INTRODUZIONE ALLA DATASCIENCE

• Dati strutturati (organizzati): sono dati tratti da osservazioni di caratteristiche, normalmente organizzati in formato tabulare (righe e colonne) • Esempio: osservazioni scientifiche registrate da ricercatori che vengono conservate in modo molto ordinato

• Dati non strutturati (non organizzati): dati che esistono come entità libere e che non seguono alcuna organizzazione standard o gerarchia • Esempi: dati che hanno una natura testuale (es. File log dei server, post di Twitter); sequenze genetiche.

I dati strutturati sono considerati più facili da elaborare e analizzare, circa il 90% dei dati in circolazione è NON strutturato e servono tecniche di pre-analisi e pre-elaborazione (preprocessing) per dare una struttura ai dati non strutturati.

• Dati quantitativi: dati che possono essere descritti tramite numeri e su cui è possibile /ha senso eseguire operazioni matematiche

• Dati qualitativi: tutto il resto, di solito è possible descriverli usando linguaggio naturale

Domande tipiche quando studiamo un dato quantitativo:

- Qual è il valore medio?
- Se il tempo è un fattore, questa quantità cresce o decresce con il trascorrere del tempo?
- Esiste una soglia oltre la quale il valore potrebbe diventare critico?

I dati quantitativi si dividono in:

Dati discreti: possono essere contati e possono assumere solo determinati valori (Esempio: numero di clienti di un caffè)

Dati continui: devono essere misurati e possono assumere una gamma infinita di valori (Esempi: peso o statura di una persona, tempo, temperatura)

Domande tipiche quando studiamo un dato qualitativo:

- Quale valore è più (o meno) frequente?
- Quanti valori univoci esistono?
- Quali sono i valori univoci?

Una specifica colonna del nostro dataset (che noi chiameremo caratteristica) può essere suddivisa in 4 livelli:

- Livello Nominale: che comprende tipicamente dati qualitativi e su cui possiamo eseguire: uguaglianza, appartenenza ad un insieme e calcolo della moda (Esempi: nazionalità, specie, classe di mammiferi…)
- Livello Ordinale: dati su cui è possibile definire strategie per collocare un'osservazione prima di un'altra e su cui possiamo eseguire: tutte le operazioni del livello precedente, ordinamento, confronto, calcolo del «centro» dei dati con la mediana (Esempio: indice di gradimento da 1 a 10)

- Livello degli Intervalli: dati esprimibili attraverso metodi quantificabili e su cui è possibile eseguire formule matematiche su cui possiamo effettuare: tutte le operazioni del livello precedente, somma, sottrazione, calcolo del «centro» dei dati con la media (Esempio: temperatura, se in Texas ci sono 37° e ad Istambul ci sono 27° significa che in Texas ci sono 10° in più)
- Livello dei Rapporti: Contiene tutti i livelli precedenti, ha senso calcolare anche moltiplicazioni e divisioni; hanno un punto iniziale naturale o uno zero naturale, ma anche una restrizione: i valori devono essere non negativi. (Esempio: il denaro depositato in banca si colloca al livello dei rapporti. È possibile avere «niente denaro sul conto» (zero naturale) ed è sensato dire che 200000 euro sono «il doppio» di 100000 euro)

Moda: In statistica, la moda (o norma) di una distribuzione di frequenza X è la modalità (o la classe di modalità) caratterizzata dalla massima frequenza. In altre parole, è il valore che compare più frequentemente.

Mediana: Data una distribuzione di un carattere quantitativo oppure qualitativo ordinabile, si definisce la mediana come il valore/modalità assunto dalle unità statistiche che si trovano nel mezzo della distribuzione.

Media: la media viene calcolata sommando tutti i valori a disposizione e dividendo il risultato per il numero complessivo dei dati.

