

Algoritmi per il Data mining

con applicazioni aziendali

Andrea Brunello ¹ Enrico Marzano ²

¹Ph.D. student at Università degli Studi di Udine

²CIO at Gap srl



Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

- ▶ Il processo di Data Mining.
- ▶ Algoritmi per il learning non supervisionato:
 - ▶ regole di associazione;
 - ▶ tecniche di clusterizzazione;
 - ▶ *caso pratico*: analisi delle note degli agenti.
- ▶ Algoritmi per il learning supervisionato:
 - ▶ problemi di classificazione e di regressione;
 - ▶ regressione lineare;
 - ▶ regressione logistica;
 - ▶ *caso pratico*: stima della propensione al sondaggio;
 - ▶ alberi di decisione;
 - ▶ *caso pratico*: classificazione delle note degli agenti.

- ▶ **Dati** \approx **fatti** memorizzati, registrati;
- ▶ L'**informazione** è costituita dall'insieme dei **concetti**, delle regolarità, degli schemi che si trovano "nascosti" fra i dati;
- ▶ Il **Data Mining** si occupa dell'**estrazione** e della presentazione di **informazione** utile, precedentemente sconosciuta, ed implicitamente contenuta in una (grande) mole di dati.
 - ▶ E' un processo di astrazione (generazione di un modello in grado di catturare tali regolarità).
- ▶ Gli algoritmi di **Machine Learning** costituiscono una "base tecnica" per il Data Mining.

- ▶ Utilizzare i modelli/pattern appresi per:
 - ▶ **conoscere**: comprendere che determinate fasce di popolazione sono più propense ad acquistare un determinato bene;
 - ▶ **inferire**: probabile guasto ad un macchinario, da un insieme di sintomi;
 - ▶ **predire**: stabilire se e quale variazione nelle vendite risulterà da un aumento del budget pubblicitario.
- ▶ Tali fini possono mescolarsi, si pensi alla ricerca di un modello che fornisca la valutazione di un'abitazione sulla base di diversi valori in input.

Riassumendo:

- ▶ il Data Mining sfrutta tecniche di Machine Learning
- ▶ per estrarre semi-automaticamente
- ▶ da (grandi) quantità di dati
- ▶ informazioni, pattern utili

Input del processo:

- ▶ insieme di istanze, tuple (esempi dei concetti che si vogliono apprendere)

Output del processo:

- ▶ modelli
- ▶ predizioni/classificazioni

Il processo di Data Mining

In genere, il processo di Data Mining si articola come segue:

- ▶ il tutto inizia con il porsi una **domanda**, ben chiara e specifica;
- ▶ in seguito, si passa alla raccolta dei **dati** da utilizzare come input;
- ▶ viene selezionato un insieme di **caratteristiche** (features) ritenute importanti su tali dati, e per il fine che si vuole ottenere;
- ▶ si applica un **algoritmo** di machine learning sul dataset così definito, in modo da “addestrare” un modello;
- ▶ dopo un’eventuale fase di tuning, il modello prodotto dall’algoritmo viene **valutato**, ed è infine pronto per essere utilizzato.

Un primo esempio: SPAM vs HAM

Ripercorriamo ora il processo di Data Mining con un esempio focalizzato sulla predizione.

- ▶ **Domanda:** è possibile distinguere automaticamente fra i messaggi email che sono indesiderati (SPAM) e quelli legittimi (HAM)?
- ▶ **Dati:** insieme di 4601 istanze di email già classificate, ed ognuna avente 57 caratteristiche indicanti la frequenza di determinate parole e caratteri nel corpo del messaggio;
 - ▶ *www.inside-r.org/packages/cran/kernlab/docs/spam*

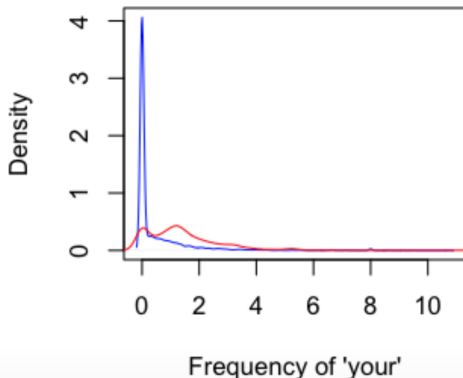
Un primo esempio: SPAM vs HAM / 2

Data mining

Brunello, Marzano

Selezione delle **caratteristiche**:

- ▶ processi di *attribute selection*;
- ▶ esplorazione e selezione manuale.



Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica

Stima della propensione

Alberi di decisione

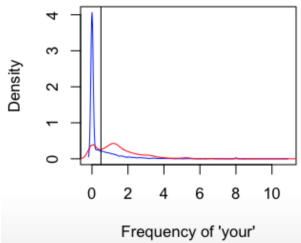
Classificazione delle note

Conclusioni

Un primo esempio: SPAM vs HAM / 3

Modello: utilizzo di un valore di **cutoff**

- ▶ se frequenza della parola "your" nel testo > 0.5 , allora l'email è SPAM



- ▶ Otteniamo la **tabella di contingenza** (valori su insieme di training):

<i>Predizione \ Classe</i>	nonspam	spam
nonspam	0.4590	0.10017
spam	0.1469	0.2923

Supervised vs Unsupervised Learning

Data mining

Brunello, Marzano

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Distinzione fondamentale, a seconda dell'obiettivo del processo:

- ▶ **Supervised Learning:** all'algoritmo di learning viene fornito un risultato noto per ciascuna istanza di training e si punta a determinare il valore per nuove istanze (attributo obiettivo).
 - ▶ *Classificazione con alberi, regressione lineare, ...*
- ▶ **Unsupervised Learning:** non si cerca di prevedere il valore di uno specifico attributo, ma viene ricercata ogni possibile associazione/correlazione fra gli attributi.
 - ▶ *Clustering, regole di associazione, ...*

- ▶ Si vogliono mettere in luce generiche relazioni fra gli attributi.
- ▶ Ciascuna regola opera in maniera indipendente.

Temperatura = fredda \rightarrow Umidità = normale

Umidità = alta \wedge Vento = no \rightarrow Pioggia = sì

Vento = sì \wedge Umidità = bassa \rightarrow Condizioni = soleggiato

 R. Agrawal, R. Srikant: *Fast algorithms for mining association rules*, 1994.

Principale metodologia di *Unsupervised Learning*:

- ▶ Cerchiamo di raggruppare fra loro istanze simili, senza considerare un determinato attributo come obiettivo.
- ▶ I gruppi individuati possono essere:
 - ▶ *esclusivi (hard clustering)*
 - ▶ *con overlap (fuzzy clustering)*
 - ▶ *probabilistici (fuzzy clustering)*
 - ▶ *gerarchici*
- ▶ Il processo di clustering può essere seguito da uno di classificazione:
 - ▶ come “conferma” dei cluster;
 - ▶ per assegnare ai gruppi nuove istanze.

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Diverse famiglie di algoritmi di clustering:

- ▶ gerarchici;
- ▶ basati su centroidi: *k-means(++)*, *k-medoids*, *FCM*
- ▶ basati sulla distribuzione: *Expectation-Maximization*
- ▶ basati sulla densità: *DBSCAN*, *OPTICS*
- ▶ ...

La scelta della tipologia dell'algoritmo riveste grande importanza per quanto riguarda il risultato finale, e dipende anche dalla distribuzione delle istanze.

- ▶ Ciascun cluster viene rappresentato da un *centroide*, elemento che può o meno appartenere al dataset iniziale;
- ▶ per gli algoritmi della famiglia *k-means*, il numero di cluster da ricercare deve essere fissato a priori;
- ▶ si cercano k centri di altrettanti cluster, e si assegnano ad essi le istanze, in modo tale da minimizzare la somma dei quadrati delle distanze;
- ▶ problema *NP-hard*, si cercano soluzioni approssimate.

Uno dei più diffusi algoritmi per il clustering *partizionale*:

1. seleziona il numero di cluster da ricercare, k ;
2. scegli casualmente k istanze, centri di altrettanti cluster;
3. assegna tutte le istanze-non-centro al punto, fra i k centri, più vicino;
4. calcola sulle istanze di ciascun cluster la media dei diversi attributi, e poni il punto così ottenuto come nuovo centro del cluster;
5. se, alla luce dei nuovi centri, almeno un'istanza cambierebbe cluster di appartenenza, ripeti dal punto 3, altrimenti la situazione è stabile e l'algoritmo termina.

