Slide 2

Dati multivariati, dati continui.

Multivariati → ogni punto è rappresentato da più valori e quindi da più variabili.

Ogni punto è un punto in uno spazio multidimensionale, composto da tante dimensioni, dette variabili, *features* o caratteristiche.

Si prende il dataset, lo si vede come tabella, dove ogni riga è un punto ed ogni colonna una feature.

1. Si prende ogni singola feature, la si plotta, per vedere quali feature non hanno senso, e si rimuovono (si rimuovono tali colonne).
2. Si normalizzazno i dati tra 0 e 1 (in modo da avere ogni feature che assume come max 1 e min 0) oppure si può sottrarre ad ogni colonna (feature) la media di tutti i punti di quella feature (colonna) e si divide per la ***standard deviation*** di questa feature; questo perché tutte le feature sono portate ad avere media pari a 0 e deviazione standard pari a 1.
3. Dato che si stanno considerando punti che appartengono a classi diverse, si possono fare dei test di ipotesi per andare a vedere se per ogni variabile le distribuzioni provenienti da classi diverse sono simili o diverse.

Se si hanno dati continui si usa il **t-test**.

Se si hanno dati categorici si usano altri tipi di test.

Dati continui → quelli descritti da numeri; tendenzialmente sono detti continui o ordinabili (possono anche essere discreti) come l’altezza.

1. Dopo aver normalizzato e fatto il test di ipotesi, si può prendere ogni varaibile, ogni classe, si separano i punti dalle classi e si plottano.

Ad esempio si può fare per ogni classe un **istogramma** e se questi istogrammi vegnono smussati diventano dei ***density plot***.

I ***box plot*** sono delle visualizzazioni di distribuzioni ed utilizzano i ***notch***.

Se i notch di due box plot differenti (rappresentanti due diverse distribuzioni) sono sovrapposti significa che le due distribuzioni sono molto simili.

Se non si sovrappongono le due distribuzioni sono diverse.

Slide 3

5. Si ha un dataset multivariato e plottare tutte le feature potrebbe essere complesso; quindi meglio rimuovere quelle feature che sono ridondanti.

Se ci sono feature che crescono o si “abbassano” insieme → si confutano delle correlazioni.

La correlazione tra coppie di feature permette di capire quali sono le feature strettamente correlate.

Feature altamente correlate sono ridondanti, ad esempio dove la correlazione è maggiore di 0.8 o 0.9.

Un’altra correlazione che si può fare è tra le feature e le label.

Feature altamente correlate alle label hanno un alto potere discriminante.

6. Si può calcolare **TSNE** (più comune per i dati multivariati al posto della **PCA**) in uno spazio multidimensionale.

Si visualizzazno i punti tipicamente tramite uno ***scatter plot*** con 3 dimensioni, (si parla quindi di un punto 3D).

Per i punti TSNE c’è una possibilità che siano separabili, anche se TSNE tiene in considerazione le label → perciò attenzione i punti TSNE (ossia i punti proiettati sullo spazio TSNE) non possono essere utilizzati per la classificazione.

Slide 4

Le *MIC statistics* sono delle statistiche che provengono dalla **MINE family** e sono molto meglio della *Pearson* e *Spearman* *correlations*.

Poiché queste due ultime riescono a identificare correlazioni lineari.

Due feature o feature e label ottengono una correlazione di Pearson elevata se la loro dipendenza è lineare.

Si plottano feature 1 e feature 2 con uno ***scatter plot***; se questi punti sono tendenzialmente lungo una retta si ha una correlazione di *Pearson* alta.

Se questi punti sono lungo una sinusoide *Pearson* fornisce un valore bassissimo.

Anche quando ci sono dipendenze non lineari Pearson dà valori bassi e questo non si vuole.

Slide 5

Vogliamo le caratteristiche di **MIC** (*Maximal Information Coefficient*):

è una misura della forza dell'associazione lineare o non lineare tra due variabili X e Y.

Slide 6

Ovvero si vuole generalità ed equitability.

Slide 7

La *generality* è utile soprattuto quando la relazione è espressa da una composizione di funzione (ossia ha funzioni più complesse che lineari).

L’*equitability* parla di coefficiente di determinazione (è una sorta di coefficiente di correlazione).

