Data Mining e Bioimmagini

Discretizzazione, Binarizzazione, Misure di similarità e dissimilarità

Prof. Pierangelo Veltri – 10/10/2023- Autori: Maturo, Accetturo - Revisionatori: Accetturo

Discretizzazione e Binarizzazione

La discretizzazione e la binarizzazione sono due tecniche comuni nell'ambito del data preprocessing, che sono utilizzate per manipolare dati numerici al fine di renderli adatti all'analisi.

La discretizzazione è il processo di trasformazione di dati numerici continui in dati categorici o ordinali, suddividendoli in intervalli o categorie discrete. Questa tecnica è utilizzata quando è necessario trattare dati continui come se fossero dati categorici.

Il processo di discretizzazione:

1. **Selezione dell'attributo:** Si seleziona l'attributo numerico che si desidera discretizzare.
2. **Definizione degli intervalli:** Si definiscono gli intervalli o le categorie in cui si desidera suddividere i dati continui.
   Ad esempio, si potrebbe suddividere un attributo che rappresenta l'età in categorie come "Giovane," "Adulto," "Anziano."
3. **Assegnazione delle categorie:** Si assegna a ciascun valore dell'attributo l'intervallo o la categoria corrispondente.
   Ad esempio, un'età di 30 anni potrebbe essere assegnata alla categoria "Adulto."

La discretizzazione può semplificare l'analisi dei dati e consentire l'applicazione di tecniche statistiche o algoritmi che richiedono dati categorici. Tuttavia, è importante scegliere con cura gli intervalli in modo da non perdere informazioni importanti o introdurre bias nell'analisi.

La binarizzazione è il processo di trasformazione di dati in attributi binari, in cui i valori sono rappresentati come 0 o 1. Questa tecnica è utilizzata quando si desidera semplificare i dati e concentrarsi su condizioni binarie o presenti/assenti.

Il processo di binarizzazione:

1. **Selezione dell'attributo:** Si seleziona l'attributo che si desidera binarizzare.
2. **Definizione di una soglia:** Si definisce una soglia o un valore limite. Ogni valore dell'attributo verrà quindi convertito in 0 se è al di sotto della soglia e in 1 se è uguale o superiore alla soglia.

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

3.**Binarizzazione dei dati:** Ogni valore dell'attributo viene convertito in 0 o 1 in base alla relazione con la soglia definita.

La binarizzazione è spesso utilizzata per convertire attributi che rappresentano condizioni o presenze in variabili binarie, rendendo più agevole l'applicazione di algoritmi di apprendimento automatico, in particolare quelli che richiedono input binari come le reti neurali.

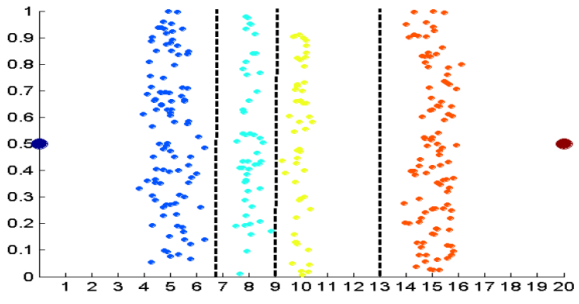
Discretizzazione non supervisionata

Questo processo è "non supervisionato" nel senso che non si basa su informazioni esterne o etichette di classificazione per definire i limiti delle categorie o degli intervalli. Invece, si basa direttamente sui dati stessi per raggruppare i valori simili in categorie.

Ecco come funziona la discretizzazione non supervisionata:

1. **Selezione dell'attributo:** Si seleziona l'attributo numerico che si desidera discretizzare.
2. **Scelta del metodo:** Si sceglie un metodo di discretizzazione non supervisionato.
   Un metodo comune è l'analisi dei cluster, in particolare il clustering **k-means**, questo algoritmo trova automaticamente i limiti dei cluster in base alla somiglianza tra i valori dell'attributo.
3. **Esecuzione dell'algoritmo:** Si applica il metodo di clustering ai dati. L'algoritmo identifica cluster di valori simili all'interno dell'attributo.
4. **Creazione delle categorie o intervalli:** I cluster risultanti vengono usati per creare categorie o intervalli. Ciascun cluster rappresenta una categoria, e i valori all'interno del cluster sono assegnati alla categoria corrispondente.
5. **Assegnazione delle categorie:** Ogni valore dell'attributo viene assegnato a una categoria in base al cluster a cui appartiene.

La discretizzazione non supervisionata è utile quando non si dispone di etichette o classificazioni predefinite per gli attributi, e si desidera comunque suddividere i dati in categorie significative in modo automatizzato. Tuttavia, va notato che i risultati della discretizzazione non supervisionata dipendono dalla scelta del metodo di clustering e dai parametri utilizzati, il che può richiedere un certo grado di sperimentazione.
Questa tecnica è spesso utilizzata per scoprire pattern nascosti o per semplificare l'analisi di dataset complessi.