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica

Stima della propensione

Alberi di decisione

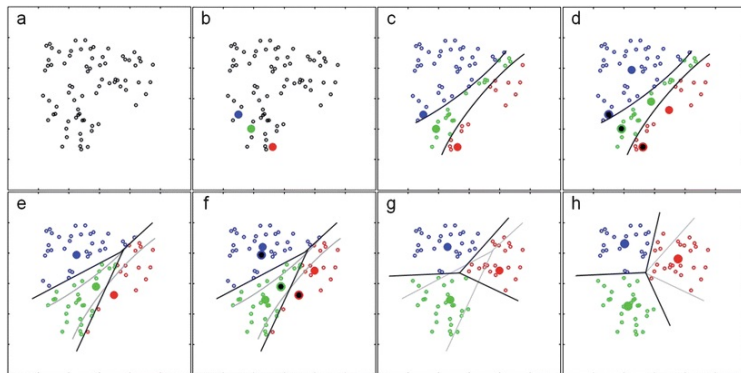
Classificazione delle note

Conclusioni

L'algoritmo *k-means* / 2

Data mining

Brunello, Marzano



- ▶ Generalmente il numero di iterazioni è minore del numero di punti;
- ▶ nel caso pessimo, su insiemi di punti costruiti appositamente, ha complessità esponenziale.

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

- ▶ Si cerca di minimizzare l'errore quadratico medio, la distanza tra punti di un cluster ed il punto designato per esserne il centro.
- ▶ L'algoritmo non garantisce l'ottimalità globale del raggruppamento.
- ▶ Aspetti riguardanti la normalizzazione degli attributi.
- ▶ Variante *k-means++*:
 - ▶ tecnica per l'inizializzazione dei centri dei cluster;
 - ▶ migliore per velocità ed accuratezza;
- ▶ Variante *k-medoids*:
 - ▶ il centro del cluster è un elemento del dataset;
 - ▶ maggiore robustezza al rumore e agli outlier.

Clustering basato sulla distribuzione

Data mining

Brunello, Marzano

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

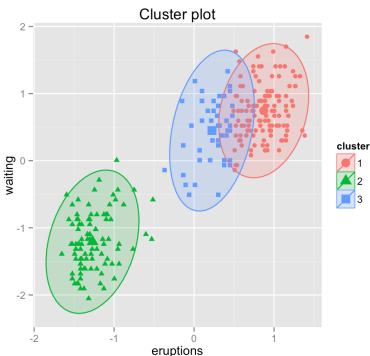
Regresione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

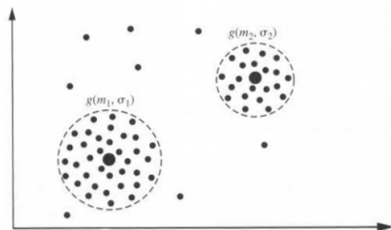
Conclusioni



- ▶ I cluster hanno un centro (es. quadrato blu).
- ▶ Ciascuno ha anche una distribuzione di probabilità, che indica la probabilità di un punto di appartenere al cluster (mixture model).
- ▶ Lungo la curva delle ellissi si ha la stessa probabilità di appartenenza.
- ▶ Generalizzazione dell'approccio *k-means*, importanza della distanza dal centro dipende dalla direzione.

Clustering basato sulla distribuzione / 2

Esempio di un mixture model. Sono presenti due cluster, ciascuno dei quali ha associata una distribuzione Gaussiana con relativa media e deviazione standard:



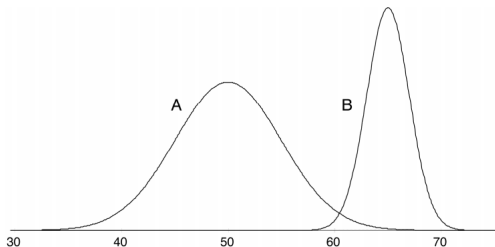
Obiettivo: partendo dall'insieme di istanze, trovare le distribuzioni corrispondenti ai cluster, più le probabilità marginali di appartenenza delle istanze ad essi (modello).