Slide 8

È un coefficiente che esprime quanto un’interpolazione (la **f** a DX) riesce ad adattarsi bene ai dati.

Dice quanto due funzioni sono vicine.

Slide 9

Si suppone di avere un dataset espresso in uno spazio bidimensonale, ossia il dataset è composto da punti bidimensionali; facendo uno ***scatter plot*** di questi punti: se si ha un’alta correlazione allora si può prendere/creare una griglia che permetta massimamente di incapsulare i dati.

Per griglia si intende che si può esprimere il dataset in uno spazio cartesiano, ne si fa lo ***scatter plot***.

Si può creare una griglia che con pochi box della griglia riesca ad includere al massimo i dati.

Si possono creare griglie X x Y che dividono lo spazio in X righe ed Y colonne, ovviamente allineate.

Si dividono in righe verticali ed orizzontali per avere una griglia di punti che riesce ad esprimere la relazione.

Quindi si possono esprimere tutte le griglie X x Y che si vuole e si sceglie la migliore, in cui si può trovare un coefficiente di relazione.

Come si determina la griglia migliore?

Si usa un concetto di “copertura” della griglia rispetto la funzione.

Slide 10

Ciò che fa MIC:

preso un dataset (2-dimensionale in questo caso), cerca tutte le griglie di dimensione xy (dove le griglie xy si intendono tutte le giriglie che dividono lo spazio2D in x righe ed y colonne), quindi si hanno X x Y quadratini.

Gli autori di MIC hanno stablito che bisogna avere **B(m)** righe e **B(m)** colonne.

Si fanno tutte le possibili griglie, per ogni possibile dimensione della griglia, per ogni scala della griglia.

Si suppone si avere tutte griglie 7x3, ossia che dividono lo spazio in 7 righe e 3 colonne.

Per ogni possibilità di questa griglia, ad esempio una griglia 2x3 nella figura si va a trovare la griglia migliore.

Quindi per ogni griglia 2x3 si calcola la qualità della griglia, ossia quanto la griglia copre la relazione che si vede nel dataset.

Si trova tra queste griglie quella migliore e si usa un coefficiente che esprime la qualità della griglia migliore, ossia il coefficiente **mi** (*mutual information*).

Per esprimere la qualità di ogni griglia si usa il coefficiente di informatione mutua tra la griglia e il dataset.

Se si sostituissero i punti del dataset con i punti della griglia si potrebbe potenzialmente spiegare il dataset.

Si salva in una matrice, in questo caso 2x3, il valore della griglia migliore, normalizzato dal **log(min(x,y))**.

Quindi nel caso delle griglie 2x3 si normalizzano tutti i valori di qualità divisi per log(2) ↑ (il minimo tra x ed y è 2).

Per tutte le possibili griglie 2x3, 7x11…si trova la griglia migliore e si registra la sua informazione mutua (il valore massimo ottenuto normalizzato) in una matrice caratteristica **M(D)xy**, dove x ed y sono minori di **B(m)**.

Fissata la griglia **grid(r,c)** con r righe e c colonne, si calcola per ogni possibile griglia (ad esempio una di queste ) tutte le informaizoni mutue e se ne calcola il massimo su quella dimensione xy.

Il massimo lo si registra poi nella matrice **M(D)xy** che ha tutte le possibili xy.

Si possono vedere le coordinate xy come la dimensione della griglia e il valore di **M(D)** come il valore 3D lungo l’asse z → si può quindi avere la matrice 3D di cui si può prendere il massimo (sostanzialmente si calcola il massimo).

Il valore di **M(D)xy** costituisce il *MIC coefficient*.

Perché per ogni griglia si prende l’informazione mutua?

Slide 12

Cos’è l’informazione mutua **I(X;Y)**?

Si supponga di avere due variabili x ed y.

Come si capisce la probabilità congiunta?

La probabilità che avvengano insieme diviso la probabilità che avvenga uno e che avvenga l’altro.

Slide 13

Applicando la probabilità congiunta con le griglie:

per il primo box si vedono quanti punti “cadono” in quel box diviso la dimensione di quel box e si somma la stessa probabilità che si ottiene per il 2° box, 3°…

Slide 14

La probabilità su x è la probabilità che tutti i punti cadono su una riga, composti dai box lungo la riga.