Sulla sinistra un esempio in cui i dati sono costituiti da quattro gruppi di punti e outliers.
I dati sono unidimensionali, ma viene aggiunto un componente y casuale per ridurre la sovrapposizione.

La linea nera rappresenta il K-means che assume i 4 valori per la divisione dei dati

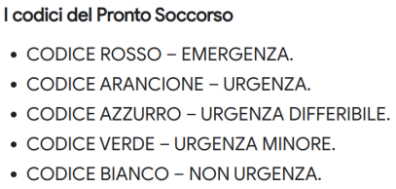
Discretizzazione Supervisionata

Questo processo è spesso utilizzato quando si hanno dati etichettati o quando si desidera creare intervalli di valori basati su una variabile target specifica.
La discretizzazione supervisionata è particolarmente utile quando si desidera creare intervalli che siano ottimali per una specifica variabile di classe o di output.
Questo approccio può migliorare l'efficacia di modelli di classificazione o regressione che utilizzano l'attributo discretizzato come input, poiché gli intervalli sono adattati alle caratteristiche della variabile target.

Trasformazione di attributo

La trasformazione di attributi è una parte essenziale del processo di preparazione dei dati nell'analisi dei dati e nell'apprendimento automatico. Consiste nell'applicare modifiche agli attributi (variabili) di un dataset al fine di migliorare la qualità dei dati, rendere i dati adatti all'analisi o ai modelli di apprendimento automatico, o per ottenere ulteriori informazioni dai dati stessi.

Trasformazioni più comuni:



- Funzioni semplici: x^k, log(x), e^x, |x|
- **Normalizzazione**: La normalizzazione è una tecnica utilizzata per ridurre le differenze di scala tra gli attributi. Questo è importante quando si lavora con algoritmi che sono sensibili alla scala, come le reti neurali o i metodi basati sulla distanza.
  Eliminare il segnale comune e indesiderato, ad esempio la stagionalità
- **La standardizzazione** è simile alla normalizzazione ma riduce la media dell'attributo a zero e la deviazione standard a uno.
  Questa trasformazione è utile quando si vuole ottenere una distribuzione normale dei dati.

Misure di similarità e dissimilarità

Le misure di similarità e dissimilarità sono utilizzate per quantificare quanto due oggetti o campioni siano simili o diversi tra loro.
Le misure di **similarità** sono espresse attraverso un numero appartenente ad un range che va da 0 a 1, dove più il valore è vicino a 1 e più sono simili.
Le misure di similarità sono: **Similarità coseno, Distanza Euclidea, Distanza di Manhattan, similarità di Jaccard, Coefficiente di correlazione di Pearson.**
Le misure di differenza misurano quanto i dati sono **dissimili** attraverso un numero appartenente ad un range che va da 0 a 1, dove più il valore è vicino a 1 e più sono diversi. (opposto della similarità).
Le misure di dissimilarità sono: **Distanza euclidea quadratica, Distanza di Mahalanobis, Distanza di Hamming, Distanza di Levenstein, Distanza di Gower.**
Il termine "distanza" è spesso usato come sinonimo di dissimilarità.

Concetto di distanza

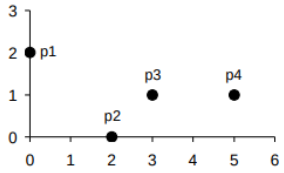
Il concetto di distanza è cruciale per misurare quanto le osservazioni, gli oggetti o i punti siano simili o dissimili all'interno di un dataset multidimensionale.
Il concetto di distanza può esser caratterizzato da alcune proprietà:

1. La distanza è sempre maggiore di 0
2. La distanza è sempre simmetrica
3. Se si considera un oggetto z, la distanza tra x e z è sempre minore della somma delle distanze passando da y.

Se una misura di distanza rispetta tali proprietà è definita metrica.
Dati i due vettori x e y, che sono rappresentati in forma binaria, una modalità di misura delle variazioni, è quello di andare a misurare il numero di cambiamenti a posizioni uguali. (Distanza di Hamming).

Distanza Euclidea

La distanza euclidea è un distanza geometrica in uno spazio **bidimensionale**. È ampiamente utilizzata in algoritmi di clustering come il k-means,
In un dataset con n attributi, la distanza euclidea tra due punti (vettori) x e y è data da:



	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distanza = $\sqrt{((x1 - y1)^2 + (x2 - y2)^2 + ... + (xn - yn)^2)}$

**Sulla destra un esempio di punti su un grafico con riportati i calcoli delle distanze.*

Distanza di Minkowski

La distanza di Minkowski è una misura di distanza generale che può essere utilizzata per calcolare la dissimilarità tra due punti o vettori in uno spazio multidimensionale.
Questa misura è una generalizzazione che include sia la distanza euclidea che la distanza di Manhattan.
La formula generale per la distanza di Minkowski tra due punti x e y in uno spazio n-dimensionale è la seguente:

Distanza = $(\sum|xi - yi|^p)^{(1/p)}$

"xi" e "yi" sono le coordinate dei punti nei rispettivi assi.
"p" è un parametro che può essere regolato per adattare la misura alle esigenze specifiche.
Quando p = 1, si ottiene la distanza di Manhattan; quando p = 2, si ottiene la distanza euclidea.

