DataVis per ***Business Intelligence***

(ad esempio fare la visualizzazione dell’andamento di una pizzeria/azienda…)

Slide 3

Dato un grafo si può voler predire la classe a cui appartiene quel nodo.

Ci sono dei metodi che permettono di predire la classe del nodo.

Altri metodi cercano di predire la label dell’edge/dell’arco.

Ci sono archi diretti o indiretti.

Ci sono altri metodi che vanno a predire se un dato arco va da un nodo di una certa classe ad un nodo di un’altra classe.

Metodi:

• *Nodel label prediction*

• *Edge label prediction*

Quei metodi che inferiscono la presenza di un arco che eventualemente c’è o meno si chiamano *Link prediction model*.

Questi metodi analizzano il grafo e trasformano ogni nodo, ogni edge del grafo in vettori.

Fanno l’embedding del grafo in uno spazio ad elevata dimensionalità.

Spazio 100-dimensionale → vuol dire che ogni nodo è espresso in un vettore 100-dimensionale

Dati i vettori, li si possono dare in input ad un modello di apprendimento (come una rete neurale) per poi fare delle predizioni.

Se si riesce a prendere il grafo e proiettarlo in uno spazio 2dimensionale o 3dimensionale, posso visualizzare i suoi nodi in modo tridimensionale ad esempio se si è in uno spazio 3dimens.

In uno spazio bidimensionale il nodo è un’immagine.

In uno spazio tridimensionale il nodo è visto in 3D.

Slide Spectral embedding for visualizations

Primo modo per fare l’embedding.

I degree dei nodi sono (in caso di arco non pesato) il numero di archi connessi ad un nodo.

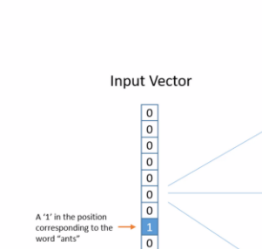
Nel caso di grafi pesati, la matrice di adiacenza ha il valore del nodo.

Più il grafo è grosso, più si hano autovettori, più è elevato l’embedding del grafo.

L’embedding del grafo deve raccontare tutto il grafo, quindi è corretto che sia elevato.

Se si hanno tanti nodi, bisogna avere uno spazio elevato per far sì che ci sia la distanza corretta tra i nodi.

Slide

Prende un nodo, con le *random walk* genera tanti dei suoi vicini, e poi si addestra uno ***SkipGram******model*** per riuscire a predire.

Slide 20

Come si fanno a visualizzare tutti questi nodi?

Si hanno dei punti che sono rappresentati da tante feature, quindi un vettore composto da tanti elementi.

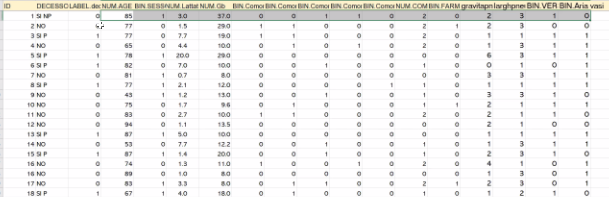
Si ha uno spazio vettoriale a due dimensioni.

Slide PCA hopefully

*Principal Component Analysis* che viene usato soprattutto nel campo delle immagini.

Slide 3

Cos’è un vettore?



Si ha un insieme di punti identificati da un numero.

Ogni punto è un vettore.

Ogni paziente è descritto da 15 numeri, quindi si tratta di ciascun punto descritto da un vettore 15-dimensionale.

Tali punti sono rappresentati in uno spazio 15-dimensionale.

Nel campo dell’analisi si hanno degli spazi (perché composte da due dimensioni) che sono delle variabili intere come l’età, variabili binarie, discrete/continue, categoriche.

Se si vuole visualizzare questi pazienti, ogni riga, ogni punto, come si fa?

Si copiano ad esempio due colonne.