Con preparazione del dato intendiamo:

Pulizia (Data Cleaning): I dati così come sono possono contenere una moltitudine di errori che vanno dai comuni typo (errori di battitura) a valori nulli, mancanti, non significativi eccetera. Quindi in sostanza con dato sporco intendiamo:

Incompleto: mancano valori (“”; null; …)

Rumoroso: contiene errori di grammatica, typo, di traduzione, o valori non compatibili con il tipo (esempio salario=-10)

Inconsistente: contiene discrepanze (esempio age=42, birthday=“03/07/1997”)

Gli errori e le inconsistenze possono essere accidentalmente introdotti in tante fasi della manipolazione dei dati, dalla raccolta, allo storage, alla trasmissione, trasformazione, integrazione… Possono essere dovuti a dati incompleti alla fonte, errori umani, incosistenze tra fonti… La fase di Data Cleaning deve essere portata a termine sempre tenendo bene a mente gli obbiettivi analitici (quali dati ci interessano per le nostre analisi statistiche e quali possiamo “buttare”)

Integrazione: Ovvero fondere le informazioni contenute in più tabelle in un unico dataset. Spesso i dati su un particolare fenomeno arrivano separati in più tabelle (Esempio: la tabella Titoli_netflix contiene tutte le info sui film di netflix mentre la tabella crediti contiene le informazioni sul cast di ogni film).

Con la parte di Data Exploration effettuiamo tutte quelle analisi dei dati che possono servire per comprendere e studiare i dati (osservando valori come moda, mediana, media, deviazione standard…)

Le statistiche su cui basiamo le nostre visualizzazioni sono:

Frequenza e moda: Si riferisce ad un possibile valore assunto da un attributo categorico, e rappresenta la percentuale di volte che quel valore appare nel data set. Dato un attributo categorico X, che può assumere valori $\{X_{1...}, X_{k...}, X_n\}$ ed un insieme M di dati, la frequenza di ogni valore V_1 è:

$$\text{frequenza}(V_1) = (\text{numero di dati con valore dell'attributo } V_1)/M$$

Percentile: Dato un attributo ordinale o continuo X, ed un valore P compreso tra 0 e 100, il p-esimo percentile X_p è un valore di X tale che il P% dei valori osservati di X sia più basso di X_p%

Media: la somma di tutti i valori assunti da un attributo X divisa la popolazione di X

Mediana: dati una serie di valori $\{X_{1...}, X_{k...}, X_n\}$ la mediana è definita come l'elemento in posizione $X_{(n/2)}$

Indici di Dispersione: ci dicono se i valori di un certo attributo sono "sparpagliati" tra il minimo ed il massimo oppure concentrate intorno ad un valore. Possono essere ad intervallo (più semplici) o a varianza (più informativa). Tuttavia dato che entrambi questi metodi risentono degli outliers (pochi elementi con valori molto diversi dalla maggior parte degli altri) esistono soluzioni più efficaci:

- Deviazione media assoluta (AAD)
- Deviazione mediana assoluta (MAD)
- Intervallo interquartile

Finora abbiamo visto la varianza di singoli elementi ma è ovviamente più interessante osservare la covarianza, ovvero quanto due elementi variano all'unisono. Essa viene rappresentata con una matrice I x J dove per ogni coppia di attributi X_{ij} abbiamo un dato valore della covarianza. Una quantità più descrittiva tuttavia è la correlazione. È sempre una matrice I x J ma sulla diagonale ci sono tutti valori posti a 1, più il valore di un attributo X_{ij} si avvicina a 1 più i due attributi sono correlati (il contrario se il valore si avvicina a -1)

Per quanto riguarda la visualizzazione dei dati, che faremo attraverso grafici e visualizzazioni OLAP, dobbiamo osservare queste caratteristiche:

• Apprehension: Capacità di percepire correttamente le relazioni tra variabili

• Clarity: Capacità di distinguere visivamente tutti gli elementi di un grafico

• Consistency: capacità di interpretare un grafico per confronto (similarità) con grafici precedenti

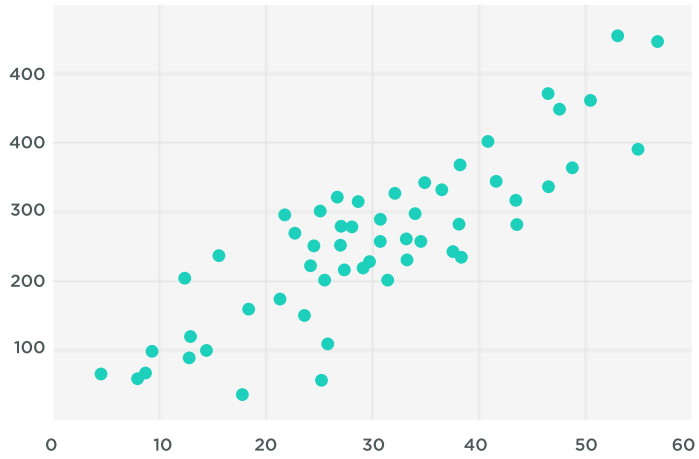
• Efficiency: capacità di rappresentare una relazione anche complessa in modo semplice

• Necessity: la necessità che si ha di usare il grafico (ci sono modi migliori per rappresentare la stessa informazione?)

