DATA MINING

La **data science** è l'intersezione delle competenze di informatica, settore specifico, matematica e statistica. Il **Data mining** (per convenzione) non richiede l'esperienza nel settore specifico: si tratta della knowledge discovery in databases.
Il punto chiave è presentare **conoscenza**, non dati.

Uno dei primi problemi del data mining è diventata l'esplosione dei dati, ovvero la dimensione impressionante della collezione di informazioni. La quantità dei dati memorizzata sui supporti informatici è in continuo aumento.
L'hardware diventa ogni giorno più potente e meno costoso.

Big data: acquisiti molto velocemente, contengono varietà di informazioni, valore informazionale molto alto, dati generalmente sporchi, volume molto elevato

Un **pattern** è una rappresentazione sintetica e ricca di semantica di un insieme di dati; esprime in genere un modello ricorrente nei dati, ma può anche esprimere un modello eccezionale. (per esempio posso raggruppare i clienti in base al comportamento d'acquisto, tramite le tessere fedeltà)
Un pattern deve essere:

- **Valido** sui dati con un certo grado di confidenza
- **Comprensibile** dal punto di vista sintattico e semantico, affinché l'utente lo possa interpretare
- **Precedentemente sconosciuto e potenzialmente utile**, affinché l'utente possa intraprendere azioni di conseguenza

Fasi del **processo** di data mining:

- **definizione e comprensione del dominio applicativo**: individuare le effettive problematiche di business e gli obiettivi da realizzare
- **creazione di un target dataset:** selezione di un sottoinsieme di variabili e di dati o di un campione di dati
- **data cleaning e pre-processing**: operazioni per attenuare il rumore dei dati, o degli outlier, selezione delle informazioni necessarie per generare il modello; decisione sul trattamento dei campi mancanti o incompleti, dei dati rari, sulla definizione della storicità e dell'aggiornamento dei dati; aggiunta di variabili derivate e indicatori che hanno valori ricavabili da dati già esistenti.
- **data reduction e projection**: definizione della modalità di rappresentazioni dei dati secondo gli obiettivi posti, utilizzo di metodi per ridurre il numero delle variabili
- **scelta del ruolo dei sistemi di data mining per l'analisi**: utilizzo dei sistemi di data mining per classificazione, regressione, clusterizzazione, etc.
- **scelta del o degli algoritmi di data mining:** selezione dei metodi per la ricerca di pattern, decidendo quali modelli o parametri possono essere più appropriati, l'integrazione dei metodi di data mining scelti con l'intero processo di scoperta della conoscenza.
- **data mining**: ricerca di modelli d'interesse per l'utente, con raffinamenti successivi, presentati secondo definite modalità di rappresentazione (classificazione, alberi di decisione, regressione, cluster analysis)

- **interpretazione dei modelli identificativi:** analisi e verifica dei risultati con possibile retroazione ai punti precedenti per ulteriori iterazioni al fine di migliorare ulteriormente l'efficacia dei modelli trovati
- **consolidamento della conoscenza scoperta:** integrazione della conoscenza e valutazione della performance di sistema, mettendo a confronto i risultati con l'effettiva realtà dei fatti e produzione della documentazione finale per l'utente

Componenti del data mining

- **classificazione**: apprendimento di una funzione per mappare oggetti in un insieme predefinito di classi
- **Regressione**: apprendimento di una funzione per mappare un oggetto in un valore reale
- **Clustering:** identificazione di una collezione di gruppi di oggetti simili
- **apprendimento di dipendenze e associazioni:** identificazione di dipendenze significative tra gli attributi dei dati
- **apprendimento di regole e sommarizzazione**: individuazione di una descrizione compatta di un insieme o sottoinsieme di dati

Training set: un insieme di punti sperimentali (dati) usati per la fase di apprendimento del metodi di Data Mining. La variabile dipendente è nota
Test set: questo dataset è utilizzato solo per l'assessment finale del metodo di classificazione o regressione, Le osservazioni contenute in questo dataset non sono state utilizzate nella prima fase.
Generalmente divido il mio dataset in un 80% per il training set e in un 20% nel test set.

Un metodo di data mining si dice **supervisionato** quando i dati di "addestramento" comprendono un insieme di esempi sulla caratteristica di interesse (variabie risposta). e.g.: classificazione/regressione
Un metodo si dice invece **non supervisionato** quando i dati di addestramento non comprendono dati sulla caratteristica di interesse (variabile risposta). e.g.: clustering

Il processo di data mining

1) Business understanding: capire gli obiettivi del progetto dal punto di vista dell'utente e tradurre il problema dell'utente in un problema di data mining
2) Data understanding: raccolta preliminare dei dati finalizzata a identificare problemi di qualità e a svolgere analisi preliminari che permettono di identificarne le caratteristiche salienti
3) Data preparation: comprende tutte le attività necessarie a creare il dataset finale
4) Modeling: diverse tecniche di data mining sono applicate al dataset per costruire il modello che sia il più possibile accurato
5) Evaluation: i modelli delle fasi precedenti sono analizzati al fine di verificare che siano sufficientemente precisi per rispondere agli obiettivi dell'utente (serviranno quindi delle metriche per valutare i modelli)
6) Deployment: il modello costruito e la conoscenza acquisita devono essere messi a disposizione degli utenti.

Data preparation
Dopo la raccolta dei dati, vanno controllati diversi aspetti delle variabili disponibili, come:

- presenza di valori estremi (creando un boxplot e verificando la presenza di outliers)
- range di variazione delle variabili; questo perchè se abbiamo variabili in range troppo diversi, la loro influenza in un modello di regressione lineare varia in base al range e non in base all'importanza a spiegare la nostra variabile risposta. Per ovviare questo problema esistono vari tipi di normalizzazione/standardizzazione.
- massimo e minimo di ogni variabile. (che ci servono per poter normalizzare la variabile). Il massimo e il minimo vanno calcolati solo sul training set, perchè tutte le scelte che facciamo le facciamo in base ai dati che abbiamo nel training set; valuteremo le nostre scelte poi sul test set.
- Individuazione dei missing values; questo è uno dei problemi più gravi durante l'analisi. Possono esserci per diverse cause: unione di due fonti di dati, fallimenti nelle misurazioni. Il problema più grave è quando i missing values compaiono nella variabile target (in quel caso solitamente si cancellano le osservazioni)
  La struttura dei dati mancanti deve essere casuale, altrimenti è difficile generalizzare i risultati che si ottengono; assumere l'andamento casuale è sempre rischioso (bisogna fare un test)
  Se in un campione abbastanza grande i dati mancanti sono pochi il problema è quasi irrilevante.

MISSING VALUES

1) MCAR: missing completely at random → il processo che ha determinato la non rilevazione è completamente indipendente dal valore mancante e da qualsiasi altra variabile disponibile
2) MAR: missing at random → il processo che ha determinato la non rilevazione è completamente indipendente dal valore mancante ma può dipendere da altre variabili disponibili.
3) NMAR: not missing at random → la probabilità che un valore sia mancante è collegata alla variabile stessa di cui stanno raccogliendo i valori.

