte le possibili situazioni e possono qualche volta originare delle previsioni contrastanti fra di loro. É necessario

quindi in tali situazioni creare dei criteri razionali che mi permettano di scegliere tra due regole in conflitto.
Altre tecniche usate nel DM riguardano le metodologie standard di:

- **Regressione lineare semplice e multipla;**
- **Regressione logistica:** previsione di variabili binarie;
- **Analisi discriminante;**
- **Generalized Additive Models (GAM):** I GAM sono una classe di modelli che nascono da una generalizzazione sia della regressione lineare che della regressione logistica così chiamati in quanto si assume che ogni modello della famiglia può essere scritto come la somma di funzioni possibilmente non lineari ciascuna per ogni predittore.
- **Generalized linear models (GLM)**

3 Valutazione dei modelli.

Quando si costruisce un modello, sia esso derivante da una procedura di DM o un qualsiasi altro modello che intende spiegare un fenomeno a partire dai dati, bisogna valutare il modello e questa valutazione deve avvenire per due differenti aspetti inerenti la valutazione interna e la valutazione esterna del modello in questione:

- VALUTAZIONE INTERNA:

Potremmo senz'altro affermare che la valutazione interna sia la più semplice da eseguire; qualunque sia il modello costruito è in generale poco impegnativo costruire degli indici che misurino l'accuratezza del modello nel descrivere i dati. Nel caso del DM il processo di costruzione di modelli predittivi prevede di per sé l'adozione di un protocollo talvolta chiamato "apprendimento supervisionato" in grado di assicurare le più accurate e robuste previsioni. L'essenza di questo protocollo consiste nello stimare il modello su una porzione di dati a disposizione (*training set*) e successivamente saggiare e, se è opportuno, validare il modello sulla base della rimanente porzione di dati (*test set*). Un modello è effettivamente costruito quando il ciclo

di stima e di valutazione del modello è concluso con la validazione di quest'ultimo. Vediamo ora alcune tecniche che permettono di validare internamente un modello:

Simple validation: consiste nello scegliere in maniera casuale una percentuale dei dati -tipicamente fra il 5% ed il 33%- del database come test set. Dopo aver costruito il modello sul corpo principale dei dati, il modello viene utilizzato per predire le classi (nel caso di classificazione) o i valori (nel caso di regressione) del test set. Dividendo il numero di classificazioni corrette col totale delle classificazioni eseguite o col totale dei casi del test set si ottiene un tasso di accuratezza (cioè accuratezza= 1-errore)[34]. Ad esempio nel caso della regressione il coefficiente di determinazione R^2 abbastanza noto esprime la bontà di adattamento cioè in che misura una dipendenza lineare fra variabili può spiegare il fenomeno in questione.

Cross validation: la cross validation viene usata nel caso l'ammontare di dati che l'utilizzatore ha a disposizione è piuttosto modesto ed è assurdo pensare di restringere ulteriormente il dataset effettuando una scissione tra training set e test set. Dunque la cross validation permette di utilizzare tutto il database. Ciò è possibile suddividendo i dati in due sets perfettamente uguali allo scopo di stimare l'accuratezza della previsione del modello costruito su tutto il dataset. Viene quindi costruito un modello sul primo set di dati ed il modello costruito viene utilizzato per prevedere i risultati sul secondo set calcolando un tasso di errore e_1. Successivamente viene costruito un modello sul secondo set di dati e questo modello viene utilizzato per prevedere i risultati sul primo set calcolando un tasso di errore e_2. Abbiamo quindi a disposizione due stime indipendenti di errore, la cui media aritmetica fornisce una migliore stima della vera accuratezza del modello costruito su tutti i dati. Generalmente viene usata la versione più generale della cross validation, che consiste nello dividere in maniera casuale il dataset originale in n set uguali e disgiunti. Si costruiscono quindi gli n modelli che si ottengono isolando a turno ciascuno degli n sets precedentemente individuati e si calcola per ogni modello il tasso di errore che si ottiene testando quest'ultimo sul set escluso fino ad ottenere la sequenza $e_1, e_2, \ldots, e_n$ di n tassi di errore indipendenti. Analogamente al caso di $n = 2$ la media aritmetica degli n tassi di errore costituisce la migliore stima della vera accuratezza del modello costruito su tutti i dati.

[34] Abbiamo costruito un tasso di accuratezza del genere per verificare la bontà di classificazione per un albero di decisione.