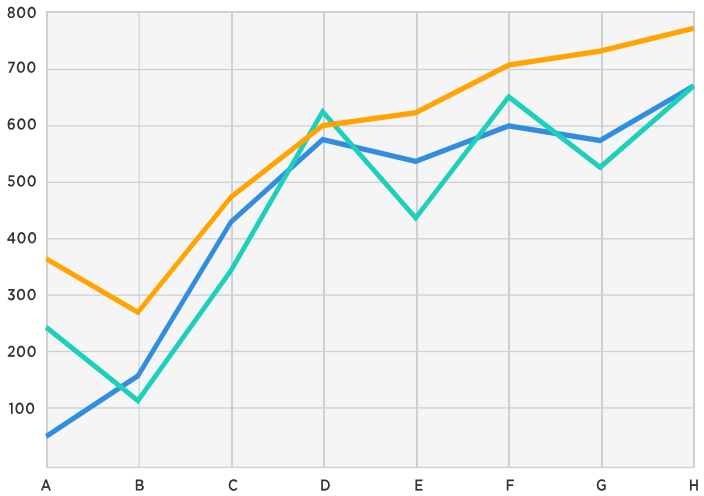
• Truthfulness: capacità di determinare il valore rappresentato da ogni elemento del grafico osservando la sua magnitude relativamente ad una scala implicita o esplicita

Per quanto riguarda i grafici per la rappresentazione ne esistono di diversi tipi:

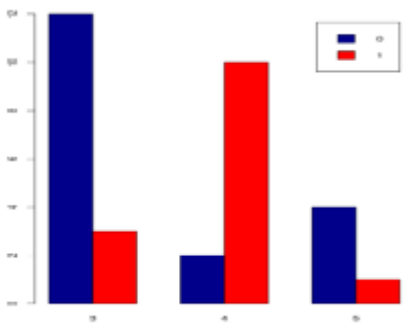
Scatter plot (grafico a dispersione): una rappresentazione dei punti su un piano cartesiano, l'obiettivo è quello di evidenziare visivamente relazioni esistenti tra variabili e se possibile rivelarne una correlazione. Viene usato per variabili quantitative



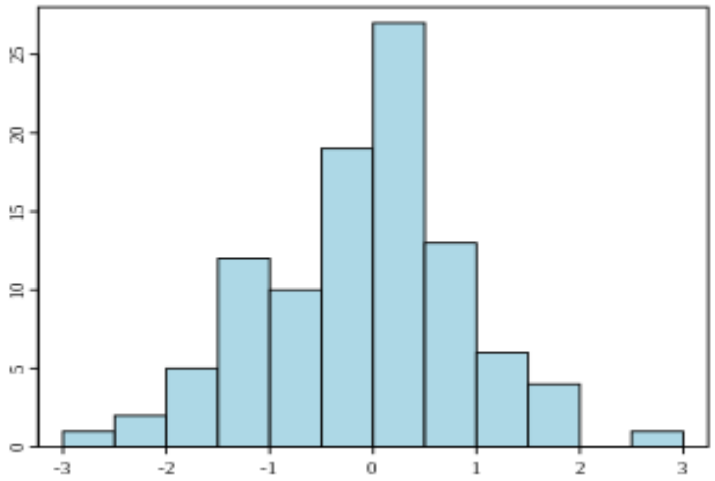
line plot (grafico a linea): vengono di solito usati per mostrare le variazioni nelle variabili con il trascorrere del tempo. Viene usato per variabili quantitative



Diagrammi a barre: Si utilizzano quando dobbiamo confrontare le variabili di vari gruppi, in genere sull'asse x troviamo una variabile categorica mentre sull'asse y una variabile quantitativa o un conteggio