Step da seguire in presenza di missing values

1) individuare i missing values e i dati anomali
2) assegnare un valore univoco ai dati mancanti, in modo da distinguerli chiaramente dai valori effettivi (un valore mancante non è per forza una cella vuora, ma può anche essere un carattere speciale o un valore senza senso)

Strategia passiva (deletion methods)
i dati missing vengono ignorati; l'analisi viene effettuata esclusivamente sui dati presenti (complete case approach) in due diversi modi:

- casewise deletion: vengono analizzati solamente casi completi (cancellare le osservazioni con dati mancanti / eliminare le variabili in cui c'è concentrazione di dati mancanti)
- pairwise deletion (l'eliminazione del dato mancante viene considerato in base all'analisi da effettuare)

Con grandi campioni e basse proporzioni di missing (5% o meno) comunemente si procede ad una casewise deletion

Strategia attiva (single imputation methods)
L'obiettivo è quello di sostituire ciascun valore mancante con uno plausibile, stimato sulla base dei valori validi delle variabili complete. Si dovrebbe considerare in modo molto cauto l'utilizzo dell'imputazione, per via del suo potenziale impatto sull'analisi dei dati.

- Imputazione della media / mediana: si inserisce la media/mediana calcolabile a partire dai dati a disposizione. Procedura molto utile in assenza di ogni altro tipo di informazione. Bisogna però considerare che imputando la media riduciamo la variabilità dei dati
- Imputazione della media / mediana condizionata: si inserisce la media calcolata sulla base di gruppi specifici, quindi condizionatamente a tali gruppi. Se la divisione in gruppi non è evidente dai dati, può essere determinata dalla conoscenza a priori del ricercatore. In questo caso diminuirà la varianza entro i gruppi e aumenterà la varianza tra i gruppi
- Imputazione attraverso metodo di regressione: la stima del dato mancante è ottenuta stimando il modello di regressione in cui la variabile con i casi mancanti è usata come variabile dipendente e le altre variabili sono utilizzate come covariate (esclusa la variabile dipendente del modello generale). Il modello è stimato sulla base dei casi completi ed è impiegata per prevedere i casi mancanti. Gli svantaggi di questo metodo sono che si migliora in modo fittizio l'adattamento dei dati, che si riduce la varianza (il valore stimato sarà molto simile alla media), il metodo "funziona" se le variabili utilizzate per la stima sono dei buoni predittori per la variabile con i valori mancanti e le stime ottenute sono utilizzabili se assumono valori nel range dei valori assunti dalla variabile nei casi completi.

Dopo la scelta del metodo per il trattamento dei dati mancanti è molto importante verificarne gli effetti. In genere questa procedura si basa sul confronto delle analisi ottenute rispettivamente sui casi completi e sui casi incompleti. Nell'eventualità vi siano forti differenze il metodo scelto potrebbe essere poco opportuno e condurre a risultati inaffidabili. In tal caso è necessario analizzare i motivi che hanno portato a tali differenze e valutare quale risultato si approssima di più alla realtà oppure riportare entrambi i risultati.

Non esiste una regola per decidere quando eliminare il record (strategia passiva) o correggerlo con un'imputazione (strategia attiva)

Normalizzazione
Normalizzare una variabile si riferisce, solitamente, all'azione di riscalare la variabile in modo tale che i valori siano poi contenuti all'interno dell'intervallo 0 e 1. Questo per riportare tutte le variabili di un dataset all'interno di una scala di variazione comune.

L'obiettivo della normalizzazione è modificare i valori delle variabili in un dataset in modo tale che, come detto, questi valori varino all'interno dello stesso intervallo senza però modificare il range di variazione. Questa operazione è richiesta quando le variabili si definiscono in intervalli diversi.

Per esempio, consideriamo un dataset contenente due variabili: età e reddito. L'età è definita nell'intervallo 0-100 e il reddito può variare tra 0 e 100000. Il reddito è circa 1000 volte più grande della variabile età.

Quando applichiamo determinati modelli, come la regressione lineare, la variabile reddito influenzerà le nostre analisi proprio perché i suoi valori sono molto più grandi rispetto a età. Questo però non vuol dire necessariamente che la variabile reddito sia più influente della variabile età.
La normalizzazione permette quindi di ricondurre le due variabile a un intervallo di definizione comune.

La normalizzazione è una tecnica che può essere utilizzata quando non è nota la distribuzione dei propri dati o la distribuzione non è normale.

Standardizzazione
La standardizzazione si riferisce alla sottrazione di una misura di posizione (della variabile considerata) da una osservazione (della stessa variabile) e successiva divisione per una misura di variabilità. Per esempio, in presenza di una variabile con comportamento normale, la standardizzazione consiste nel sottrarre la media e dividere per la deviazione standard così da ottenere una variabile con media 0 e varianza 1.

Standardizzare quindi trasforma una variabile centrandola nel valore 0 e con deviazione standard pari a 1.
Questo approccio può essere utile quando valutiamo variabili con unità di misura differenti. I motivi sono simili a quelli già presentati per la normalizzazione.
Per esempio, una variabile che varia tra 0 e 1000 avrà maggior peso di una variabile che varia tra 0 e 1. L'applicazione della standardizzazione rende equiparabile la variabilità delle due variabili così da avere un effetto inferiore sulle analisi.

Applicare la standardizzazione prevede che i dati abbiano una distribuzione normale.
Questo non deve essere sempre vero ma la tecnica ottiene migliori risultati sotto tale condizione.

La normalizzazione è una tecnica utile quando vengono utilizzati approcci che non fanno assunzioni sulla distribuzione dei dati, mentre la standardizzazione è utile quando le variabili hanno variabilità diversa e i metodi che si intendono usare assumono la normalità.

In presenza di outlier, queste due tecniche vanno applicate dopo aver eliminato gli outlier dal dataset

CLASSIFICAZIONE

Human perception
L'uomo ha sviluppato competenze relative alla comprensione dell'ambiente e la capacità di agire in relazione agli eventi circostanti:
- riconoscere oggetti e persone
- comprendere parole
- leggere documenti scritti
- distinguere cibo fresco dall'odore

Nel data mining e machine learning si vogliono far apprendere queste capacità alle macchine

Classification può essere visto come lo studio

- dell'ambiente circostante
- di rappresentazioni statistiche di interesse
- di approcci per prendere decisioni sulla base di queste rappresentazioni

Un esempio di classificazione è quello di insegnare a una macchina a distinguere due diverse specie di individui, a partire da caratteristiche oggettive quali possono essere il peso, la lunghezza, la larghezza e la forma. Non sempre questa distinzione è solare
Dobbiamo raccogliere alcuni esemplari (un training set sufficiente a rappresentare l'intera popolazione) e analizziamo con degli istogrammi le distribuzioni marginali di ogni caratteristica scelta (**feature**). Ogni individuo viene quindi rappresentato da un vettore che associa al pesce il valore di ogni feature.
Il decision boundary è quella soglia superata la quale cambio la mia scelta tra un individuo e l'altro.

Numero di features
Verrebbe automatico pensare che un aumento del numero di features migliori sempre i risultati, ma bisogna stare attenti a certe cose:

- evitare informazioni non necessarie
- controllare la correlazione tra features
- stare attenti ai costi di misurazione
- stare attenti al rumore dovuto a nuove osservazioni

Dobbiamo inoltre considerare il costo di commettere errori nel nostro processo decisionale (**cost of different errors**)
In base alle feature disponibili, la complessità del nostro modello cambierà come cambiano anche le linee di confine (**decision boundary**)
Un punto chiave della classificazione è di identificare modelli che prevedono correttamente la classe di appartenenza di una nuova istanza. La parola chiave è **generalizzazione**
Esiste un trade off tra la complessità della regola decisionale e l'affidabilità della previsione.

