

KNN

Caso di studio di Metodi Avanzati di
Programmazione

AA 2021-2022

Corso A

Data Mining

Lo scopo del **data mining** è l'*estrazione* (semi) automatica di *conoscenza* nascosta in voluminose basi di dati al fine di renderla disponibile e direttamente utilizzabile



Aree di Applicazione

1. previsione

utilizzo di valori noti per la previsione di quantità non note (es. stima del fatturato di un punto vendita sulla base delle sue caratteristiche)

2. classificazione

individuazione delle caratteristiche che indicano a quale gruppo un certo caso appartiene (es. discriminazione tra comportamenti ordinari e fraudolenti)

3. Regressione

Predizione del valore di un attributo numerico associato a un esempio sulla base di valori osservati per altri attributi dell'esempio medesimo

3. segmentazione

individuazione di gruppi con elementi omogenei all'interno del gruppo e diversi da gruppo a gruppo (es. individuazione di gruppi di consumatori con comportamenti simili)

4. associazione

individuazione di elementi che compaiono spesso assieme in un determinato evento (es. prodotti che frequentemente entrano nello stesso carrello della spesa)

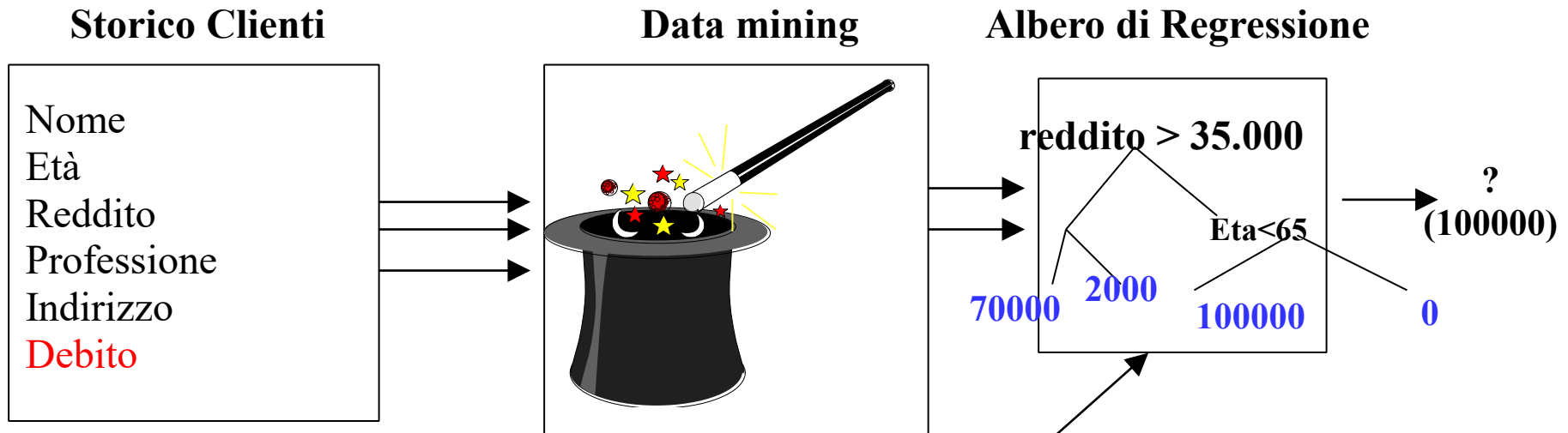
5. sequenze

individuazione di una cronologia di associazioni (es. percorsi di visita di un sito web)

...

Regressione

Considerando dati storici relativi a passati clienti e pagamenti, predire l'ammontare del debito del cliente con la banca



Dati di un nuovo cliente: **Paolo Rossi, 35,37.000,architetto,Bari, ?**

Regressione

- Apprendimento induttivo da esempi per **imparare** la definizione di una **funzione di regressione**
- Gli esempi usati per l'apprendimento sono descritti come **vettori di coppie attributo-valore** per i quali è nota l'attributo classe (target)
- Nella regressione l'attributo target è numerico

Regressione: KNN

Dato

- un training set (\mathbf{X}, Y)
- un esempio \mathbf{x} denominato query per il quale il valore y sia sconosciuto
- un intero $k > 0$

Predice il valore sconosciuto di y associato ad \mathbf{x} identificando i k esempi del training set più vicini ad \mathbf{x} e restituendo la media dei valori y nei k vicini selezionati

KNN

Input

Input: una collezione di esempi di apprendimento (**training set**), ciascun esempio è una tupla di valori per un prefissato insieme di attributi (**variabili indipendenti**)

$$\mathbf{X} = \{X_1, X_2, \dots, X_m\}$$

e un attributo di classe numerico (**variabile dipendente/target**). L'attributo X_i è descritto come **continuo** o **discreto** a seconda che i suoi valori siano numerici o nominali.