Ipotesi di fondo: le istanze del dataset sono generabili a partire da una *mixture* di modelli probabilistici.

Clustering basato sulla distribuzione / 3

Esempio con singola variabile numerica e due distribuzioni gaussiane. Dati etichettati dalle “classi” reali (per comodità, sopra) e modello (sotto):

A	51	B	62	B	64	A	48	A	39	A	51
A	43	A	47	A	51	B	64	B	62	A	48
B	62	A	52	A	52	A	51	B	64	B	64
B	64	B	64	B	62	B	63	A	52	A	42
A	45	A	51	A	49	A	43	B	63	A	48
A	42	B	65	A	48	B	65	B	64	A	41
A	46	A	48	B	62	B	66	A	48		
A	45	A	49	A	43	B	65	B	64		
A	45	A	46	A	40	A	46	A	48		
model											



- ▶ Il modello è dato da due cluster A, B con medie $\mu_A = 50, \mu_B = 65$ e deviazioni standard $\sigma_A = 5, \sigma_B = 2$;
- ▶ la probabilità marginale che ha un'istanza di appartenere ad A è pari a $P(A) = \frac{32}{50} = 0.64$, dunque $P(B) = 1 - P(A)$.

⇒ Date le istanze (senza alcuna conoscenza di A e B), vogliamo trovare i 5 parametri: $\mu_A, \mu_B, \sigma_A, \sigma_B, P(A)$.

Problema: non conosciamo né i 5 parametri, né l'effettiva appartenenza delle istanze ai cluster.

- ▶ *EM* è un algoritmo di raffinamento iterativo che può essere usato per trovare le stime dei parametri;
- ▶ può essere visto come un'estensione di *k – means* operante *fuzzy clustering*;
- ▶ esistono varianti in cui il numero di cluster viene determinato automaticamente;
- ▶ non viene garantita l'ottimalità del risultato.

Funzionamento:

1. imposta casualmente i valori di: $\mu_A, \mu_B, \sigma_A, \sigma_B, P(A)$;
2. itera:
 - 2.1 **Expectation step** - calcola la (densità di) probabilità di appartenenza ai cluster A, B per ogni istanza x_i , cioè $P(x_i \in A)$ e $P(x_i \in B)$, a partire dai parametri (assegnamento “soft” delle istanze ai cluster);
 - 2.2 **Maximization step** - stima, in maniera pesata in base alle probabilità di appartenenza delle istanze ai cluster, i parametri.

⇒ dunque, ad ogni passo *E-M*: riassegna le istanze tenendo conto del modello dato dai parametri (*expectation*); gli oggetti assegnati vengono quindi usati per generare nuove stime dei parametri (*maximization*).

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Possiamo calcolare $P(x_i \in A)$ a partire dai parametri:

$$P(x_i \in A) = P(A|x_i) = \frac{P(x_i|A) * P(A)}{P(x_i)}$$

dove:

$$P(x_i|A) = \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{(x_i - \mu_A)^2}{2\sigma_A^2}}$$

ossia la funzione di densità di probabilità nel caso della distribuzione normale.

Non conosciamo $P(x_i)$ (*probabilità di un'istanza dati i cluster*); tuttavia, $P(A|x_i) + P(B|x_i) = 1$, dunque:

$$\frac{P(x_i|A) * P(A) + P(x_i|B) * P(B)}{P(x_i)} = \frac{P(x_i \wedge A) + P(x_i \wedge B)}{P(x_i)} = 1$$

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

I parametri possono essere invece ricalcolati a partire dalle (densità di) probabilità come segue:

$$\mu_a = \frac{\sum_i P(x_i \in A) * x_i}{\sum_i P(x_i \in A)}$$

$$\sigma_a = \sqrt{\frac{\sum_i P(x_i \in A) * (x_i - \mu_A)^2}{\sum_i P(x_i \in A)}}$$

$$P(A) = \frac{\sum_i P(x_i \in A)}{\text{numero_totale_istanze}}$$

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Quando terminare?