Overfitting: comprendo la relazione tra i dati ma non la funzione sottostante. Questo rende il mio modello estremamente affidabile sul training set ma probabilmente sbagliato nella previsione di nuove istanze

Il processo di classificazione

- data collection: La parte più lunga di un processo di riconoscimento di un pattern. Dobbiamo innanzitutto capire quante osservazioni sono "abbastanza" per creare un modello
- feature extraction or collection: la conoscenza a priori di colui che compie l'analisi può aiutare in questo caso. Se il nostro modello non sembra essere affidabile, aggiungere nuove feature è una scelta ragionevole
- Model choice: metodi supervisionati e non supervisionati. In questa fase si scelgono i parametri
- training, validation and test: dividiamo il nostro dataset in training set (set di osservazioni in cui la variabile dipendente è nota), validation set (set di osservazioni dove ottimizzare i parametri del modello) e test set (utilizzato unicamente per verificare la capacità di previsione; non è mai utilizzato durante le altre fasi). Compariamo l'effetto su tutti i dati di diversi modelli.

Funzioni obiettivo
Ci sono diverse funzioni obiettivo che possono essere ottimizzate in un processo di classificazione. La più comune è la minimum error rate: la previsione deve minimizzare la percentuale di istanze assegnate al cluster sbagliato

Il problema della dimensionalità
è dimostrato che sopra una certa soglia aggiungere features non migliora il modello selezionato (*curse of dimensionality*)
Le potenziali ragioni per una bassa accuratezza del modello sono la selezione del modello sbagliato e un numero basso di osservazioni rispetto al numero di features. Le possibili soluzioni sono ridurre la dimensionalità e semplificare il modello.

Feature reduction
La dimensionalità può essere ridotta:
- riconsiderando le features
- selezionando un sottoinsieme di features
- combinando le features esistenti

Tutti i modelli di classificazione possono soffrire del problema curse of dimensionality. Buona regola è avere a disposizione un numero di punti sperimentali tale che $n/p > 10$. Più il modello è complesso più sono necessari punti sperimentali
Combinazioni lineari delle features sono spesso utili perchè sono semplici da calcolare e analiticamente trattabili. I metodi lineari proiettano spazi ad alta dimensionalità in spazi più piccoli. Questo ci aiuta a ridurre la dimensionalità dei modelli di classificazione.
Uno dei classici approcci possibili è quello della Principal Component Analysis (PCA): proiezione delle osservazioni (punti sperimentali) da uno spazio p dimensionale a uno spazio (p - k) dimensionale

Feature selection
Un'alternativa alla future reduction è rappresentata da tutti gli approcci di selezione delle variabili atti a individuare sottoinsiemi significativi di features.
Sequential forward selection
- viene selezionata la feature migliore
- vengono formate delle coppie tra la feature migliore e tutte le altre feature, e si sceglie la coppia migliore
- vengono formate delle triplette di feature tra la coppia migliore e una delle rimanenti feature, e si sceglie la tripletta
- si continua il processo fino a scegliere un numero predefinito di features

Sequential backward selection
- la funzione è calcolata su tutte le features
- successivamente, la funzione viene ricalcolata escludendo una feature alla volta (d - 1 features) e viene scartata la feature peggiore
- il processo continua fino a quando si rimane con un numero predefinito di features

La scelta tra feature reduction e feature selection dipende dal dominio di applicazione. Feature selection è meno intensiva computazionalmente e mantiene il significato fisico delle features d'origine. Feature reduction può portare a una maggiore capacità di discriminare tra classi ma si perde il significato fisico delle features d'origine.

Approcci supervisionati e non supervisionati

Approcci supervisionati

- I training data includono sia features di input sia le features di output
- La target feature (variabile risposta o feature di output) è usata nel processo di learning del modello
- La costruzione del dataset di training, validation e test è cruciale
- Questi metodi sono spesso veloci e accurati
- Devono saper generalizzare

Approcci non supervisionati

- Il modello non ha come input la target feature
- Possono essere usati per raggruppare le features di input in classi

CLASSIFICATORE DI BAYES

La teoria delle decisioni basata sull'approccio bayesiano è un approccio statistico fondamentale per quantificare il trade off tra l'assegnazione di una classe e il costo di questa azione.

Esempio di classificazione

Supponiamo di voler dividere le orate dai salmoni con un nastro trasportatore.

Definiamo y come il tipo di pesce (classe, variabile casuale) dove:

$y = y_1$ orata

$y = y_2$ salmone

$P(y_1)$ è la probabilità a priori che il prossimo pesce sia un'orata. Analogamente, $P(y_2)$ è la probabilità a priori che il prossimo pesce sia un salmone. La probabilità a priori rappresenta la nostra conoscenza rispetto al tipo di pesce prima ancora di poterlo osservare.

Il problema ora diventa la scelta del valore di $P(y_1)$ e $P(y_2)$. Assumendo che non esistano altri tipi di pesce, deve valere che $P(y_1) + P(y_2) = 1$ (esclusiva e esaustiva)

Naturalmente, la nostra scelta basata solo sulla conoscenza a priori è y_1 se $P(y_1) > P(y_2)$ e y_2 viceversa. La probabilità di sbagliare è

$P(\text{error}) = \min\{P(y_1), P(y_2)\}$

Assumiamo di voler migliorare la nostra regola decisionale sulla base di una feature x. Assumiamo x come una variabile casuale continua.

Definiamo $P(x \mid y_j)$ come *probabilità condizionata* (ovvero la probabilità di x data la classe di appartenenza del pesce)

$P(x \mid y_1)$ e $P(x \mid y_2)$ descrivono le differenze tra orate e salmoni rispetto alla feature x

Supponiamo ora di conoscere $P(x_j)$ e $P(x \mid y_j)$ per j=1, 2, e il valore della feature x.

Definiamo $P(Y_j|x)$ come la **probabilità a posteriori** (probabilità che un pesce sia y_j dato il valore della feature x. Utilizzando il teorema di Bayes possiamo calcolare la probabilità a posteriori di ogni classe

$$P(y_j|x) = \frac{P(x|y_j)P(y_j)}{P(x)}$$

dove

$$P(x) = \sum_{j=1}^{2} P(x|y_j)P(y_j)$$

Dunque la regola decisionale dopo aver osservato x diventa y_1 se $P(Y_1|x) > P(Y_2|x)$ e y_2 viceversa.
Equivalentemente, la regola decisionale può essere y_1 se $P(x \mid y_1) / P(x \mid y_2) > P(y_2) / P(y_1)$ per ogni x $P(Y_1|x) + P(Y_2|x) = 1$

Classification setting
Training data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Feature vector $X = (X_1, X_2, \ldots, X_p)$
La variabile risposta Y è qualitativo/categorica → $y \in Y = \{1, 2, \ldots, k\}$
Funzione f(x) che assegna un valore Y sulla base di X
f(x) divide lo spazio in un insieme di regioni.
In un problema di classificazione per valutare le performance di un modello, viene comunemente calcolato l'error rate come proporzione della quantità di errori che vengono effettuati nel training set:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

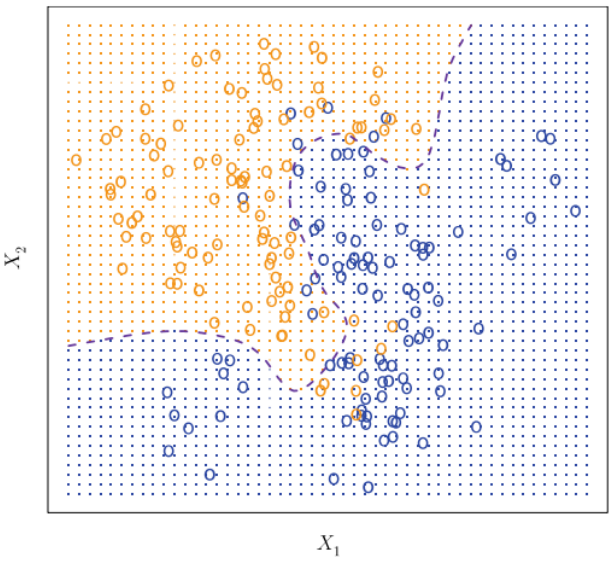
Quando viene calcolato sul training set, lo chiamiamo training error.
è possibile dimostrare che il test error è minimizzato, in media, da un classificatore che assegna ad ogni nuova osservazione la classe più probabile dati i valori delle variabili indipendenti. Per fare ciò, utilizziamo il Classificatore di Bayes, che si basa sulla probabilità a posteriori.