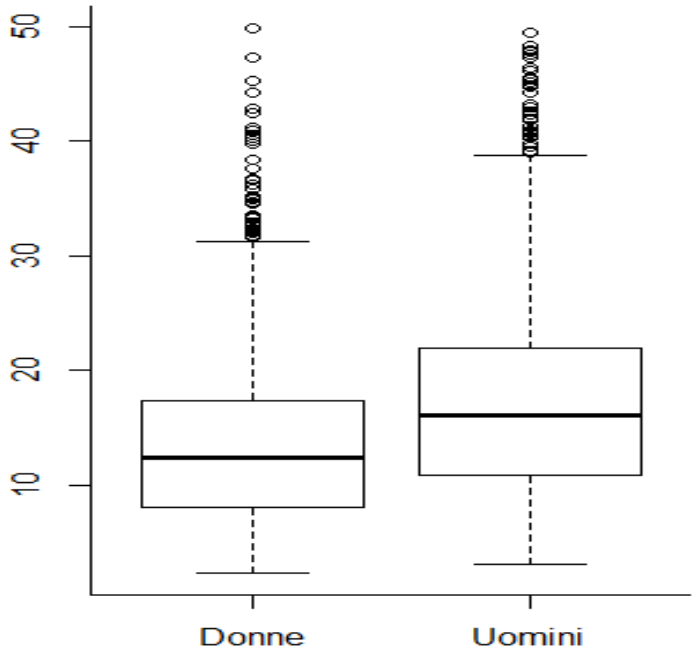


Istogrammi: Servono per rappresentare e visualizzare la distribuzione di frequenza di una variabile quantitativa, raggruppando i dati in intervalli (bin) equidistanti. Possono anche essere bidimensionali

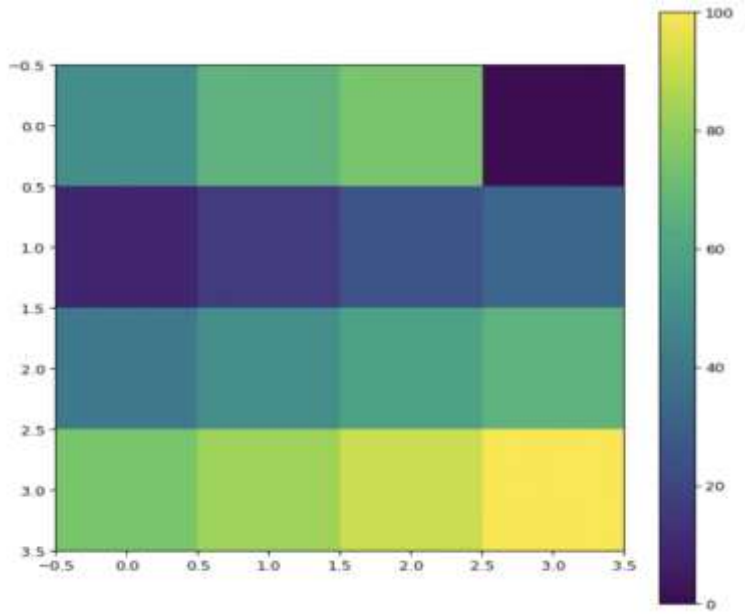


Grafici Box-plot: Vengono utilizzati per mostrare una distribuzione di valori e mettono in evidenza cinque diverse quantità

- Il valore minimo
- Il primo quartile → Valore che separa il 25% di valori più bassi da tutti gli altri → è detto anche 25esimo percentile
- La mediana → è il secondo quartile
- Il terzo quartile → Valore che separa il 25% di valori più alti da tutti gli altri → è detto anche 75esimo percentile
- Il valore massimo



Matrici: Utili quando dobbiamo rappresentazione similarità/distanze tra dati, come nel caso della covarianza o della correlazione



Quando sospettiamo che la dimensione dei dati sia troppo alta rispetto alla loro quantità possiamo affidarci a tecniche di riduzione della dimensionalità, che hanno ulteriori effetti positivi

• Ci permettono di poter visualizzare i dati

• Ci permettono di interpretare meglio i dati

Per una rappresentazione più efficiente essa dovrà avere 3 principali caratteristiche: Varianza alta, Non correlazione e un buon bilanciamento tra numero di dati e dimensionalità

Dato un insieme di dati esiste un modo per poterli "osservare" in modo migliore ovvero tramite un analisi delle componenti principali (PCA):

• Si tratta di un algoritmo che possiamo applicare a matrici di dati di qualunque dimensione NxD

• Consiste nell'identificare una nuova base le cui componenti catturano quanta più varianza possibile dai dati originali

• Un ingrediente chiave è la decomposizione in valori singolari

Per la PCA dovremmo calcolare la SVD ovvero la sua decomposizione in valori singolari:

Data una matrice X la sua decomposizione in valori singolari è

$$X = U\Sigma V^T$$

Dove:

• U è una matrice ortogonale (se X è una matrice reale)