Slide 15

Stesso ragionamento per y.

Slide 17

L’*entropia* è la misura dell’incertezza (***Shannon***).

Se si ha un dataset, l’entropia indica quanto è sparso (“incasinato”) il dataset.

**H(X|Y)** → L’entropia condizionata degli eventi indica quanto è l’entropia su X dato che si è osservato Y.

Slide 18

L’informazione mutua è l’entropia sul dataset D [ se si sostituisce X, in **H(X)**, a D ] dato che si ha la griglia Y.

Perché MIC usa l’entropia?

Immagine che contiene testo

Descrizione generata automaticamenteSe si ha la configurazione di griglie X x Y, quanto la griglia riesce a spiegare il dataset?

Se si usa la griglia 2x3, quanto questa griglia riesce a spiegare il dataset sottostante?

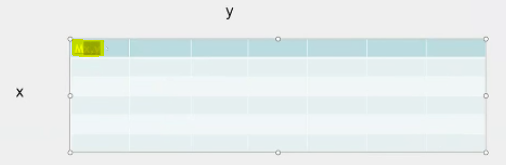
Si calcola l’informazione mutua.

Poi si passa alla griglia sotto Immagine che contiene testo

Descrizione generata automaticamente e sotto ancora.

Dei valori ottenuti di informazione mutua, si sceglie il massimo, che è alla fine questo: **mxy= max(min(x,y, D))**

Questo valore **mxy** lo si registra nella matrice **M(D)xy**, dopo averlo opportunamente normalizzato.



Indipendentemente dal grafico, si è ottenuto per ogni possibile griglia che divide l’immagine in x righe ed y colonne, per ogni possibile coppia di valori xy, si è ottenuto un valore massimo, ossia un valore **mxy**.

Se si calcola il massimo di tutti questi valori, si trova il *MIC coefficient* e non solo si trova lo *splitting* del dataset ottimale ma si trova anche una griglia (normalizzata) in grado di spiegare il dataset e la relativa informazione mutua.

Slide 20

Il **MAS** esprime la monotonicità della relazione.

Se si ha una relazione monotona (ossia che va sempre verso l’alto o verso il basso) si avrebbe una matrice simmetrica, quindi invece di andare a prendere il massimo, si vede quanto è simmetrica la matrice che si ottiene.

Slide 21

Se invece si vuole capire quanto quel dataset, se plottato con uno ***scatter plot***, si distribuisce lungo una funzione che è continuamente derivabile si utilizza il **MEV**.

Slide 22

**MCN** → Alternativamente si può vedere qual è la dimensione che ci ha portati a calcolare il massimo.

Slide 23

**TIC** → si calcola la somma dei valori della matrice – è più potente del MIC, in quanto MIC a volte fornisce un valore alto anche con l’indipendenza, ossia quando non c’è una relazione forte.

Slide 24

Gli autori dicono di calcolare la somma della matrice e i valori che superano una certa soglia si possono analizzare con MIC.

Slide 25

Diversi tipi di relazione come reagiscono ai vari tipi di statistiche: **MIC**, **MAS**, **MEV** ed **MCN**.

MIC fornisce un valore minore.

Slide 27

Sull’asse x è indicato il coefficiente di determinazione.

Il MIC ha un andamento praticamente sempre uguale per divesi tipi di relazione; mentre altri tipi di coefficcienti di correlazione danno magari lo stesso valore a relazioni che sono completamente diverse.

MIC è abbastanza coerente per le relazioni percepite dalla visione umana.

*Time Series Data*

Slide 32

Si tratta di una ***Time Series*** → si ha un incremento nel tempo o comunque una funzione che varia nel tempo in base ad un dato parametro.

Si supponga di avere due distribuzioni **y1** ed **y2** generate con MatLab: sono una serie di esperimenti e per ognuno si ha un valore di accuratezza.

Ogni funzione è esprimibile con un incremento.

Ad occhio quale delle due funzioni ha un incremento rispetto al punto precedente?

Ad esempio il punto 600 è ottenuto da un incremento del punto precedente.

La **y2** ha un incremento.