- ▶ *EM* converge verso un punto fisso;
- ▶ concetto di verosimiglianza globale (*likelihood*), calcolata ad ogni iterazione:
 - ▶ misura della bontà del clustering;
 - ▶ aumenta ad ogni iterazione, fino ad un massimo locale;
 - ▶ intuitivamente, quanto è verosimile che il modello descriva/generi il dataset originario.
- ▶ l'algoritmo si arresta quando la *likelihood* rimane invariata (o varia entro una determinata threshold).

Come calcolare la *likelihood*?

$$\mathcal{L} = \prod_i P(x_i) = \prod_i (P(x_i|A) * P(A) + P(x_i|B) * P(B))$$

In pratica si calcola il logaritmo della verosimiglianza.

↪ la *likelihood* può essere anche utilizzata per confrontare la bontà di diversi risultati di clustering.

Caso pratico: analisi delle note degli agenti

- ▶ La valutazione degli agenti riveste grande importanza nell'ambito contact center;
- ▶ Gap srl utilizza una funzione analitica in grado di sintetizzare un punteggio $\in [0, 1]$ per ogni coppia agente/servizio;
- ▶ tale informazione *quantitativa* è complementata da una serie di analisi *qualitative*;
- ▶ fra esse, la valutazione delle note redatte dagli agenti nel corso delle sessioni inbound.

⇒ Come analizzare, valutare le note scritte?

Data mining

Brunello, Marzano

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Fase 1: estrazione delle feature dal testo

Per ogni nota, calcoliamo (script *R*, libreria *openNLP*):

- ▶ numero di caratteri / parole;
- ▶ indice *Gulpease* di leggibilità;
- ▶ articoli su parole;
- ▶ congiunzioni su parole;
- ▶ verbi su parole;
- ▶ sostantivi su parole;
- ▶ aggettivi su parole;
- ▶ avverbi su parole;
- ▶ preposizioni su parole;
- ▶ quantificatori su parole;
- ▶ pronomi su parole;
- ▶ codici numerici su parole;
- ▶ nomi propri su parole;
- ▶ fraz. delle parole / abbr. nel dizionario italiano;
- ▶ fraz. delle parole nel dizionario di dominio;
- ▶ fraz. delle parole non riconosciute.

Fase 2: gruppi di note simili

- ▶ Selezione casuale di 1000 note (su ≈ 400.000);
- ▶ applicazione dell'algoritmo *E-M*;
- ▶ sono emersi 6 cluster:
 - ▶ note articolate, generiche, ben scritte;
 - ▶ note non articolate, generiche, ben scritte;
 - ▶ note abbreviate;
 - ▶ note con linguaggio del dominio;
 - ▶ note nonsense;
 - ▶ note con caratteristiche ibride.
- ▶ [dettaglio output di *E-M*]

Problemi di regressione e di classificazione

Nel *Supervised Learning*, a seconda della tipologia dell'attributo obiettivo, distinguiamo problemi di

- ▶ **classificazione**: si vuole assegnare a ciascuna istanza uno di un insieme finito di valori;
- ▶ **regressione**: si vuole assegnare a ciascuna istanza un valore numerico.

Alcune famiglie di algoritmi di learning si adattano ad entrambi i problemi (alberi di classificazione e di regressione, regressione lineare e logistica, ...).

Modelli di regressione lineare

Output numerico (*variabile dipendente*), sulla base di attributi in input numerici (*variabili indipendenti*):

- ▶ esprimiamo la variabile dipendente come combinazione lineare delle variabili indipendenti:

$$X = w_0 + w_1 a_1 + w_2 a_2 + \dots + w_k a_k$$

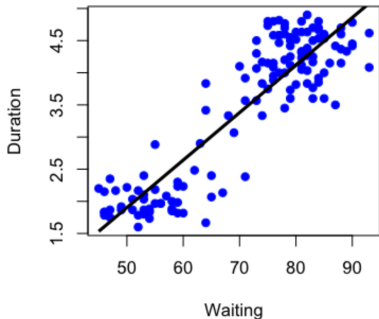
- ▶ calcolo con il metodo dei minimi quadrati, complessità $O(nm^2)$, n istanze, m feature.
- ▶ metodo semplice e di intuitiva interpretazione, tuttavia ha delle limitazioni:
 - ▶ la relazione che si vuole modellare deve essere lineare;
 - ▶ le variabili indipendenti non devono essere “correlate” fra loro (no *multicollinearità*);
 - ▶ Wiki: *linear regression*.