Sull’asse delle x l’età, sull’asse y la variabile binaria.

Si è fatto uno ***scatter plot*** (direttamente in Excel).

[1 sta per femmina, 0 per maschio]

Si possono fare tutti i possibili incroci tra le diverse variabili disponibili per ciascun paziente.

Riduciamo la dimensionalità per vedere i punti, per poterli visualizzare.

Se riusciamo a visualizzare tali punti/dati ad elevata dimensionalità (nel .xls medico 15-dimensioni ad esempio) si possono vedere tante informazioni.

Nell’esempio si hanno 3 classi di pazienti: SI NP, NO, SI P.

Si plottano le variabili e si cerca di capire quale o quali variabili hanno generato la classificazione dei vari dati.

Se si riuscisse a capire quali sono, tra le varibaili che descrivono il paziente, quelle che influenzano dei dati sintomi, magari si potrebbe dal punto di vista medico fare in modo che quelle variabili non raggiungano certi valori in modo da ridurre la sintomalogia.

Quindi si vogliono vedere dei dati rilevanti, riducendo la dimensionalità.

Slide

Riducendo la dimensionlità, non si vuole perdere alcuna informazione, ma anzi guadagnarne di più.

Slide Cosa significa normalizzare dei dati?

Ci sono le varibaili binarie che hanno valore 0 oppure 1.

L’età siccome ha un range di variabilità maggiore, può raggiungere un valore maggiore la distanza tra i due elementi (pazienti).

Se quei due pazienti hanno diverse caratteristiche, ad esempio per la distanza del sesso: un maschio ed una femmina hanno distanza 1 in quanto la femmina è 1 e maschio 0.

Quindi si normalizzano i dati.

Prima normalizzazione, la più semplice, ossia la min\_max normalization.

Dopo la *min\_max\_normalization*, tutte le colonne hanno minimo 0 e massimo 1:

nella min\_max, il minimo valore della feature viene mappato su 0, mentre il massimo valore della feature è mappato ad 1, gli altri valori si trovano nel mezzo.

**STD** → deviazione standard

Si deve pensare che si sta lavorando con valori reali, quindi si fa il **rounding**.

Con la Z-normalization si fa in modo che tutte le varibaili/feauture ossia queste:



(ogni paziente è descritto da queste feature: lattici….)

abbiano lo stesso range di variabilità, perché sono centrate in zero e deviazione standard pari a 1.

L’importante è che tutte le variabili, per comparare feature diverse che descrivono i punti, siano normalizzate per fare in modo di non dare più importanza ad una feature piuttosto che ad un’altra.

Se si ha una dimensione non numerica?

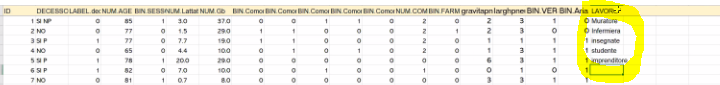
Ad esempio per la feature “sesso” posso mappare i suoi valori “femmina” e “maschio” a 0 e 1.

Con le variabili categoriche?

Sono variabili che identificano delle classi che tante volte non sono ordinabili.

Ci sono **variabili categoriche ordinabili** e **non ordinabili**.

Ad esempio, ci sono delle malattie che derivano dal lavoro che il paziente fa:



Il lavoro è una variabile categorica.

Il “muratore” si può ripetere, ci possono essere più pazienti muratori.

Come faccio a calcolare quanti muratori ci sono?

Per dimostarre che tutti questi pazienti, che fanno un lavoro pesante, hanno una data malattia, si può stabilire di dare un valore al lavoro in base alla forza ad esempio fisica che si impiega per fare quel lavoro.

Quindi si “mappa” ogni lavoro su uno sforzo fisico richiesto: ad esempio insegnante ed impiegato sono allo stesso livello e la loro distanza è zero.

In questo caso le classi quindi possono essere ordinate per sforzo fisico.