Si supponga che si abbia un incremento costante, ma come si fa a capire di quanto è l’incremento?

Le due funzioni si impennano, ci sono dei valori diversi da zero ma non li vediamo.

Slide 33

Se si pensa alla funzione logaritmica, i numeri bassi vengono impennati e i numeri alti vengono appiattiti.

Slide 34

Si applica la funzione logaritmica ad **y1** ed **y2**, che si suppone siano fatte come in slide.

L’incremento della funzione **y1** è piuttosto basso, ma di quanto?

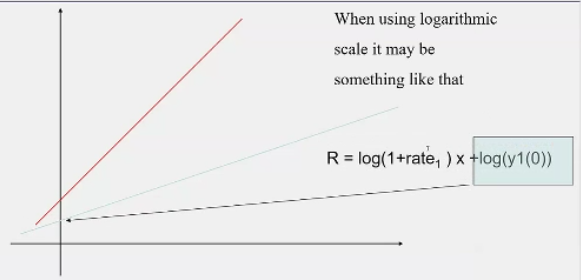
Slide 35

Se invece di plottare **y(t)** si plottasse **log(y(t))**, ciò che succede è che si vuole fare in modo di avere un valore che sia sempre maggiore di 0.

Il ***rate*** si suppone costante.

**t** è il valore del parametro che si suppone avere nel tempo.

Slide 36

Si è supposto che **rate1 > rate2**.

**R** è l’equazione della retta (mx+q), dove m è **log(1+rate1)** mentre il valore di intercetta è **log(y1(0))**.

Si vede qual è la retta più pendente e quella più pendente è quella che ha m più elevato, ossia che ha l’incremento più elevato.

Slide 37

Se si vuole studiare la percentuale di incremento di una Time Series, com’è descritto il trend bisogna stare attenti: l’incremento prima era pari allo 0.01.

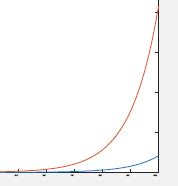
Ossia il trend di ascesa è sempre lo stesso.

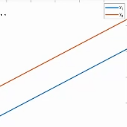
Finché questo era la scala non logaritmica, appena si passa alla scala logaritmica, le percentuali sono molto più alte.

Quindi quando si visualizza bisogna fare attenzione alla scala che si sta utilizzando!

Quando si hanno delle pendenze elevate o troppo basse, se si avessero tutti valori prossimi allo zero, si potrebbe utilizzare una scala esponenziale, così da ingrandire le distanze.

Quando si ha un incremento elevato si usa la scala logaritmica per vedere meglio la variazione.

Prima, con la scala normale, ci sembrava che y2 avesse un incremento maggiore.

 Ma invece utilizzando una scala logaritmica si può vedere che l’incremento è lo stesso.

ATTENZIONE ALLE SCALE!

Slide 38

Invece di esprimere in modo assoluto i valori, esprimerli con il **log** che fornisce la percentuale di cambiamento rispetto ai valori precedenti o addirittura utilizzare direttamente le percentuali.

Slide 40

Come si analizzano le time series?

In slide c’è la distribuzione di un valore nel tempo.

Nella time series si cerca di vedere se c’è un fenomeno che si ripete nel tempo, se la time series è ciclica o magari ci si concentra in dati periodi per vedere qual è l’andamento in quel periodo.

Se però si vede tutta la time series nel tempo, non si capisce molto.

Slide 41

Generalmente si prendono i periodi che si intendono analizzare e li si superimpongono utilizzando colori differenti.

Si può riuscire così a vedere se il picco è sempre nello stesso periodo, ad esempio a Settembre c’è sempre un picco nei vari anni.

Dal punto di vista della visualizzazione per vedere la periodicità dei dati, l’unico modo è analizzare dei dati periodi oppure si visualizzano i picchi, i minimi ed i massimi, e si vedono le distanze tra questi minimi e massimi.

Slide 47

È stato fatto il LOG per vedere se gli *outlier* (in questo caso i valori troppo alti) dessero fastidio, e in effetti lo danno, non fanno vedere bene i picchi (che si possono vedere zommando: ).

Slide 49

Quindi sono stati abbassati ed alzati rispettivamente gli ***outlier*** e i ***lowlier*** per vedere meglio le variazioni.