• Σ è una matrice diagonale, con elementi non negativi in ordine decrescente

• V è una matrice ortogonale (se X è una matrice reale)

U e V sono matrici aventi come colonne vettori ortonormali, detti, rispettivamente, vettori singolari sinistri e destri di X mentre gli elementi diagonali di Σ sono detti valori singolari di X

Con la nostra SVD possiamo creare l'algoritmo della PCA:

• Costruiamo la matrice di covarianza dei dati: $C = X^T X$

• Calcoliamo la sua SVD: $X = U\Sigma V^T$

• Identifichiamo la nuova base: le prime K colonne di V, se vogliamo utilizzare K componenti principali

Una versione in python di come dovrebbe risultare la varianza

```
def explainedVariance(eig_vals):
    tot = sum(eig_vals)
    var_exp = [(i / tot)*100 for i in eig_vals]
    cum_var_exp = np.cumsum(var_exp)
    return cum_var_exp
```

Quando analizziamo i dataset sono principalmente riportati in formato tabellare, abbiamo quindi bisogno di un metodo per rappresentarli in modo multidimensionale. Con rappresentazione multidimensionale intendiamo:

Dobbiamo prima identificare quali attributi rappresentano le dimensioni (cioè i parametri) dell'analisi e quali le misure da analizzare

- Gli attributi utilizzati come dimensioni hanno in genere valori discreti
- Il valore delle misure è sempre numerico

Dobbiamo poi calcolare il valore di ogni entry dell'array multidimensionale aggregando i valori (della misura considerata) di tutti gli oggetti che hanno come valori per le dimensioni i valori corrispondenti a quell'entry

In parole povere dobbiamo costruire una visualizzazione in qui i tre attributi scelti assumano uno tra n valori possibili (precedentemente quantizzati) con il quale andiamo a creare quella che è la nostra visualizzazione OLAP (idealmente è come se costruissimo un cubo le cui 3 dimensioni sono i valori dei 3 attributi scelti)

Le operazioni che di visualizzazione che possiamo effettuare tramite la OLAP sono:

• Slicing: significa selezionare un gruppo di celle dalla struttura multidimensionale selezionando uno specifico valore per una dimensione

• Dicing: significa selezionare un sottoinsieme di celle specificando una combinazione di condizioni per le diverse dimensioni

Prima di cominciare la parte di analisi statistica è bene riprendere qualche concetto base. Cominciamo col teorema di Bayes:

$$P(A|B) = [P(A)P(B|A)]/P(B)$$

• A e B sono eventi

• P(A)[P(B)]: la probabilità che si verifichi evento A [B]

• P(B|A): la probabilità che si verifichi evento B sapendo che si è verificato A

• P(A|B): la probabilità che si verifichi evento A sapendo che si è verificato B

In statistica si utilizza il calcolo probabilistico per fare delle predizioni studiando dati che già sono in nostro possesso.

Stime dei punti e intervalli di confidenza:

• Una stima di un punto è una stima di un parametro della popolazione sulla base dei dati di un campione (es la media)

• Esempio: supponiamo che esista una società con 9000 dipendenti e che siamo interessati a determinare la durata delle loro pause

• Non è possibile in generale ottenere la risposta da tutti i dipendenti quindi si considera un campione per calcolare poi, ad esempio, la media

• La media del campione è la nostra stima del punto

Facciamo finta di avere già i valori della popolazione totale nella quale vediamo come le pause seguano una distribuzione binomiale(due picchi distinti nel grafico) con una media di 39.99 minuti. Vogliamo simulare la situazione in cui chiediamo a solo 100 dipendenti scelti a caso la durata solita delle loro pause quindi campioniamo a caso 100 risposte dalla popolazione

Per effettuare questo che è un campionamento si deve: campionare N volte un insieme di M campioni e costruire un istogramma delle N stime calcolate (es. media). Torniamo all'esempio precedente: campioniamo 500 volte un insieme di 100 durate della pausa. Grazie al teorema centrale del limite, se aumentiamo il numero di campioni, la distribuzione delle stime approssima una distribuzione normale, in più se i dati sono «abbastanza», la media della distribuzione si avvicina alla media della popolazione! (la media misurata campionando = 40.01194 minuti)

Intervalli di confidenza:

• Spesso non è facile (a volte è impossibile) ottenere delle stime abbastanza precise anche da campionamenti della popolazione