L'attributo di classe Y è numerico e ha valori **nell'insieme dei numeri reali**

KNN

Input & query

X1	X2	Y
A	B	1
A	B	2
E	B	2
E	C	3
F	C	4
A	C	2

QUERY

X1=A & X2= B Y= ?

K=1

KNN

- Calcolo le distanze tra ciascun esempio nel training set e l'esempio query rispetto ai valori assunti dalle variabili indipendenti **X**

X1	X2	Y
A	B	1
A	B	2
E	B	2
E	C	3
F	C	4
A	C	2

Distanza di Hamming con variabili categoriche

$$D(\langle A, B \rangle, \langle A, B \rangle) = 0$$

$$D(\langle A, B \rangle, \langle A, B \rangle) = 0$$

$$D(\langle E, B \rangle, \langle A, B \rangle) = 1$$

$$D(\langle E, C \rangle, \langle A, B \rangle) = 2$$

$$D(\langle F, C \rangle, \langle A, B \rangle) = 2$$

$$D(\langle A, C \rangle, \langle A, B \rangle) = 1$$

KNN

- Identifico i k esempi di training più vicini rispetto a X1 e X2 e restituisco la media della variabile dipendenti nei vicini selezionati

X1	X2	Y
A	B	1
A	B	2
E	B	2
E	C	3
F	C	4
A	C	2

Distanza di Hamming con variabili categoriche

$$D(\langle A, B \rangle, \langle A, B \rangle) = 0$$

$$D(\langle A, B \rangle, \langle A, B \rangle) = 0$$

$$D(\langle E, B \rangle, \langle A, B \rangle) = 1$$

$$D(\langle E, C \rangle, \langle A, B \rangle) = 2$$

$$D(\langle F, C \rangle, \langle A, B \rangle) = 2$$

$$D(\langle A, C \rangle, \langle A, B \rangle) = 1$$

$$Y = (1 + 2) / 2 = 1.5$$

KNN

Input & query

X1	X2	Y
A	B	1
A	B	2
E	B	2
E	C	3
F	C	4
A	C	2

QUERY

X1=A & X2= B Y= ?

K=2

KNN

- Identifico i k esempi di training più vicini rispetto alle variabili indipendenti X1 e X2 e restituisco la media della variabile dipendenti nei vicini selezionati

X1	X2	Y
A	B	1
A	B	2
E	B	2
E	C	3
F	C	4
A	C	2

Distanza di Hamming con variabili categoriche

$$D(\langle A, B \rangle, \langle A, B \rangle) = 0$$

$$D(\langle A, B \rangle, \langle A, B \rangle) = 0$$

$$D(\langle E, B \rangle, \langle A, B \rangle) = 1$$

$$D(\langle E, C \rangle, \langle A, B \rangle) = 2$$

$$D(\langle F, C \rangle, \langle A, B \rangle) = 2$$

$$D(\langle A, C \rangle, \langle A, B \rangle) = 1$$

$$Y = (1 + 2 + 2 + 2) / 4 = 1.75$$

KNN

Input & query con variabili indipendenti numeriche?

X1	X2	Y
1	10	1
2	50	2
4	100	2
5	60	3
8	20	4
4	40	2

QUERY

X1=1 & X2=50= B Y= ?

K=1

1-distanza con minMax Scaler

$$d(< x_1, x_2, \dots, x_m >, < x'_1, x'_2, \dots, x'_m >) = \sum_{i=1, \dots, m} |x_i - x'_i|$$

Min-max scaler

- Usato per trasformare il training set in modo che tutte le variabili indipendenti abbiano lo stesso range [0, 1]

$$\text{newValue} = (\text{value} - \text{min}) / (\text{max} - \text{min})$$

- Dove min e max sono rispettivamente minimo e massimo nella variabile indipendente da scalare

KNN

Input & query

X1	X2	Y
1	10	1
2	50	2
4	100	2
5	60	3
8	20	4
4	40	2

QUERY

X1=1 & X2=50= B Y= ?