I trattini più bassi sono gli *outlier*.

Slide 50

I ***box plot*** vicino sono relativi alla stessa feature.

I ***notch*** sono ben separati tra di loro.

Se si facesse il test di ipotesi statistica si otterebbero dei valori alti di ***pvalue*** e non è necessario considerare il test di ipotesi nulla.

A questo punto si possono scegliere le feature da considerare.

Slide 51

Altrimenti se si volessero considerare le singole feature si può fare lo ***small multiple***: si fa lo stesso plot per tutte le feature, in modo da comparare le diverse feature.

Slide 52

Sono stati considerati **CHIUSO** e **APERTO** in un unico slot temporale.

Slide 53

Questo è un intervallo occhio chiuso e occhio aperto.

Slide 54

È la linea di trend per una feature.

È più utile fare questo plot?

Slide 55

Oppure il ***cycle plot***?

Si divide la time series in intervalli di interesse.

Si prendono tutti gli intervalli e si decide di dividere ognuno di essi in sottointervalli.

Si prende la media di ogni sottointervallo e la si plotta.

Il ***cycle plot*** dice qual è stato l’andamento nel tempo per il nostro periodo di interesse.

Tendenzialmente in questo plot si salva la media di questi valori.

Si può vedere come la media cambia nel tempo in un dato sottointervallo.

Questo tipo di plot viene utilizzato soprattutto nel campo delle azioni finanziarie.

Si osserva per diversi periodi la stessa azione e si fa un ***cycle plot*** per vedere ad esempio per la stagione primaverile, estiva, autunnale ed invernale, come quell’azione si comporta nei diversi anni.

Slide 56

Alternativamente un modo per visualizzare le serie temporali è utilizzare i ***radar plot***.

I ***radar plot*** per ogni istante temporale plottano il valore di una data variabile.

Perché il radar plot è importante quando si vogliono andare ad analizzare dei cicli nelle time series?

(Pensare alla spiegazione del seno e del coseno, funzioni periodiche, che sono state fatte con una circonferenza.)

Se si visualizza una funzione sul *radar plot*, si può visualizzare la media dei cicli o la separazione degli intervalli, il loro trend nel tempo.

Siccome un *radar plot* è diviso in “spicchi”, viene spontaneo confrontarli.

Slide 58

Alternativamente si può stampare tutta la time series sul radar plot, proprio tutta.

Slide 59

L’inizio e la fine del radar plot corrispondono rispettivamente con l’inizio e la fine dell’intervallo che si intende considerare.

Slide 61

In giallo fa vedere quando il figlio è sveglio ed in blu quando dorme.

Per una time series si può anche usare una ***heatmap*** che fa vedere con un colore che cosa è successo nel tempo.

In questo caso è una heatmap **categorica** (si ha giallo e azzurro), ma si possono anche usare solo le tonalità del giallo o solo le tonalità dell’azzurro.

NELLE TIME SERIES SI RICERCANO DEI PICCHI oppure DEI CICLI.

 ***RAG PLOT***: è una sorta di istogramma in cui il numero di valori presenti viene rappresentato da una “stanghetta”. Tante stanghette vicine diventano chiare.

Slide 62

Dati categorici → dati che sono divisi in classi

Esempio di dataset: si è voluto sapere se l’approvazione o disapprovazione del pagamento fosse dovuta ad una qualche variabile categorica.

Slide 63

Si ha un dataset multivariato.

Si hanno 4 variabili categoriche: job, housing, sposato o meno ed education.

Si possono distribuire tali variabili ed analizzarle.

Ad esempio si fa un bar chart verticale della variabile job.

Era meglio farlo orizzontale per i nomi dei lavori.

Nel bar chart si è visualizzato per ogni tipo di lavoro l’istogramma normalizzato.

Le categorie sono ordinate per ordine alfabetico di default, qui è stato dato un ordine in base all’altezza della colonna.

Percettivamente si vede anche subito qual è la classe più rappresentata e quale meno.

Si tratta di una serie di istogrammi normalizzati rispetto al totale.

Si poteva normalizzare anche rispetto alla classe “pagamento approvato” oppure “pagamento non approvato”.

Cosa significa istogramma normalizzato?