Bootstrapping: anche in questa tecnica, utile per stimare l'errore di un modello, quest'ultimo viene costruito utilizzando l'intero dataset. Numerosi datasets chiamati "campioni bootstrap", di numerosità uguale, sono costruiti operando un campionamento con reimmissione dal dataset di partenza. In tal caso ciascun record può essere presente più di una volta nel campione bootstrap, ma ciò mi assicura l'indipendenza dei tassi di errore calcolati sui singoli campioni bootstrap. La media di tali tassi di errore (si possono individuare anche 1000 campioni bootstrap con 1000 tassi di errore) costituisce la stima finale dell'errore del modello costruito su tutti i dati.

- VALUTAZIONE ESTERNA:

- **Valore netto del modello:** un modello che sia internamente valido non può essere esteso immediatamente ad altri datasets. Infatti il tasso di accuratezza trovato per mezzo del test set è valido solo per il dataset con il quale è costruito il modello. Ciò implica che l'accuratezza è non necessariamente la giusta metrica per selezionare il modello migliore, ma è necessario avere ulteriori informazioni dul tipo di errori dei vari modelli, nonché dei costi associati ai modelli candidati.
Nel caso, ad esempio, di problemi di classificazione uno strumento noto in letteratura come "matrice di confusione" può essere utile per misurare le potenzialità di un modello. Associare ad ogni modello una matrice di confusione permette di scegliere il modello migliore che non necessariamente coincide con quello avente il tasso di accuratezza maggiore. Un esempio di matrice di confusione è qui rappresentato:

		C. di appartenenza	*C. di appartenenza*	*C. di appartenenza*
		Classe A	**Classe B**	**Classe C**
C. prevista	**Classe A**	45	2	3
C. prevista	**Classe B**	10	38	2
C. prevista	**Classe C**	4	6	40

Essa è una matrice di dimensione $k * k$, con k numero di classi nella quale sulle colonne si hanno il numero reale di records appartenenti a ciascuna classe e sulle righe il numero previsto di records appartenenti ad una data classe. In questo modo i valori presenti sulla diagonale principale sono quelli che rappresentano il numero di casi classificati correttamente dall'algoritmo, mentre ogni valore fuori dalla dia-

gonale principale rappresenta un errore di classificazione. Ad esempio, la cella (2 1) della matrice ci dice che vi sono 10 casi della classe A classificati non correttamente nella classe B. Dalla matrice si può egualmente ricavare il tasso di accuratezza, dato dal rapporto della somma degli elementi della diagonale principale e dl totale delle celle (123/150= 82%) ma non solo, essa è molto più informativa del semplice tasso di accuratezza; si può vedere ad esempio che il modello riesce a prevedere correttamente 38 dei 46 casi appartenenti alla classe B, classificando i rimanenti 8 non correttamente rispettivamente 2 nella classe A e sei nella classe C. Tuttavia l'utilità di questa forma di rappresentazione sta nella possibilità di associare a questa matrice un'analoga matrice non necessariamente simmetrica (ricordiamo ancora una volta che classificare un oggetto in B invece che in A può avere un costo diverso dell'errore opposto). In tal modo nel caso di differenti costi associati con differenti errori, un modello con una bassa accuratezza può essere preferito ad uno con un'alta accuratezza, ma anche con un alto valore del costo dovuto ai tipi di errori che esso compie nella classificazione. Per capire ciò consideriamo oltre a quella già vista una ulteriore matrice di confusione:

		C. di appartenenza	*C. di appartenenza*	*C. di appartenenza*
		Classe A	**Classe B**	**Classe C**
C. prevista	**Classe A**	40	12	10
C. prevista	**Classe B**	6	38	1
C. prevista	**Classe C**	2	1	40

A queste matrici associamo la medesima matrice di costo:

10	5	5
10	10	20
20	20	10

la quale attribuisce un costo pari a 5 per un errore di classificazione nella classe A, un costo pari a 10 per la classe B ed a 20 per la C. Inoltre i valori della diagonale principale rappresentano, come è ovvio, i vantaggi di una giusta classificazione piuttosto che dei costi. Il valore netto del modello associato alla prima matrice è:
(123*10)-(5*5)-(12*10)-(10*20)=885

mentre quello del secondo modello associato alla seconda matrice è:
(118*10)-(22*5)-(7*10)-(3*20)=940.
nonostante la seconda matrice abbia un tasso di accuratezza pari a 118/150=79%<82% della prima matrice. Dunque per massimizzare il valore del modello si sceglierà il modello meno accurato, ma che presenta un più alto valore netto (totale vantaggi - totale costi).