• In questi casi è opportuno usare il concetto di intervallo di confidenza, un intervallo di valori basato su una stima che sappiamo contenere il vero parametro della popolazione con un certo grado di confidenza

• Importante: il livello di confidenza rappresenta la frequenza con cui la risposta ottenuta è accurata (probabilità % che l'intervallo contenga il valore esatto) Esempio: per avere il 95% di probabilità di catturare il vero parametro della popolazione usando la stima, dobbiamo impostare il livello di confidenza a 95%

Usiamo il concetto di livello di confidenza per parlare della verifica delle ipotesi statistiche. Una verifica delle ipotesi è un test statistico per valutare se possiamo presumere che una determinata condizione sia vera per l'intera popolazione dato un campione limitato dei dati. Il test ci dice se possiamo accettare l'ipotesi o rigettarla. Una verifica delle ipotesi di solito esamina due diverse ipotesi su una popolazione:

- Ipotesi nulla à Ipotesi da verificare
- Ipotesi alternativa

Usiamo un valore p (p-value) che si basa sul livello di significatività per giungere alla conclusione del test (legato al concetto di confidenza). Ora come conduciamo la verifica delle ipotesi?

1. Specificare l'ipotesi
    - Formuliamo le due ipotesi: quella nulla e quella alternativa
    - Usiamo la notazione H0 e Ha
2. Determinare le dimensioni del campione per il test
3. Scegliere un livello di significatività
    - Di solito fissato a 0.05
4. Raccogliere i dati
5. Decidere se rigettare o accettare l'ipotesi nulla (dipende dal tipo di test…)

T-TEST

Il t-test per un campione è un test statistico per determinare se un campione di dati numerici (quantitativi) differisce in modo significativo da un altro dataset (come la popolazione o un altro campione). Rimaniamo sul nostro esempio precedente, vogliamo vedere se la media della durata della pausa di un reparto differisce in modo statisticamente rilevante dalla media di un altro campione (in questo caso la media della popolazione totale. Seguiamo i 5 passaggi scritti sopra e dal valore del p-value (la probabilità che i dati osservati si comportino in questo modo) risultante decidiamo se rigettare o meno l'ipotesi nulla H0.

Esistono due condizioni importanti per questo t-test ovvero:

- La distribuzione della popolazione deve essere normale e il campione deve essere ampio (n>=30)
- La dimensione della popolazione deve essere almeno 10 volte superiore a quella del campione (10n<N) → Questo garantisce che il campione sia tratto in modo indipendente

Quindi con valori del p-value inferiori al solito 0.05 (il livello di significatività) la probabilità che i dati si comportino come ipotizzato da H0 è quasi nulla, possiamo quindi rigettarla a favore dell'ipotesi alternativa Ha.

Test chi-quadrato dell'idoneità

- Lavora su dati QUALITATIVI ragionando in termini di conteggi
- Si usa quando
    - Vogliamo analizzare una variabile categorica da una popolazione
    - Vogliamo determinare se una variabile segue una certa distribuzione, specificata oppure attesa

Confrontiamo quanto osservato con quanto previsto. Anche qui abbiamo dei requisiti da rispettare:

- Tutti i conteggi previsti devono essere almeno 5
- Le singole osservazioni devono essere indipendenti e le dimensioni della popolazione devono essere almeno 10 volte quelle del campione

Immaginiamo di volere verificare se le distribuzioni dei gusti di caramelle nei sacchetti commercializzati è quella attesa. Ogni sacchetto contiene 100 caramelle che dovrebbero essere equamente distribuite tra i 5 gruppi. Usiamo un campione di 10 sacchetti e contiamo i gusti presenti all'interno di ciascun sacchetto

Dalla conta ricaviamo il numero effettivo di caramelle per sacchetto ed effettuiamo le nostre ipotesi:

- H0 : la proporzione tra i gusti in ogni sacchetto di caramelle è la stessa
- Ha : la proporzione tra i gusti in ogni sacchetto di caramelle NON è la stessa
- Eseguiamo un test con livello di significatività pari a 0.05
- I gradi di libertà sono 5-1=4

$$X = \Sigma\ [(\text{osservato} - \text{atteso})^2]/\text{atteso}$$

- Il valore del test risulta essere 52,75

Per trarre una conclusione avremo bisogno di confrontare il valore ottenuto dal test ed il valore della distribuzione del chi-quadrato corrispondente a livello di confidenza e gradi di libertà del nostro problema. Cosa può succedere:

- La statistica di test è inferiore al valore del chi-quadrato → non possiamo rifiutare ipotesi nulla
- La statistica di test è superiore al valore del chi-quadrato → rifiutiamo ipotesi nulla

Per controllare il valore della distribuzione occorre cercare il valore corrispondente al nostro livello di confidenza (95%) con il gradi di libertà del nostro esempio (4) nella tabella del chi-quadrato(https://it.wikipedia.org/wiki/Distribuzione_chi_quadrato) e osservare se il valore che abbiamo ottenuto è maggiore o minore del suddetto. (nel nostro caso 52,75 > 9.488 e possiamo rigettare H0)

[Rivedi le slide prima dell'esame]

Metodi predittivi

Con il termine Machine Learning ci riferiamo ad una classe di metodi in grado di imparare da esempi invece di essere esplicitamente programmati a fare qualcosa

Due macro-tipologie

- Machine Learning supervisionato (simula l'imparare con un insegnante)
- Machine Learning non supervisionato (simula l'imparare senza un insegnante)

Machine Learning supervisionato: abbiamo un insieme $S = \{(x_1,y_1), (x_2,y_2)\ldots, (x_n,y_n)\}$ che sarà il nostro training set dalla quale l nostro algoritmo imparerà. Ogni coppia rappresenta un possibile (input-output) del nostro problema e Lo scopo del Machine Learning supervisionato è stimare una funzione f(x) tale che:

$$f(x_k) = y_k$$

Cosa sia y dipende dal tipo di problema che stiamo studiando, mentre la funzione che cerchiamo di stimare deve avere due proprietà importanti:

- Capacità di rappresentare i dati di training (fitting)
- La capacità di generalizzare ai dati che avremo a disposizione in futuro (stability/generalization)

Ogni algoritmo di Machine Learning ambisce a trovare delle soluzioni che siano un compromesso tra queste due proprietà, è tale compromesso che ci garantisce la giusta capacità predittiva

Machine Learning supervisionato

Regressione lineare

Immaginiamo di dover valutare la bontà di una funzione stimata $\hat{f}(x)$, tale funzione ci fornirà una predizione su ogni input del training set: $f(x_k) = \hat{y}_k$ A questo punto possiamo valutare quanto la predizione si discosta dalla verità (ecco dove usiamo la supervisione, ossia l'insegnante!) calcolando i residui:

$$R = \Sigma(\hat{y}_k - y_k)^2$$

Il processo di identificare, tra tante soluzioni possibili, la soluzione migliore a partire dai dati a disposizione prende il nome di fase di training di un metodo di Machine Learning. In cosa consiste il training del metodo di regressione lineare? Dobbiamo trovare i parametri della f migliore, ossia i coefficienti.

Alla fine del processo di training, è importante avere una misura della bontà della funzione stimata (la migliore, che abbiamo trovato con la fase di training stessa) su dati nuovi, ossia dati non utilizzati durante il training → questo consente di verificare la stabilità/generalizzazione. E' buona pratica tenere da parte un po' di dati da usare solo in questa fase! La metrica che viene di solito utilizzata in regressione è l'errore quadratico medio

MSE = $1/N\ \Sigma(\hat{y}_k - y_k)^2$

RMSE = $\text{sqrt}(1/N\ \Sigma(\hat{y}_k - y_k)^2)$

Regressione logistica

Si tratta di una generalizzazione del modello di regressione lineare ai problemi di classificazione, con la regressione logistica prevediamo la probabilità che il dato appartenga ad una certa classe. La funzione che stimiamo in questo caso può assumere valori tra [0,1], come arriviamo a questa funzione? Introduciamo il concetto di odds: l'odds di un evento è il rapporto tra la sua probabilità p e la probabilità che non accada, cioè 1 − p (evento complementare)

Classificazione Bayesiana naïve

Supponiamo che H sia l'ipotesi relativa a determinati dati e che D siano i dati disponibili possiamo usare il teorema di Bayes per ottenere la probabilità che la nostra ipotesi sia corretta sulla base dei dati disponibili

$$P(H|D) = [P(H)P(D|H)]/P(D)$$

Qual è la probabilità che la mia ipotesi sia vera sulla base dei dati che ho?