$$Pr(Y = j \mid X = x_0)$$

Assegnamo l'osservazione x_0 alla classe che j per cui questa probabilità è più alta.
Supponendo di avere due classi, classe 1 e classe 2, il classificatore di Bayes assegnerà la nuova osservazione x_0 alla classe 1 se:

$$Pr(Y = 1 \mid X = x_0) > 0.5$$

Bayes decision boundary



Il classificatore di Bayes produce sempre il più basso *test error rate*, che viene chiamato **Bayes error rate**. Viene sempre scelta la classe di appartenenza in base a Pr (Y = j | X = x_0) Il valore più alto determina la classe di appartenenza. In generale l'error rate per una nuova osservazione X = x_0 sarà:

$$1 - \max_j \Pr(Y = j \mid X = x_0)$$

Considerando un certo numero di nuove osservazioni, il Bayes error rate totale è

$$1 - E\left(\max_j \Pr(Y = j|X)\right)$$

Dove il valore atteso è calcolato calcolando la media di tutte le probabilità di tutti i valori in X
Il Bayes error rate è un **errore irriducibile**

BIAS - VARIANCE TRADE-OFF

Consideriamo di osservare una variabile target di tipo quantitativo Y e un numero p di covariate indipendenti X_1, X_2, …, X_p
Assumiamo che esista una relazione tra Y e X

$$Y = f(X) + \varepsilon$$

Il nostro obiettivo è

$$\hat{Y} = \hat{f}(X)$$

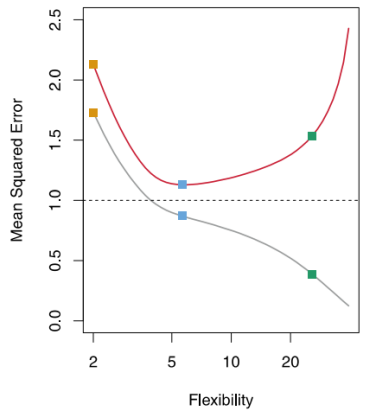
Dobbiamo valutare quando il nostro modello è in grado di prevedere il valore vero di una nuova osservazione

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Una possibilità è quella di calcolare MSE sul training set. Ci aspettiamo che MSE abbia valori piccoli e che quindi, nel training set, il modello che utilizziamo commenta pochi errori. Il nostro principale interesse però non è rivolto al training set, ma al test set, ovvero alla capacità del modello di prevedere correttamente il valore della variabile target per nuove osservazioni.
Calcoliamo quindi il test MSE

$$\text{Ave}(y_0 - \hat{f}(x_0))^2$$

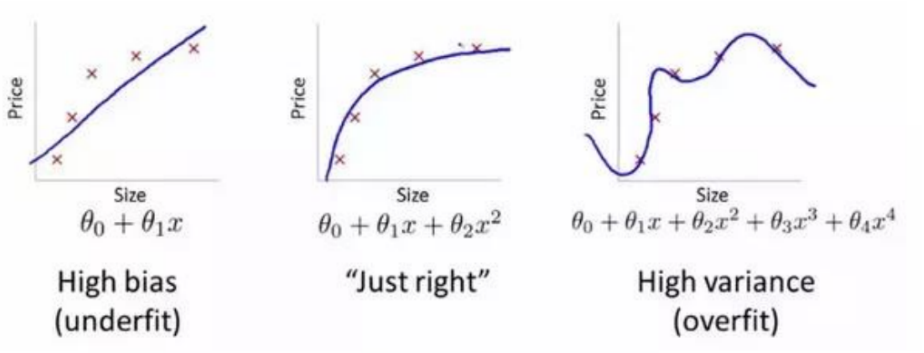
Il training MSE è monotono e decrescente. L'andamento del test MSE prende il nome di U shape

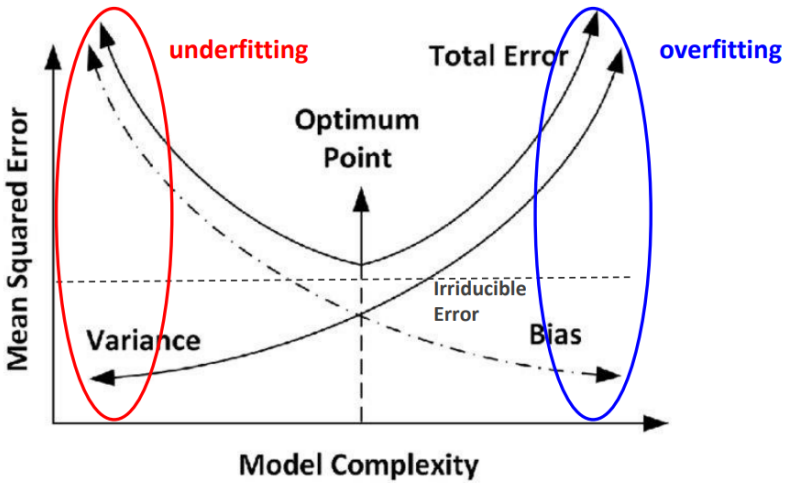


All'aumentare della flessibilità del modello il training MSE diminuisce ma non è detto che valga la stessa cosa per il test MSE
Il test MSE atteso, per una nuova osservazione x_0, può essere scomposto in 3 quantità

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

La quantità a sinistra dell'uguale rappresenta l'MSE che otterremo se potessimo stimare ripetutamente *f* su diversi training set e testare f^ sullo stesso x_0
Var(ε) rappresenta l'errore irriducibile
La varianza di f^ si riferisce alla quantità per la quale il nostro modello cambierebbe se venisse stimato utilizzando un training set diverso
Il bias si riferisce all'errore che si introduce approssimando una funzione reale (per esempio, applicare un modello lineare a un problema non lineare porterà a un test MSE molto alto)





ALGORITMO K - NN (classificatore K nearest neighbours)

KNN è un algoritmo non parametrico chiamato lazy learning algorithm . Un metodo si dice non parametrico quando non sono necessarie assunzioni sulla distribuzione dei dati di partenza. Questo tipo di classificatori non richiedono che nessun modello venga stimato. L'unico parametro presente nel K - NN è il parametro *k*, detto parametro di tuning e rappresenta la dimensione del vicinato.
Il classificatore K - NN assegna la nuova osservazione al gruppo che ha più osservazioni tra le k osservazioni più vicine alla nuova osservazione x_0

Dunque il punto ora diventa definire un metodo per calcolare la distanza tra le osservazioni (spesso useremo la distanza euclidea)

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Procedura

1. Determinare il parametro k
2. Calcolare la distanza tra un nuovo punto e tutti i punti del training
3. Ordinare le distanze e individuare le k osservazioni più vicine
4. Individuare la categoria Y di tutte le k osservazioni più vicine
5. Individuare la categoria maggiormente presente e utilizzarla come previsione per il nuovo punto

L'unico parametro che può influenzare la difficoltà dell'algoritmo è il k numero di osservazioni più vicine da considerare. Più grande è il parametro k più le linee di confine saranno smooth. Basandosi solo sull'osservazione più vicina, l'overfitting può essere un grosso problema. Per evitare l'overfitting, possiamo rilassare le linee di confine considerando più di un vicino. Consideriamo quindi le k osservazioni più vicine.
Da un punto di vista computazionale, questo algoritmo può essere molto intensivo. Quando il training dataset è molto grande calcolare la distanza tra ogni nuovo punto e ogni punto del training e successivamente ordinare le distanze può richiedere tempo.