- **Costi del modello:** anche se un modello risulta valido sia internamente che esternamente non è detto che esso venga subito adottato. É infatti necessario, visti i costi del DM, tenere sotto controllo il rapporto profitti/costi ovvero un indicatore conosciuto in contabilità con il nome di ROI (return on investment).

- **Validità del modello nel mondo reale:** non vi è alcuna garanzia che un modello, per quanto accurato possa essere e per quanto garantisca un alto valore netto ed un valore accettabile del ROI, rifletta il comportamento del mondo reale. I modelli in generale sono semplificazioni della realtà che comportano spesso e volentieri delle assunzioni forti difficilmente realizzabili; ad esempio costruire un modello che predice la propensione marginale al consumo degli individui spesso prescinde dall'inserire come variabile predittore anche il tasso di inflazione, ma sappiamo benissimo che brusche variazioni di quest'ultimo influenzeranno il comportamento dei consumatori. Anche i dati usati per costruire i modelli possono portare ad un modello che non rispecchia per nulla la realtà osservata. Dunque un modello valido non è necessariamente un modello corretto ed è soprattutto in tal senso che si tenta di dare una validità esterna al modello, provando continuamente il modello nel mondo reale.

-**Monitoraggio continuo del modello:** se un modello presenta tutte la caratteristiche che gli assicurano la possibilità di adottarlo per fini di business da parte di un'organizzazione, ciò non implica che il modello sarà sempre valido. Spesso all'interno di organizzazioni commerciali nascono problemi dovuti alle continue adozioni di modelli ritenuti validi. Bisogna quindi monitorare continuamente le prestazioni di un modello, ritestarlo, riprovarlo e possibilmente ricostruirlo, sapendo benissimo che, qualunque sia il sistema al mondo che il modello cerca di ricostruire, presenta un certo grado di evoluzione che, in poco tempo può eliminare completamente la validità di tale modello ritenuto per certi versi "immortale".

4 Considerazioni finali e aspetti critici.

Abbiamo più volte accennato ad una significativa distinzione tra il DM e gli altri strumenti analitici (si veda a tal proposito capitolo 1 paragrafo 1.4) consistente in un differente approccio che le due tecniche utilizzano nell'esplorare le relazioni nei dati. Molti degli strumenti analitici disponibili seguono un approccio basato sulla verifica, in cui l'utilizzatore ipotizza specifiche relazioni nei dati ed usa gli strumenti per verificare o rifiutare quelle ipotesi[35]. Questo approccio si basa sull'intuizione che ha l'analista di porre le domande iniziali e di impostare l'analisi sulla base dei risultati che si ottengono ponendo le domande al database. Tuttavia un simile modo di operare, molto diffuso in statistica[36], presenta dei limiti tra i quali l'abilità dell'analista nel porre al database le domande appropriate, la velocità con cui si ottengono i risultati e le difficoltà nel gestire la complessità dello spazio degli attributi, soprattutto nel caso di grossi database. Gli ultimi due limiti diventano basilari soprattutto in un contesto aziendale caratterizzato da un' estrema velocità nei cambiamenti (che rende sempre più difficile al management effettuare operazioni di just in time) e da database sempre più grandi nelle dimensioni. Il DM "puro" ovvero quello di tipo previsivo, in contrasto con i tradizionali metodi analitici, usa un approccio basato sulla scoperta, in cui diversi algoritmi (alcuni dei quali studiati in questa tesi) sono appositamente costruiti per determinare velocemente le relazioni chiave tra la variabili in grossi database.

Tuttavia questa enorme fiducia nelle tecniche di DM, o quanto meno nella filosofia del DM, è piuttosto esagerata: è vero che il DM può dare molto alle organizzazioni che ne fanno uso (a tal proposito nel capitolo seguente vedremo in che modo

[35] Anche in tal senso potremmo parlare di DM ed in realtà lo abbiamo già fatto considerando un approccio al DM di tipo top down. In generale qualunque procedura che porti all'estrazione di conoscenze utili da un insieme di dati o da un database aziendale può essere nel complesso considerata un'operazione di DM, indipendentemente dalla filosofia delle tecniche usate. Inoltre non è per nulla strano che software normalissimi, nati per la statistica, vengano usati in contesti di DM.