Esempio
P1 e P2, con le seguenti coordinate: P1 = (3, 4) P2 = (1, 2)
Si vuole calcolare la distanza di Minkowski tra questi due punti con un parametro p uguale a 2 (che corrisponde alla distanza euclidea).

Distanza = $((|3 - 1|^2 + |4 - 2|^2)^{(1/2)})$ = **$2\sqrt{2}$**

La distanza di Minkowski tra i punti P1 e P2 è quindi $2\sqrt{2}$, che è equivalente alla distanza euclidea tra questi due punti nello spazio bidimensionale.

Se avessimo utilizzato un parametro p diverso, avremmo ottenuto una misura di distanza diversa.
Ad esempio, con p = 1, otterremmo la distanza di Minkowski di tipo Manhattan, che tiene conto solo delle differenze orizzontali e verticali tra i punti.
La scelta di "p" influenzerà il modo in cui le differenze tra le coordinate sono pesate nella misura di distanza.

SMC e Jaccard

Il coefficiente di Jaccard e il Simple Matching Coefficient (SMC) sono due diverse misure utilizzate per valutare la similarità tra due insiemi.

Il coefficiente di **Jaccard** è spesso utilizzato per misurare la similarità tra insiemi, ed è particolarmente utile quando ci interessa sapere quanti elementi sono in comune rispetto al totale degli elementi nei due insiemi. È più appropriato quando la presenza o l'assenza degli elementi nei due insiemi sono importanti.

Il **SMC**,considera sia gli elementi in comune che quelli senza corrispondenza nei due insiemi. È più appropriato quando si desidera tenere conto anche degli elementi che non sono presenti in entrambi gli insiemi.

Esempio

X = 1 0 0 0 0 0 0 0 0 0
Y = 0 0 0 0 0 0 1 0 0 1

F01= 2 (numero di variazioni dove x valeva 0 e y valeva 1)
F10= 1 (numero di variazioni dove x vale 1 e y vale 0)
F00= 7 (numero di variazioni dove x vale 0 e y vale 0)
F11= 0 (numero di variazioni dove x vale 1 e y vale 0)

Smc= $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$
$= (0+7) / (2+1+0+7) = 0.7$

Jaccard = $= (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$

Con questo metodo si misurano i cambiamenti rispetto a tutto ciò che può accadere, quindi tutte le possibili combinazioni.

Similarità del coseno

La similarità del coseno è una misura utilizzata per valutare quanto due vettori siano simili(sovrapponibili) in uno spazio multidimensionale.

La similarità del coseno misura l'angolo tra due vettori, piuttosto che la distanza euclidea tra di essi. È particolarmente utile quando si desidera valutare la similarità tra documenti, utenti o oggetti in base ai loro attributi o alle frequenze delle parole.

Formula= $\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|},$

"x” “ y" rappresentano il prodotto scalare tra i vettori x e y.
"||x||" rappresenta la norma euclidea (la lunghezza) del vettore x.
"||y||" rappresenta la norma euclidea del vettore y.

La similarità del coseno produce un valore compreso tra -1 e 1, dove 1 indica che i due vettori sono perfettamente allineati (cioè, identici), un valore di 0 indica che i due vettori sono ortogonali tra loro (cioè, completamente diversi), e un valore di -1 indica che i due vettori sono opposti.

La similarità del coseno è una misura di similarità robusta e ampiamente utilizzata, ma è importante notare che non tiene conto delle dimensioni dei vettori o delle differenze nella magnitudine dei valori nei vettori.

Esempio

$\mathbf{d_1} = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$\mathbf{d_2} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$\langle \mathbf{d_1}, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\| \mathbf{d_1} \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d_2} \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{d_1}, \mathbf{d_2}) = 0.3150$

Concetto di prossimità

Il concetto di "prossimità" si riferisce alla vicinanza o alla distanza tra oggetti, punti o entità in uno spazio multidimensionale. La misura della prossimità può essere utilizzata per valutare quanto due oggetti siano simili o dissimili tra loro, la loro relazione spaziale o quanto siano "vicini" in termini di attributi specifici.

PENTAHO

Pentaho è un insieme di strumenti open source per l'integrazione, l'elaborazione, l'analisi e la visualizzazione dei dati. Tra i componenti principali di Pentaho troviamo “Pentaho Data Mining”. Questo modulo supporta le attività di data mining e machine learning ed è possibile creare modelli di previsione e di classificazione utilizzando vari algoritmi.