Formalizziamo nella notazione nota il K-NN

- modello non parametrico per il calcolo della $p_g(x) = P(Y = g \mid X = x)$
- Il modello k-nearest neighbors stima le probabilità come

$$\hat{p}_{kg}(x) = \hat{P}_k(Y = g \mid X = x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x, \mathcal{D})} I(y_i = g)$$

- In altre parole: la probabilità di ogni classe g è proporzionale al numero di x_i con quella classe nell'intorno di dimensione k

A questo punto la classificazione sarà effettuata sulla base della probabilità stimata più alta

$$\hat{C}_k(x) = \underset{g}{\operatorname{argmax}} \ \hat{p}_{kg}(x)$$

Corrisponde a classificare rispetto alla classe più presente nell'intorno di dimensione k
Nel caso di un problema a 2 classi la regola decisionale sarà

$$\hat{C}_k(x) = \begin{cases} 1 & \hat{p}_{k0}(x) > 0.5 \\ 0 & \hat{p}_{k1}(x) < 0.5 \end{cases}$$

La standardizzazione (e normalizzazione) sono spesso usate nella fase di pre-processing. Poiché K-NN è basato su una misura di distanza dobbiamo sempre ricordarci di trasformare le nostre variabili.
Utilizziamo il validation set per valutare le performance dei diversi k applicati al training set

REGRESSIONE LOGISTICA

Si vuole descrivere la relazione di dipendenza del possesso di un attributo dicotomico da una o più variabili indipendenti (X_1, X_2, .., X_p) = X, di natura qualsiasi (sia qualitative che quantitative)

Gli obiettivi possono essere molteplici:

- individuare tra le variabili indipendenti quelle che sono maggiormente esplicative (vanno dunque interpretate come *determinanti* per il possesso o meno della variabile target). A seconda che siano correlate positivamente o negativamente possono essere detti **fattori di rischio** o **fattori di protezione**
- ricercare la combinazione lineare delle variabili indipendenti che meglio discrimina fra il gruppo delle unità che possiedono l'attributo e quello delle unità che non lo possiedono
- stimare la probabilità del possesso dell'attributo per una nuova unità statistica su cui è stato osservato il vettore di variabili X e, fissato per tale probabilità un valore soglia, classificare l'unità alla categoria delle unità che possiedono l'attributo oa quello delle unità che non lo possiedono

Si tratta di costruire un modello di regressione per Y, variabile risposta dove Y dicotomica a valori 0 e 1, corrispondenti rispettivamente all'assenza e alla presenza dell'attributo.

In un modello di regressione la quantità che si ipotizza funzione di X è il valore medio atteso della variabile dipendente Y condizionato ad un dato x,

$$E(Y \mid \mathbf{x})$$

Modello di *regressione logistica*

valor medio condizionato $E(Y \mid \mathbf{x})$

corrisponde a

$P(Y = j \mid \mathbf{x})$, con j = 1, 2,

cioè alla probabilità di possedere uno degli attributi in esame condizionata al fatto che il vettore delle variabili indipendenti assume valore x.

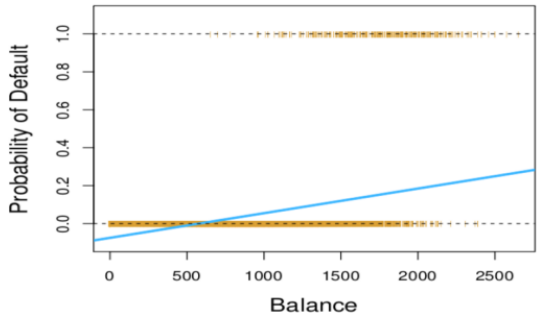
Si vuole descrivere la funzione che lega tale probabilità, che indicheremo con π(x), alla combinazione delle variabili indipendenti.

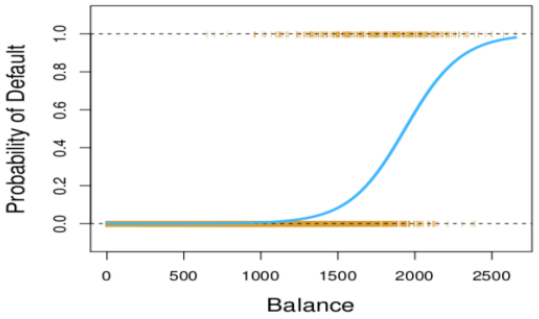
Perchè il modello di regressione lineare non va bene? Il problema è che noi vogliamo che la variabile risposta sia compresa tra 0 e 1, ma la funzione lineare di X, essendo non limitata (né superiormente né inferiormente) potrebbe dare luogo a valori stimati di $\pi(x)$ esterni all'intervallo [0, 1], e quindi privi di senso.

Per descrivere la relazione di dipendenza della probabilità $\pi(x) = P(Y = 1 \mid x)$ dai valori di $X = (X_1, X_2, .., X_p)$ si può usare la **distribuzione logistica:**

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}} = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}.$$

(a col pallino è la sommatoria)





Il grafico di tale funzione descrive una curva monotona a forma di S allungata (detta "sigmoide"), limitata superiormente dalla retta $y = 1$ e inferiormente dalla retta $y = 0$, alle quali tende asintoticamente.

Nel modello di regressione lineare l'errore si distribuisce normalmente, con media nulla e varianza costante. Questa assunzione non è valida quando Y è una variabile dicotomica, perchè in tal caso l'errore può assumere solo due valori

$$\varepsilon = Y - \pi(x) = \begin{cases} 1-\pi(x) & \text{con probabilità } \pi(x) \\ -\pi(x) & \text{con probabilità } 1-\pi(x) \end{cases},$$

con media e varianza

$$E(\varepsilon) = [1-\pi(x)]\pi(x) - \pi(x)[1-\pi(x)] = 0 \quad V(\varepsilon) = [1-\pi(x)]^2\pi(x) + \pi(x)^2[1-\pi(x)] = \pi(x)[1-\pi(x)],$$

che dipende dal valore di $\boldsymbol{X}$, quindi non è costante.

La variabile aleatoria $Y|\boldsymbol{x}$ segue quindi la distribuzione di Bernoulli $Ber(\pi(\boldsymbol{x}))$ con

$$f(y|x) = \pi(x)^y [1-\pi(x)]^{(1-y)} \text{ con}$$
$$E(Y|x) = \pi(x) \quad \text{e} \quad V(Y|x) = \pi(x)[1-\pi(x)]$$

Si consideri, ora, la seguente funzione di $\pi(x)$, detta logit,

$$logit(\pi(x)) = \ln\left\{\frac{\pi(x)}{1-\pi(x)}\right\};$$

Che è il logaritmo naturale del rapporto della probabilità condizionata di possedere l'attributo diviso la probabilità condizionata di non possederlo.

Il rapporto fra probabilità associate ad una dicotomia, cioè fra probabilità complementari, è detto **odds**.

Sostituendo: $$logit(\pi(x)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

(le é e le ù sono parentesi graffe, la a col pallino una sommatoria)

I parametri, in un modello logistico, vengono stimati attraverso il metodo della massima verosimiglianza:

$$l(\beta_0, \beta) = \prod_{i:y_i=1} \pi(x_i) \prod_{i:y_i=0} (1 - \pi(x_i))$$

Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti.