Modelli di regressione lineare / 2

Data mining

Brunello, Marzano

Durata eruzione	Tempo attesa
2.883	55
1.883	54
2.167	52
1.600	52
1.750	47
1.967	55



Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Esistono varianti del modello di regressione (ed ancora più algoritmi che le implementano):

- ▶ *regressione polinomiale*: per relazioni non lineari
- ▶ *regressione multivariata*: per predire il valore di più di una variabile
- ▶ *regressione logistica*: per classificazione binaria
- ▶ *regressione logistica multinomiale e analisi discriminante lineare*: per classificazione multiclasse

$$P(y = 1 \mid x_1, \dots, x_p) = 1 / (1 + e^{-(a + \sum_k b_k x_k)})$$

Caratteristiche:

- ▶ modello lineare generalizzato:
 - ▶ no relazione lineare fra variabile dipendente e variabili indipendenti (funzione link *logit*) ...
 - ▶ ... ma si considerano ancora solamente somme pesate (no interazione fra i parametri);
- ▶ output $\in [0, 1]$, interpretabile come una probabilità;
- ▶ classificazione binaria;
- ▶ è possibile analizzare il “contributo” di ciascun parametro studiando i relativi coefficienti e *p-value*.

Viene risolto un problema di ottimizzazione:

- ▶ coefficienti stimati attraverso *Maximum likelihood estimation*;
- ▶ processo iterativo (es. *glm* in *R*, cap a 25 iterazioni).

Condizioni di applicabilità:

- ▶ numero di istanze di training, \approx minimo 10 casi per parametro;
- ▶ i parametri possono essere numerici o nominali;
- ▶ evitare fenomeni di *multicollinearity*;
- ▶ evitare la cosiddetta (*quasi*-)complete separation.
- ▶ Wiki: *Logistic regression*.

Caso pratico: stima della propensione

Data mining

Brunello, Marzano

Servizio outbound, sondaggi di opinione:

- ▶ un numero di telefono, nel tempo, viene chiamato più volte, attraversando diversi stati (di transizione / finali);
- ▶ **regola generale:** più stati finali sono legati ad un numero, minore è la probabilità di effettuare un nuovo sondaggio;
- ▶ **eccezioni:** es., se nel passato è stato svolto con successo un sondaggio ad un determinato numero telefonico, sarà più probabile effettuarne un altro in futuro.

⇒ Dato un numero di telefono, possiamo stabilire, conoscendo lo storico, la sua propensione ad effettuare un ulteriore sondaggio?

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Criteri:

- ▶ un'istanza per ciascun numero di telefono, attributi:
 - ▶ cumulatori relativi agli esiti storici (finali);
 - ▶ il risultato dell'ultima chiamata, *positivo* o *negativo*;
- ▶ **training:** $\approx 1.300.000$ istanze;
- ▶ **test:** ≈ 250.000 istanze;

Funzione *glm*, linguaggio *R*:

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.502068	0.001524	-985.639	< 2e-16	***
n_fallito_il_contatto	-0.741944	0.012898	-57.525	< 2e-16	***
n_numero_inesistente	-3.044570	0.025546	-119.180	< 2e-16	***
n_non_interessato	0.596997	0.075487	7.909	2.60e-15	***
n_rifiuto_conversazione	-0.216311	0.002114	-102.330	< 2e-16	***
n_sondaggio	0.715526	0.005188	137.913	< 2e-16	***
n_ditta	-1.576644	0.174214	-9.050	< 2e-16	***
n_deceduto	-0.400226	0.055515	-7.209	5.62e-13	***
n_trasferito	-0.406277	0.048669	-8.348	< 2e-16	***
n_numero_sbagliato	-0.858462	0.012623	-68.009	< 2e-16	***
n_non_parla_italiano	-0.476779	0.047350	-10.069	< 2e-16	***
n_troppo_vecchio	-0.487005	0.004547	-107.116	< 2e-16	***
n_troppo_giovane	0.392120	0.040756	9.621	< 2e-16	***
n_sacerdote_o_laureato	-1.236127	0.389908	-3.170	0.00152	**
n_blacklist	-1.298493	0.322074	-4.032	5.54e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring iterations: 6

- Modello facilmente implementabile nei sistemi informativi esistenti, es. in SQL.