- P(H) è la probabilità dell'ipotesi (probabilità a priori)
- P(H|D) è la probabilità dell'ipotesi dopo aver osservato i dati (probabilità a posteriori)
- P(D|H) è la probabilità dei dati sotto l'ipotesi data (probabilità)
- P(D) è la probabilità dei dati sotto qualsiasi ipotesi (costante di normalizzazione)

Possiamo usare il teorema di Bayes per classificare un dato come appartenente o no ad una certa classe calcolando

$$P(\text{classe}|x) = [P(\text{classe})P(x|\text{classe})]/P(x)$$

Machine Learning NON supervisionato

In questo caso gli algoritmi di machine learning si dividono in due approcci differenti, dato che non vengono usati training set:

Clustering:

L'obiettivo degli algoritmi di clustering è di individuare strutture (gruppi di dati "coerenti" rispetto ad una qualche misura) all'interno dei dati. Con cluster intendiamo un gruppo di dati che si "comporta in modo analogo" mentre con centroide il "centro" del cluster (ad esempio il punto medio, o centroide). Un primo esempio di algoritmo di clustering è il k-means, che si sviluppa in questi passaggi:

1. Scegliere k centroidi iniziali, ovvero elementi mediani tra gruppi di valori contingenti (vedi grafici slide)

2. Per ogni punto: o Assegnare il punto al centroide più vicino

3. Per ogni centroide: o Aggiornare la posizione del centroide

4. Ripetere i passi 2 e 3 fino a raggiungere un criterio di arresto

Idealmente questo algoritmo continua a generare centroidi (casualmente) e gli associa i valori più vicini fino a che non crea classi d valori distinte tramite le quali può confrontare i dati successivi. Vediamo ora un esempio di implementazione di un algoritmo k-means (python);

```
def K-Means(X, centers, maxiter):
        # X: n x d
        # centers : k x d
        n, d = X.shape
        k = centers.shape[0]
        for i in range(maxiter):
                # Compute Squared Euclidean distance (i.e. the squared
                # distance) between each cluster centre and each observation
                dist = all_distances(X, centers)
                # Assign data to clusters:
                # for each point, find the closest center in terms of euclidean
                # distance
                c_ass = np.argmin(dist, axis=1)
                # Update cluster center
                for c in range(k):
                        centers[c] = np.mean(X[c_ass == c], axis=0)
        return c_ass, centers
```

Siccome la scelta dei centroidi è random non sempre i risultati sono soddisfacenti, infatti occorre usare dei "metodi" per controllare l'attendibilità dei risultati ottenuti, uno di questi è il coefficiente di silhouette:

prendiamo b = distanza media tra i cluster e b = distanza media tra gli elementi di un cluster

$$SC = (b - a)/max(a,b)$$

I valori di SC vanno da -1 (modello molto scarso) a 1 (modello eccellente), per una maggiore attendibilità del modello è fondamentale effettuare una standardizzazione dei dati che forniamo al nostro algoritmo.

Riduzione della dimensionalità:

Quando i dati vengono rappresentati con un numero di caratteristiche troppo alto rispetto al numero di dati potremmo incontrare la cosiddetta curse of dimensionality. Maggiore è il numero di caratteristiche usate per la rappresentazione di un dato, più i dati stessi risultano essere distanti gli uni dagli altri. Quando le caratteristiche sono troppe rispetto alle reali necessità rischiamo di non migliorare la capacità descrittiva della rappresentazione, e anzi peggiorare i risultati. Per curse of dimensionality intendiamo che: Quando la dimensionalità (quante caratteristiche vengo usate per descrivere un singolo dato) dei dati (D) cresce, il volume dello spazio dei dati cresce velocemente ed i dati diventano presto sparsi… Perchè i risultati siano affidabili il numero di dati (N) deve crescere esponenenzialmente con la loro dimensionalità. Quando sospettiamo che la dimensione dei dati sia troppo alta rispetto alla loro quantità possiamo affidarci a tecniche di riduzione della dimensionalità, che hanno ulteriori effetti positivi

• Ci permettono di poter visualizzare i dati

• Ci permettono di intepretare meglio i dati

Se avessimo la possibilità di progettare una rappresentazione quali sono le proprietà fondamentali sarebbero

- Varianza alta: caratteristiche con varianza alta contengono "molto segnale"
- Non correlazione: caratteristiche correlate le une alle altre sono ridondanti e poco informative
- Non troppe: deve essere sempre un buon bilanciamento tra numero di dati e dimensionalità

Un modo per ridurre la dimensionalità dei dati è quello di usare la PCA, che abbiamo già visto nel caso della visualizzazione OLAP.