Statistica-test sui parametri "test rapporto di verosimiglianza"

$$G = -2\ln\left[\frac{\text{verosim. modello senza la variabile}}{\text{verosim. modello con la variabile}}\right] = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_i}\left[1-\hat{\pi}(x_i)\right]^{(1-y_i)}}\right]$$

Sotto l'ipotesi zero H0: $\beta_1=0$ che l'inserimento della variabile X nel modello non apporti un contributo significativo, nell'universo dei campioni la variabile campionaria G si distribuisce asintoticamente come una variabile aleatoria.

Wald Chi-square: il quadrato del rapporto tra stima e standard error.
Se il p-value è piccolo, cioè < del livello di significatività α fissato a priori, allora rifiuto H_0 (ossia, rifiuto l'ipotesi di coefficiente nullo) → il regressore a cui il coefficiente è associato è rilevante per la spiegazione del fenomeno (il coefficiente stimato è significativamente diverso da zero)

Nel modello semplice di regressione lineare il valore di β_1 rappresenta la variazione media di Y al crescere di un'unità di X.
Nel modello di regressione logistica

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}},$$

in termini di logit

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

β_1 esprime la variazione del logit corrispondente ad un incremento unitario di X.

Nella regressione logistica possiamo trattare variabili qualitative come variabili di input (attraverso l'introduzione di diverse variabili dummy)

Un grafico utile per valutare l'adeguatezza del modello è quello contenente i valori stimati in ascissa e i residui in ordinata: in un buon modello tali punti dovrebbero essere disposti casualmente intorno all'asse delle ascisse. Se invece si evidenziano andamenti particolari potrebbe non essere corretta la scelta del logit come funzione legame. Questa eventualità può rappresentare una spiegazione anche per comportamenti difformi dall'atteso nel grafico che controlla la normalità dei residui.

La ricerca di valori anomali può essere effettuata anche valutando la differenza della stima dei parametri conseguente all'esclusione dal dataset di un'unità alla volta.
L'ispezione dei residui consente in primo luogo di controllare la validità delle assunzioni preliminari.

> Per esempio, è possibile controllare l'ipotesi di linearità della relazione fra il *logit(P(Y=1|X=x))* e un dato regressore continuo X attraverso la rappresentazione grafica dei punti di coordinate $\left(x_k, \hat{y}_k\right)$, per $k = 1, ..., J$ (con $J \leq n$).

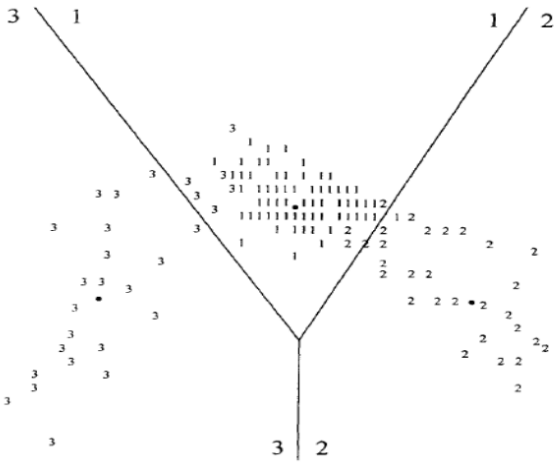
Se la numerosità campionaria non è troppo elevata, può essere utile analizzare un semplice grafico dei residui (in ordinata) corrispondenti alle varie unità statistiche (elencate in ascissa). Dato che in un buon modello i residui dovrebbero essere prossimi allo 0, l'utilità di questo grafico sta nella possibilità di evidenziare la presenza di residui "grandi" (in valore assoluto, di solito esterni all'intervallo [-2, 2]), cioè di valori che il modello non è in grado di spiegare.

LE REGOLE DI CLASSIFICAZIONE

Ci sono diversi tipi di metodi di classificazione in letteratura
Essenzialmente si differenziano tra di loro sulla base del modo in cui si definiscono le regole di classificazione. Una prima e importante differenziazione è tra i metodi che si concentrano sulla discriminazione tra le classi e quelli che invece cercano di modellare le classi stesse.

La prima classe di metodi - detti di classificazione pura - – cercano in maniera implicita o esplicita di trovare i confini che separano le differenti classi nello spazio multidimensionale. In questi casi la risposta che si ottiene in termini di classificazione è sempre l'assegnazione ad una delle K classi disponibili.

Questi metodi si concentrano sul trovare dei confini ottimali tra le classi da discriminare
e.g. la discriminaizione di paziente eu, iper, e ipo tiroidei
In questo caso, la classificazione dei campioni può essere fatta misurando alcune variabili sui campioni in esame, identificando i baricentri delle distribuzioni dei campioni, e tracciando delle superfici a metà strada tra i diversi baricentri.



Come detto, questi metodi procedono alla classificazione dei campioni in una di K classi disponibili.
Per costruire il modello di classificazione, in tutti i casi si parte dalla Regola di Bayes:

"un campione va assegnato alla classe per la quale sia maggiore la sua probabilità di appartenenza"

Il processo di classificazione è quindi un processo a due stadi:
1. Calcolo della probabilità che un campione incognito appartenga a ciascuna delle K classi (o di una qualsiasi funzione monotona di questa probabilità – detta funzione di classificazione)
2. Assegnazione del campione alla classe corrispondente alla probabilità più alta.

Consideriamo il dataset default e supponiamo di usare una sola variabile indipendente (essere studente o meno), per determinare

$$P(default = Yes|student = Yes)$$

Incrociando i dati

	student		
default	No	Yes	Sum
No	6850	2817	9667
Yes	206	127	333
Sum	7056	2944	10000

Possiamo stimare direttamente $\hat{P}(default = Yes|student = Yes) = \frac{127}{2944} = 0.043$

Tuttavia nella pratica, volendo prevedere la probabilità di default sulla base di variabili qualitative e quantitative, risulta più semplice passare attraverso l'utilizzo del teorema di Bayes.

Indichiamo con $\pi = P(default = Yes)$ e corrispondentemente $(1 - \pi) = P(default = No)$. Inoltre,

$$P(S|D) = P(student = Yes|default = Yes)$$
$$P(S|\bar{D}) = P(student = Yes|default = No)$$

Utilizzando il teorema di Bayes

$$P(default = Yes|student = Yes) = \frac{\pi \cdot P(S|D)}{\pi \cdot P(S|D) + (1 - \pi) \cdot P(S|\bar{D})}$$

In assenza di qualsiasi informazione ulteriore, selezionata un'unità a caso, utilizzando la tabella, possiamo stimare

	student		
default	No	Yes	Sum
No	6850	2817	9667
Yes	206	127	333
Sum	7056	2944	10000

$$\hat{\pi} = \frac{333}{10000} = 0.0333$$
$$\hat{P}(S|D) = \hat{P}(student = Yes|default = Yes) = \frac{127}{333} = 0.38$$
$$\hat{P}(S|\bar{D}) = \hat{P}(student = Yes|default = No) = \frac{2817}{9667} = 0.29$$

Inserendo i valori ottenuti nella formula di Bayes otteniamo

$$\hat{P}(default = Yes|student = Yes) = \frac{0.0333 \cdot 0.38}{0.0333 \cdot 0.38 + 0.9667 \cdot 0.29}$$
$$= 0.043$$

L'analisi discriminante utilizza il teorema di Bayes per produrre delle stime che una variabile risposta Y qualitativa appartenga ad una certa categoria k = 1, …, K sulla base di informazioni fornite da p variabili indipendenti.