Accuratezza globale del modello sull'insieme di test (h = numero di stati storici):

	# inst	$Ratio_{actual}$	$Ratio_{pred}$	e_{abs}	e_{rel}
All	758944	0.162	0.162	0.000	0.000
$h = 0$	469506	0.181	0.182	0.001	0.004
$h = 1$	188089	0.146	0.144	0.002	0.017
$h = 2$	73156	0.111	0.109	0.002	0.015
$h = 3$	18403	0.097	0.101	0.004	0.039
$h = 4$	6077	0.083	0.084	0.001	0.015
$h = 5$	2217	0.062	0.071	0.009	0.134
$h = 6$	1179	0.034	0.060	0.026	0.760

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

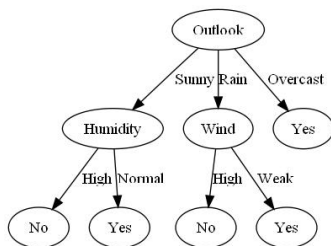
Alberi di decisione

Classificazione delle note

Conclusioni

Metodo semplice e compatto per rappresentare l'output di un processo di **classificazione**:

- ▶ nodo interno = valutazione attributo;
- ▶ classificazione di un'istanza: dalla radice, si scende sino a giungere ad una foglia etichettata con una determinata classe (insieme di classi, distribuzione di probabilità).



Costruzione di un albero di decisione

Presentiamo una metodologia generale per la creazione di un albero di decisione, simile a quella adottata da *ID.3*:

- ▶ costruzione dell'albero ricorsiva, partendo dalla radice;
- ▶ ad ogni passo si sceglie l'attributo su cui effettuare lo split, intuitivamente quello che porta al maggior *information gain* (alberi più bassi, metodo *greedy*);
- ▶ sino alla generazione di una foglia, in corrispondenza di un caso base:
 - ▶ gli elementi nel nodo hanno stessa classe (attenzione all'overfitting, *i.e.* modello costruito ad-hoc sui dati, che non generalizza), oppure
 - ▶ si è raggiunto un sufficiente grado di purezza del nodo .

Costruzione di un albero di decisione / 2

Come misurare il guadagno di informazione dato dallo split su un attributo?

- ▶ insieme T di istanze partizionate nelle classi $C = \{C_1, \dots, C_k\}$ dall'attributo obiettivo;

- ▶ distribuzione di probabilità associata a T :

$$P = (|C_1|/|T|, |C_2|/|T|, \dots, |C_k|/|T|)$$

- ▶ definiamo l'informazione necessaria ad identificare la classe di un elemento di T come:

$$Info(T) = H(P) = - \sum_{i=1}^k p_i * \log(p_i)$$

- ▶ intuitivamente, se quasi tutti gli elementi fanno parte di una stessa classe, $Info(T)$ è bassa, ed aumenta all'aumentare della "confusione" (entropia).

Cosa succede partizionando l'insieme di istanze sulla base di un attributo?

- ▶ supponiamo di suddividere T in sottoinsiemi T_1, \dots, T_n sulla base del valore di una delle *feature*, diciamo X ;
- ▶ l'informazione necessaria ad identificare la classe di un elemento di T è la media pesata dell'informazione necessaria ad identificare la classe dell'elemento all'interno di ciascun sottoinsieme:

$$Info(X, T) = \sum_{i=1}^n (|T_i|/|T|) * Info(T_i)$$

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Diamo infine la definizione di *information gain*:

$$Gain(X, T) = Info(T) - Info(X, T)$$

- ▶ rappresenta la differenza fra l'informazione necessaria ad identificare la classe di un elemento di T e l'informazione necessaria dopo la suddivisione di T in sottoinsiemi attraverso l'attributo X ;
- ▶ in altre parole, il guadagno d'informazione dovuto all'attributo X (alto è meglio).

Costruzione di un albero di decisione / 5

Data mining

Brunello, Marzano

Considerazioni finali:

- ▶ utilizzare l'information gain può portare a preferire la scelta di attributi con un gran numero di valori (*highly branching attributes*), rischio di *overfitting*;
- ▶ l'*information gain ratio* considera anche il numero e la dimensione dei vari sottoinsiemi che verrebbero generati.