Ogni valore identifica una classe specifica e sono ordinabili.

Altro esempio con una variabile categorica non ordinabile: Tumore

Il tumore si può ordinare in base alla sua malignità ma è difficile, quindi come fare per ordinare variabili categoriche non ordinabili?

if categoria1 == categoria2 distanza = 0

Si chiama ***giacart difference*** e si applica ai valori logici .

Salvo che nel problema specifico, le variabili categoriche pesano diversamente.

Se si va a vedere la *min\_max normalization* su variabili booleane (binarie), queste variabili rimangono uguali perché il minimo è 0 e massimo è 1.

Slide 5

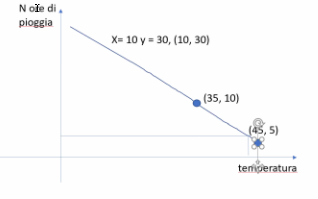
Correlazione tra due variabili

Possono essere correlazioni lineari e non lineari.

Correlazione lineare diretta: vuol dire che quando una varibaile aumenta in valore, aumenta anche l’altra.

Ogni retta rappresenta una correlazione lineare diretta tra i punti presenti sugli assi.

Correlazione lineare inversa: all’aumentare della x, diminuisce la y.

La retta esprime una correlazione lineare inversa.

Se sull’asse delle x si mette ad esempio la temperatura in Italia:

più si alza la temperatura e più sull’asse delle y (su cui ci sono i n° di ore di pioggia) le ore di pioggia diminuiscono (quindi in estate la pioggia diminuisce).

Se si parte dall’inverno (temperatura bassa), si ha tanta pioggia.

Si tratta di una relazione inversa.

Esempio di correlazione lineare diretta

Più sono alti gli assembramenti più sono alti i contagiati da covid.

Più si va avanti e più amentano.

La retta che esprime la correlazione tra i due elementi è descritta da una funzione lineare.

Ci sono due modi per valutare la correlazione:

* Correlazione di Pearson → correlazione di tipo parametrica
* Correlazione di Spearman → non parametrica
* Correlazione di Kendal → non parametrica

Slide 6

La correlazione di Pearson non vuole *outlier*, ossia un punto che si stacca ed è completamente fuori dalla distribuzione.

Slide 7

Una funzione monotona crescente è una funzione per cui all’aumentare della x cresce la y.

Un funzione non monotona decrescente è una funzione per cui all’aumentare dellla x decresce y.

Una funziona non monotona presenta una correlazione lineare inversa.

La correlazione di Spearman funziona come la Pearson solo che fa il **rank** di x.

Spearman va a verificare che il *ranking* delle due variabili sia correlato.

Slide 11

Se si hanno n vettori, la media degli vettori come si calcola? Elemento per elemento.

Slide 12

Si hanno N vettori in uno spazio D-dimensionale molto grosso che si vuole ridurre.

PCA richiede la *Z-normalization*.

Dopo aver normalizzato si calcola la matrice di covarianza nxn dove n sono gli elementi ed il valore x,y è la covarianza tra le due variabili.

Slide 14

Questa matrice di covarianza la si può vedere come una **trasformazione**.

Fa vedere quanti punti covariano tra di loro, quanto gli elementi dei due vettori variano insieme.

La covaraianza dei due vettori dice quanto gli elementi dei due vettori sono allineati tra di loro.

Se si hanno tanti valori ad elevata covarianza vuol dire che si hanno tanti valori nella matrice di partenza che sono simili.

Questo vuol dire che le variabili in qualche modo sono correlate.

Si vede la matrice di covarianza come una trasformazione.

PCA DICE:

visto che la covarianza dice quanto sono simili gli elementi di due vettori;

se la matrice di covarianza contiene valori alti, vuol dire che si ha un’alta correlazione tra i dati e ciò non si vuole.