La regressione logistica viene utilizzata per calcolare:

$$\Pr(Y = k|X = x)$$

Questo è possibile stimando:

$$\Pr(Y = k|X = x) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Attraverso il metodo della massima verosimiglianza.

La regressione logistica modellizza la distribuzione della variabile dipendente Y date le variabili indipendenti X.

L'approccio alternativo dell'analisi discriminante lineare consiste nello stimare la distribuzione delle variabili indipendenti X separatamente per ogni classe di risposta presente in Y. Date le distribuzioni così calcolate, ottenere Pr(Y = k | X = x) attraverso il teorema di Bayes.

ANALISI DISCRIMINANTE LINEARE

Vantaggi:

- Se le classi della variabile Y sono "ben separate", la stima dei parametri nella regressione logistica diventa poco stabile
- Se n è piccolo e la distribuzione delle X è approssimativamente normale allora l'analisi discriminante lineare risulta un approccio migliore
- Analisi discriminante lineare è molto usata quando il numero di classi K è maggiore di 2

Supponiamo di voler classificare una osservazione in una delle K classi disponibili in Y, con K ≥ 2. La variabile risposta Y, di tipo qualitativo, può prendere uno dei K possibili valori distinti e non ordinabili

Consideriamo π_k la probabilità a priori che una osservazione scelta casualmente venga da una delle k classi. Per semplicità, consideriamo X discreta perciò definiamo

$f_k(x)$ = Pr (X = x | Y = k) come la funzione di densità di X per una osservazione x data la sua classe di appartenenza k

Grazie al teorema di Bayes possiamo scrivere:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Per calcolare Pr(Y = k | X = x), da ora in poi abbreviata in p_k (X), dobbiamo trovare delle valide stime di π_k e $f_k(X)$
Stimare π_k è molto semplice: rapporto tra la numerosità della classe k e il totale delle osservazioni
Stimare $f_k(X)$ è molto complesso: una possibile soluzione è assumere delle funzioni di densità note
La probabilità che cerchiamo è una probabilità a posteriori: la probabilità che un'osservazione x appartenga alla classe k dato il valore osservato di x. Più è alta questa probabilità, più sarà verosimile che la classe di osservazione x sia k

Questo è detto classificatore di Bayes, ed è il classificatore con il più basso error rate possibile.

Analisi discriminante lineare per p = 1
Supponiamo p = l

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

dobbiamo fare delle ipotesi sulla funzione di densità di $f_k(X)$

Supponiamo che $f_k(X)$ sia:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

Distribuzione normale (caso univariato per semplicità), con:
μ_k media della k - esima classe
σ^2_k varianza della k - esima classe

In questo caso assumiamo che la varianza sia costante, ovvero $\sigma^2_1 = \ldots = \sigma^2_k = \sigma^2$

Sostituendo in $p_k(x)$, otteniamo:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}$$

Considerando il *log* della precedente eguaglianza si può dimostrare che è equivalente to assegnare l'osservazione x alla classe k quando $\delta_k(x)$ è massima.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

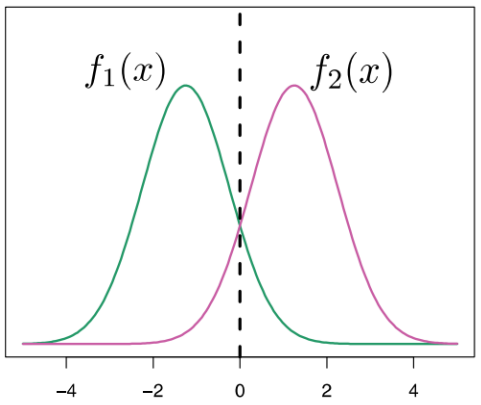
Ipotizziamo sia possibile avere un classificatore di Bayes.
In tal caso, con K = 2 e $\pi_1 = \pi_2$, il classificatore assegnerà l'osservazione x nel seguente modo

Classe 1 se $2x\,(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$
Classe 2 altrimenti

In questo caso, è possibile calcolare la linea (regola) di confine (o di decisione) di Bayes.

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Consideriamo noti $\mu_1 = -1.25$, $\mu_2 = 1.25$, $\sigma^2 = 1$



Sostituendo in:

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Otteniamo la seguente regola di decisione:

Classe 1 se $x < 0$
Classe 2 altrimenti

Nell'analisi discriminante lineare dobbiamo stimare i valori di $\mu_1, \ldots, \mu_k$, di $\pi_1, \ldots, \pi_k$ e di σ^2

Date le stime di questi parametri, possiamo sostituirli a:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Gli stimatori utilizzati sono i seguenti:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

dove il primo è la media condizionata alla classe calcolata sul training set, il secondo può essere visto come la media pesata delle varianze di ogni classe k (training set)
$\pi_1, \ldots, \pi_k$ se non sono note le probabilità a priori di ogni classe possono essere calcolate semplicemente come:

$$\hat{\pi}_k = n_k/n$$

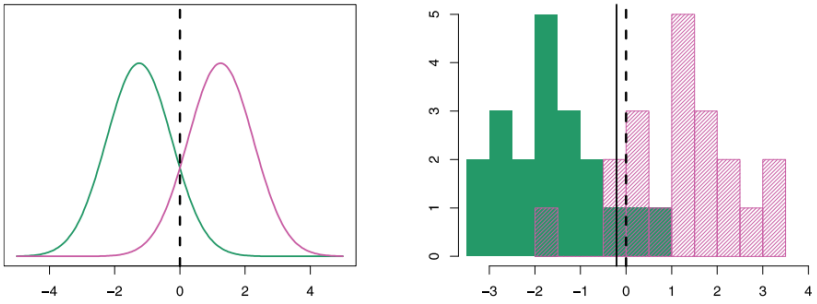
Numerosità della classe considerata diviso il numero totale delle osservazioni nel training set

A questo punto assegneremo l'osservazione x alla classe per la quale:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

è maggiore

$\hat{\delta}_k(x)$ sono chiamate *funzioni discriminanti* e sono funzioni lineari di x.



Importante

Analisi discriminante lineare

- Le osservazioni in ogni classe provengono da una distribuzione normale.
- Ogni classe è caratterizzata da un proprio vettore di medie.
- La varianza è considerata uguale per ogni classe.

Fino ad ora abbiamo considerato una sola variabile indipendente p = 1.
Estendiamo ora l'analisi discriminante lineare a più variabili indipendenti p > 1.
Consideriamo X = (X_1, X_2, .., X_p)
dove X ha una distribuzione normale multivariata con media uguale a un vettore di medie condizionate alla classe e matrice varianza-covarianza comune per tutte le classi.

Un variabile casuale p-dimensionale si distribuirà come una normale multivariata.

$$X \sim N(\mu, \mathbf{\Sigma})$$

Dove:

$E(X) = \mu$ è la media di $\boldsymbol{X}$ (vettore p-dimensionale)

$\mathbf{\Sigma} = \mathrm{Cov}(X)$ è la matrice di varianza-covarianza di $\boldsymbol{X}$ di dimensione p x p

La distribuzione di densità di una normale multivariata si può scrivere come:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)$$

Nell'analisi discriminante lineare nel caso p >1 si assume che le osservazioni derivino da una distribuzione normale multivariata

$N(\mu_k, \boldsymbol{\Sigma})$

dove

$\boldsymbol{\Sigma}$ vettore di medie condizionate alla classe di dimensione p

μ_k matrice varianza-covarianza comune a tutte le classi K

Ora considerando

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

e

$$f(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu)\right)$$

otteniamo che l'osservazione x verrà assegnata alla classe k con il più alto valori

$$\delta_k(x) = x^T \boldsymbol{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}^{-1}\mu_k + \log \pi_k$$

Metriche per la valutare l'accuratezza di specifiche classi sono molto importanti in differenti settori come quello medico.
In questo settore due possibili metriche sono:

- Sensitivity: percentuali di osservazioni correttamente classificate come Default
- Specificity: percentuale di osservazioni correttamente classificate come No-Default

L'analisi discriminante lineare (Linear Discriminant Analysis – LDA) cerca di approssimare il classificatore di Bayes.
Il classificatore di Bayes cerca di minimizzare il Bayes rate.
Il Bayes rate è una misura generale di accuratezza e non mira a migliorare le performance su determinate classi di un problema.