[36] Comunque si imposti il problema, un tale modo di procedere conduce sempre ad un problema di verifica di ipotesi nel quale si individua un parametro od un vettore di parametri μ, una distribuzione di probabilità uni o multivariata e sulla base dei dati a disposizione ed attraverso delle opportune statistiche test, funzioni dei dati, si saggiano sistemi di ipotesi abbastanza noti in statistica.

il DM può assistere il management aziendale), ma è anche vero che le operazioni di pre-analisi (segmentazione, pulitura dei dati, ...) e di post-analisi (valutazione dei modelli) che il DM necessita, possono più volte far sorgere il dubbio di quanto effettivamente convenga usare tecniche di DM al posto delle tradizionali tecniche di analisi dei dati. E' necessario dunque che l'eccessivo entusiasmo sulle effettive capacità del DM, nonostante sia comprensibile visto anche la novità delle tecniche che esso solitamente impiega, venga in qualche modo smorzato[37]. Per cui le seguenti affermazioni, qualcuna in realtà già accennata, costituiscono da una parte delle critiche al DM, mentre dall'altra servono ovviamente a dare un'idea delle giuste potenzialità di esso:

- il DM contribuisce a migliorare un'organizzazione, già in precedenza di successo, fornendo piccole ed utili indicazioni che sommate tra di loro nel contesto di più anni possono dare degli enormi vantaggi competitivi. Esso non produce, come alcuni sostengono, risultati sorprendenti tali da trasformare totalmente un'attività commerciale;

- il DM non è così sofisticato da poter sostituire la conoscenza e l'esperienza di manager ed analisti di mercato. Piuttosto esso si arricchisce del contributo di questi conoscitori dei sistemi economici ed assume un ruolo complementare anziché alternativo;

- gli strumenti di DM non trovano automaticamente (così è addirittura riportato in talune definizioni di DM) i modelli utili per i fini preposti, ma devono essere costantemente diretti agli scopi specifici che l'utilizzatore si propone. In altre parole l'utilizzatore non è per nulla assente nel processo di scoperta di conoscenza, ma al contrario interagisce continuamente con lo strumento di DM indirizzandolo all'obiettivo prefissato;

- l'utilizzo di metodologie di DM, visto anche i costi che essi comportano, necessita, più di altre tecniche, di una continua visione della forbice costi/ricavi. Che senso ha, in un'ottica aziendale e di salvaguardia di un bilancio, utilizzare tecniche di DM il cui costo non è adeguatamente supportato dai ricavi?

[37] In quest'ottica faremo riferimento ad un articolo il cui titolo "Debunking Data Mining myths", ovvero Ridimensionare i miti del DM, si commenta da sé.

- Le tecniche di DM sono spesso usate sfruttando i database che il più delle volte rappresentano dei campioni sul totale della popolazione. Tali campioni possono non essere rappresentativi della popolazione, sia per motivi probabilistici (cioè i campioni non sono campioni probabilistici ma si tratta di dati che possiede l'organizzazione e dunque riguardanti i clienti di quell'organizzazione e non tutti i potenziali clienti), sia per motivi che potremmo definire storici (e cioè le condizioni che hanno portato a quei dati non sono più valide). Per cui sarebbe assolutamente errato costruire un modello di DM basato su quei campioni. Ed anche quando si ha a disposizione l'intero database della popolazione, e quindi assenza di problemi di campionamento, si preferisce non usare l'intero database ma operare una divisione dei dati in modo da effettuare una sorta di valutazione interna del modello costruito su parte dei dati del database.

In ogni caso il DM non è una moda del momento, né il ritorno alle tecniche standard di analisi dei dati, che qualche critico del DM ipotizza, sembra probabile. La realtà è che nessuno può negare a priori la validità delle tecniche di DM; ma è opportuno individuare in maniera precisa le sue potenzialità e mai affidarsi ad esso ad "occhi chiusi". É anche vero che le organizzazioni commerciali che si affideranno ad esso, non tenendo conto dei limiti su accennati e di tanti altri che abbiamo trascurato, si troveranno presto in un serio svantaggio competitivo nei confronti di quelle organizzazioni che si affideranno ad un approccio più misurato, anche con una non piena fiducia, alle tecniche di DM.