Ogni vettore deve covariare con sé stesso, ma con gli altri valori si vuole che il valore della covarianza sia pari a 0, cioè si vuole che i vettori non siano per niente similari tra di loro.

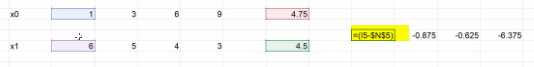
[ I punti sulla matrice sono le variabili. ]

Slide 15

Come si può calcolare la correlazione tra due variabili, si può calcolare la correlazione tra i due vettori.

Si calcola il vettore medio, ossia la media degli elementi del vettore (4.75 e 4.5).

Di seguito tutto il procedimento:



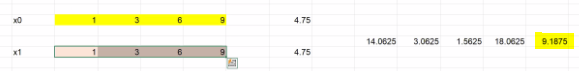
 Questi due vettori hanno una correlazione negativa ed è -3.375.

Quando il vettore giallo cresce, il vettore rosa decresce:

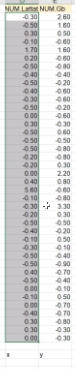
quindi in questo caso il valore -3.375 indica la covarianza.

Vuol dire che c’è un’elevata correlazione tra i due vettori, ma è inversa.

In questo caso invece si ha covarianza positiva:



Slide 16

 **vT** → vettore trasposto

Gli autovettori sono quei vettori V tali per cui la loro trasformazione tramite la matrice, su cui si calcolano gli autovettori, non li cambia radicalmente, ma li ***stretcha***.

*Lambda* è una costante ed indica il valore di stretch degli autovettori, quanto li allunga.

Più lambda è alto più i punti si distribuiscono lungo l’asse.

[ Si hanno due variabili, che si considerano come due punti e si misura la covarianza:

←

Covarianza:  che è una matrice 2x2.

Questi sono i due vettori di proiezione: 

Gli autovettori e gli autovalori dicono come stretchano i dati. ]

Quindi si sta diagonalizzando la matrice di covarianza.

Si sta decomponendo la matrice di covarianza in una matrice diagonale che è composta dai suoi valori ed una matrice che è comporta da vettori che sono ortognali tra di loro la cui covarianza è zero.

Quindi si sta trovando una base per la matrice di covarianza che è composta dalla matrice di correlazione che è ortogonale (moltiplicata per la matrice diagonale).

Slide 17

PCA fa la costruzione di un nuovo spazio che riassume bene i punti, ottiene dei nuovi punti che sono degli autovettori ottenuti come combinazione lineare dei vettori di partenza.

Gi autovettori sono una combinazione dei vettori che si aveva nel file Excel.

Ciò signfica che nel nuovo spazio vettoriale le variabili sono completamente indipendenti le une dalle altre.

Slide 18

Questi sono dei punti in uno spazio generato a caso.

I punti sono stati già normalizzati (in quanto hanno due soli valori), quindi hanno già subìto la *Z-normalization*.

SI tratta quindi di vettori che esprimo ad esempio lattici (asse y) e globuli bianchi (asse x).

C’è una correlazione lineare: al crescere della x crescono anche i valori della y.

A tutti gli effeti la PCA cosa fa?

Se devo calcolare la correlazione tra due punti, si vede che c’è una bella correlazione.

Questo è l’asse ottimale.

Per ridurre lo spazio vettoriale, dato che si hanno 2 dimensioni, si vuole quindi passare ad uno spazio 1D.

Slide 20

Si sta andando a cercare un nuovo asse che dipende da questi punti e fa sì che i punti prioiettati sul nuovo spazio siano espressi su variabili completamente scorrelllate, quindi su variabili massimamente distanti.

Python PCA: libreria ***sklearn*** per fare PCA

<https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA.transform>

Progettino “Test PCA applied on RGB image”.

La PCA è un metodo di *data embedding* che consente di ridurre la dimensionalità, creando un nuovo spazio vettoriale.

Ha la caratteristica di mantenere le covarianze.