Per ogni categoria si è contato ad esempio per la categoria “*management*” si è contato quanti avessero quel lavoro, si è ottenuto ad esempio 50; se si hanno 100 campioni bisogna fare 50/100.

Normaizzato vuol dire che ogni conta che si fa, bisogna divdiderla per il totale.

Ciò vuol dire che si sta facendo una stima della probabilità: tra tutti i campioni che si hanno qual è la probabilità di trovare un lavoratore in *management* con pagamento accettato (barra verde in corrispondenza di *management*)?

Così si riesce a capire come si distribuisce la variabile categorica singola.

Slide 64

Ciò che piace fare per ciascuna variabile categorica è un ***pareto chart***.

Non solo si fa un bar chart verticale, ma si fanno partire delle linee che sono una distribuzione **cumulativa**.

Di tutte le probabilità ottenute nella slide precedente si può fare la distribuzione cumulativa, che dice qual è la probabilità di avere un valore minore o uguale ad un certo valore.

Dato l’ordinamento scelto si può fare quindi la probabilità cumulativa.

Ad esempio si può voler capire quando si raggiunge il 50% di pagamento non accettato, quali professioni sono considerate?

La distribuzione cumulativa permette di considerare tutti i valori che sono minori di un certo valore.

PARETO CHART → rappresenta l'importanza delle differenze causate da un certo fenomeno. Contiene al suo interno un bar chart ed un line chart.

Il ***pareto chart*** consente di vedere subito quali sono le catgorie che generano più o meno problemi.

Tendenzialmente è utile quindi vedere sia l’istogramma normalizzato (che è la stima di probabilità) che la distribuzione cumulativa.

Slide 65

Quando si hanno classi per cui si hanno pochissimi campioni, bisognerebbe aggregare tali classi.

Ad esempio si possono aggregare coloro che non hanno lavoro.

Slide 67

Una volta studiato le variabili come si vede la loro combinazione?

Si hanno due classi: rossa (pagamento rilasciato) e verde (non rilasciato).

Si vuole vedere come si distribuiscono i campioni della classe; per tutte le variabili si crea un cammino che descrive il punto (il singolo lavoro in questo caso).

Ad esempio ci sono lavoratori in *admin* che erano single, che non avevano la casa (no housing), avevano education e avevano ottenuto il pagamento.

Per ogni lavoro si potrebbe utilizzare un colore diverso, in modo da vedere bene i diversi cammini.

I ***parallel sets*** dicono come sono distribuiti i punti per classe, fa vedere i diversi cammini.

Slide 68

Un altro modo per visualizzare i dati categorici multivariati è utilizzando i glifi (*glyphs*).

Ad esempio si può voler usare questi omini per visualizzare i campioni.

Slide 69

Si possono aggiungere poi delle informazioni:

la larghezza della testa può indicare la tipologia di lavoro ad esempio.

Con i glifi si possono visualizzare anche delle variabili continue → se si fa un corpo più grosso vuol dire che si ha un conto elevato in banca.

Si utilizzano diverse forme e colori per identificare diverse categorie.

Slide 70

Si includono i glifi in un box (ovviamente se si hanno le label): si fanno vedere in un box i glifi che hanno ottenuto il pagamento e nell’altro quelli che non lo hanno ottenuto.

Slide 71

Le *Chernoff faces* sono dei glifi famosi.

***Chernoff*** codifica i diversi punti/le informazioni utilizzando delle determinate caratteristiche del volto, in modo che le differenze fossero lampantemente visibili.

Slide 72

Altrimenti si usano *whiskers* e *stars*.

Tutti gli assi rappresentano il valore assunto da una delle variabili categoriche.

A volte i *whiskers* vengono congiunti per dare una forma ai dati categorici.

Slide 73

Anche per i dati multivariati categorici si possono utilizzare le heatmap.

La media intesa come la media di quella classe per quella categoria.

Visualizzazioni di questo tipo creano dei problemi:

1. Tendenzialmente la media non si visualizza col nero – lo associamo al negativo, meglio grigio.
2. Se un *color blind* vede questi colori non distingue verde e rosso.

Slide 74

Sono state utilizzate tonalità che i *color blind* vedono ed è stato utilizzato il grigio per la media.