Oltre all'entropia, esistono altre misure di impurità dei nodi.

- ▶ indice di impurità di *Gini*;
- ▶ se l'attributo da predire è numerico, si può utilizzare *RMSE* rispetto alla media del nodo;
- ▶ ...

Algoritmi: *ID.3*, *C4.5* R8 (*J48*), *C5.0*, *CART*, ...

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni

Caso pratico: classificazione delle note degli agenti

Ricordiamo i 6 cluster che erano emersi:

- ▶ note articolate, ben scritte;
- ▶ note non articolate, comunque ben scritte;
- ▶ note abbreviate;
- ▶ note con linguaggio del dominio;
- ▶ note nonsense;
- ▶ note con caratteristiche ibride.

⇒ Data una nuova nota, vogliamo classificarla in uno dei 6 cluster noti.

- ▶ A ciascuna delle 1.000 note clusterizzate, aggiungiamo un attributo: *etichetta del cluster*.
- ▶ Attraverso un processo di *feature selection*, selezioniamo gli attributi più opportuni su cui basare la predizione:
 - ▶ articoli_su_parole;
 - ▶ congiunzioni_su_parole;
 - ▶ preposizioni_su_parole;
 - ▶ sostantivi_su_parole;
 - ▶ indice_gulpease;
 - ▶ non_riconosciuti_su_parole;
 - ▶ riconosciuti_dict_su_parole;
 - ▶ riconosciuti_abbr_su_parole;
 - ▶ riconosciuti_dominio_su_parole.
- ▶ Tramite l'algoritmo *J48*, effettuiamo learning con obiettivo l'etichetta del cluster;
- ▶ [dettaglio dell'output di *J48*].

Esempio di applicazione del modello

Data mining

Brunello, Marzano

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regresione lineare

Regresione logistica


Stima della propensione


Alberi di decisione


Classificazione delle note


Conclusioni

	valore text	gruppo_nota text
1	info voltura	hybrid
2	invio del f24 unico	articulated
3	informazioni per appunt sub e comunica dati catastali	articulated
4	info posizione pagamenti mensa scolastica	hybrid
5	NON RISPONDE	non-articulated
6	Info	abbreviated
7	VIA [REDACTED] MQ 37 C'è SCRITTO 43 BOLLETTAZIONE SBAGLIATA. DEVE PASSARE AGLI SPORTELLI PER RETTIFICA DI METRATURA CON PIANTINA SCALA 1:100. RIFERISCO. C	articulated
8	SIGNORA CHIAMA PER SAPERE SE È STATA APPLICATA LA DETRAZIONE DI 25 euro per figlio sul calcolo	articulated
9	la signora aveva chiamato il 23/05 per una verifica posizione per la TARES: ha un locale come	articulated
10	chiede quanto deve pagare per la tassa. Parlati con esatto: deve pagare 61 euro.	articulated
11	info boll	abbreviated
12	rimborso ud	non-articulated
13	tasi	domain-specific
14	INFO GENERICHE IMU, TASI	domain-specific
15	info su avv sosp	hybrid
16	chiede se può rateizzare l'importo da versare per la mensa. Riferito che deve fare richiesta	articulated
17	invio copia boll	hybrid
18	chiede il saldo mensa. Riferito che abbiamo problemi tecnici tecnici al server	articulated

 Data Science Lab @ Uniud:
<http://datasciencelab.dimi.uniud.it/it/>

 I. H. Witten, E. Frank, M. A. Hall: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edition, 2011.

 G. James, D. Witten, T. Hastie, R. Tibshirani: *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

 T. Fawcett: *An introduction to ROC analysis*. Pattern Recognition Letters 27, pp. 861-874, 2006.

Introduzione

Cos'è il Data Mining

Il processo di Data Mining

Approcci per il learning

Learning non supervisionato

Regole di associazione

Tecniche di clustering

Analisi delle note agenti

Learning supervisionato

Regressione lineare

Regressione logistica

Stima della propensione

Alberi di decisione

Classificazione delle note

Conclusioni