K=1

Applicare minmax scaler a ciascuna variabile indipendente numerica del training set

KNN

Training set	X1	X2	Y		Scaled Trainign set	scaledX1	scaledX2	Y
	1	10	1			0	0	1
	2	50	2			0,142857	0,444444	2
	4	100	2			0,428571	1	2
	5	60	3			0,571429	0,555556	3
	8	20	4			1	0,111111	4
	4	40	2			0,428571	0,333333	2
min	1	10						
max	8	100						

QUERY

X1=1 & X2=110= B Y= ?

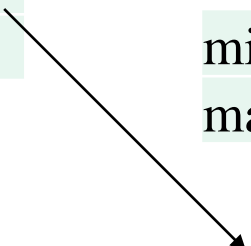
K=1

KNN

- Scalare il query point applicando minimo e massimo determinati per ciascuna variabile indipendente nel training set

X1	X2	Y
1	50	???

min	1	10
max	8	100



scaledX1	scaledX2	Y
0	0,444444	???

KNN

- Calcolo le distanze tra ciascun esempio nel training set (scalato con minmax scaler) e l'esempio query (scalato con il medesimo minmax scaler usato sul training set) rispetto ai valori assunti dalle variabili indipendenti

$d(<0,0>,<0,0.444>)$	0,4444444444
$d(<0.142,0.444>,<0,0.444>)$	0,1428571429
$d(<0.428,1>,<0,0.444>)$	0,9841269841
$d(<0.571,0.555>,<0,0.444>)$	0,6825396825
$d(<1,0.111>,<0,0.444>)$	1,3333333333
$d(<0.428,0.333>,<0,0.444>)$	0,5396825397

Training set	X1	X2	Y	Scaled Trainign set				scaledX1	scaledX2	Y
	1	10	1					0	0	1
	2	50	2					0,142857	0,444444	2
	4	100	2					0,428571	1	2
	5	60	3					0,571429	0,555556	3
	8	20	4					1	0,111111	4
	4	40	2					0,428571	0,333333	2
min	1	10			scaledX1	scaledX2	Y			
max	8	100			0	0,444444	???			

- $d(<0,0>,<0,0.444>)$
0,4444444444
- $d(<0.142,0.444>,<0,0.444>)$
0,1428571429
- $d(<0.428,1>,<0,0.444>)$
0,9841269841
- $d(<0.571,0.555>,<0,0.444>)$
0,6825396825
- $d(<1,0.111>,<0,0.444>)$
1,3333333333
- $d(<0.428,0.333>,<0,0.444>)$
0,5396825397
- K=1 → Y=2

K=2 → Y=1.5

K=3 → y=1.666

KNN

- Nel caso di variabili indipendenti miste
 - Si applica la **distanza di Hamming** alle variabili discrete
 - Si applica **minmax scaler + 1-distanza** alle variabile continue

Training set	X1	X2	Y	Scaled Trainign set	X1	scaledX2	Y
	A	10	1		A	0	1
	A	50	2		A	0,44444444444	2
	E	100	2		E	1	2
	E	60	3		E	0,55555555556	3
	F	20	4		F	0,11111111111	4
	A	40	2		A	0,33333333333	2

QUERY

X1=A & X2=50= B Y= ?

K=1

min 10
max 100

X1	X2	Y
A	50	???

min		10
max		100

X1	scaledX2	Y
A	0,444444	???

D(<A,0.333>,<A,0.444>)	0,1111111111
------------------------	--------------

K=3 → Y=1.666

Caso di studio

Progettare e realizzare un sistema **client-server** denominato “KNNMiner”.

Il server include funzionalità di **data mining** per l'apprendimento di **modelli KNN** e uso degli stessi come strumento di previsione.

Il client è un applicativo Java che consente di effettuare previsioni usufruendo del servizio di previsione remoto

Istruzioni



- Non si riterrà sufficiente un progetto non sviluppato in tutte le sue parti (client-server, serializzazione,...)
- Le estensioni aggiungono funzionalità, non le rimuovono