Nel caso di un problema a due classi, come l'esempio Default, il classificatore di Bayes assegnerà una osservazione a Default quando:

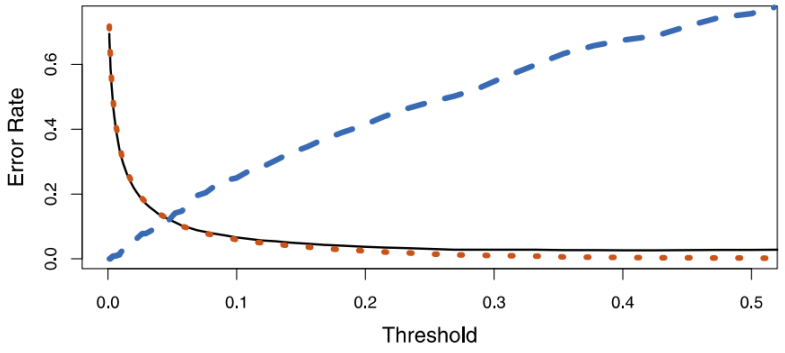
$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5$$

Anche LDA utilizza la stessa soglia:

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5$$

Nel caso fossimo interessati a classificare più correttamente la classe Default potremmo ridurre la soglia.

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.2$$



Frazione di Default classificati erroneamente

Frazione di No-Default classificati erroneamente

LDA assume che le osservazioni all'interno di ogni classe provengano da una distribuzione normale multivariata con media condizionata a ogni classe e matrice di varianza-covarianza in comune per ogni classe K.

L'**analisi discriminante quadratica** (**QDA**) estende questo approccio.

QDA assume che ogni classe ha la sua matrice di varianza-covarianza.

$$X \sim N(\mu_k, \mathbf{\Sigma}_k)$$

Dove,

$\mathbf{\Sigma}_k$ è la matrice di varianza-covarianza della k-esima classe.

In questo caso l'osservazione x verrà assegnata alla classe k quando

$$\begin{aligned} \delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\mathbf{\Sigma}_k^{-1}x + x^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k \end{aligned}$$

è il più alto.

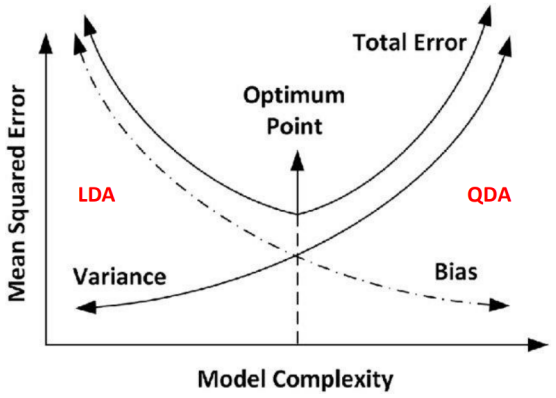
LDA vs QDA

Supponiamo di avere un problema con p = 50.

In generale stimare una matrice varianza-covarianza con 50 variabili indipendenti richiede di calcolare *p(p+1)/2* coefficienti cioè 1275.

In QDA, è necessario calcolare una matrice di varianza-covarianza per ogni classe ottenendo così un numero di coefficienti da calcolare pari a *Kp(p+1)/2* .

In LDA, dove la matrice varianza-covarianza è considerata in comune tra tutti le classi *K*, i coefficienti lineari sono *Kp*.



LDA è meno flessibile di QDA ed ha quindi meno varianza.
Ma c'è un tradeoff: se l'assunto di LDA che le classi condividano la stessa matrice di varianza-covarianza è totalmente errato, c'è un bias elevato.
LDA tende ad essere la scelta migliore se ci sono relativamente poche osservazioni nel training set e ridurre la variabilità è cruciale.
Al contrario, QDA è consigliata se il training set è molto grande, o se l'assunzione di una matrice di covarianza comune per le classi è chiaramente insostenibile.

CONFRONTO TRA METODI DI CLASSIFICAZIONE
Date le variabili indipendenti X e la variabile dipendente Y ci sono diversi modi per costruire un metodo di classificazione

- Metodi discriminanti: consistono nel mappare direttamente lo spazio delle feature X nello spazio della risposta Y e successivamente stimare la $p(y \mid x)$
- Approccio alternativo: costruire un modello che studia le variabili indipendenti rispetto a una classe di risposta → analisi discriminante
  I modelli generativi cercano di modellizzare $p(x \mid y)$

Regressione logistica e analisi discriminante lineare, pur partendo da presupposti differenti, sono metodi simili. Infatti, sia la regressione logistica sia la LDA individuano delle linee di confine lineari. L'unica differenza è determinata da come sono individuati $_0$, $_1$, c_0 e c_1. I primi sono stimati con il metodo di massima verosimiglianza, gli ultimi sono calcolati usando le stime di media e varianza della distribuzione normale.

LDA e regressione logistica si comportano spesso allo stesso modo.
LDA assume che le osservazioni provengano da una distribuzione normale con una comune matrice di covarianza. Quando queste ipotesi sono vere LDA è più affidabile della regressione logistica. Nel caso in cui queste ipotesi non fossero confermate la regressione logistica è avvantaggiata.

KNN è un approccio completamente differente. Una nuova osservazione X = x viene associata alla classe più presente tra le k osservazioni più vicine. KNN è un approccio non parametrico dove nessuna assunzione è fatta sulla forma delle linee di decisione. Ci aspettiamo che KNN sia migliore di LDA e della regressione logistica quando le linee di confine sono marcatamente non lineari.

KNN non permette però l'interpretazione delle variabili d'interesse più significative. QDA è un compromesso tra KNN e LDA/regressione logistica. Le linee di decisione quadratica possono essere adatte ad un numero maggiore di problemi differentemente da quanto succede per LDA.

Curve ROC

La sensibilità e la specificità sono le due misure principali che vengono impiegate per valutare la capacità del test di individuare, fra gli individui di una popolazione, quelli provvisti del <<carattere>> ricercato e quelli che invece ne sono privi.

- Sensitivity: percentuale di osservazioni correttamente classificate come la classe di interesse
- Specificity: percentuale di osservazioni correttamente classificate come l'altra classe

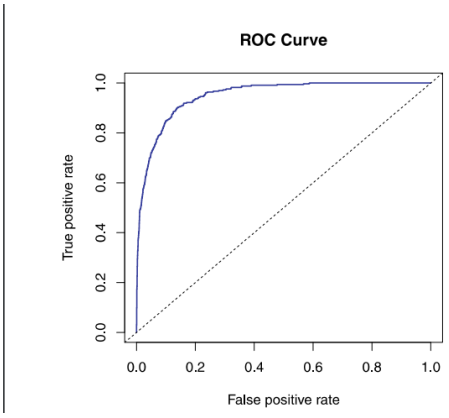
Ad esempio, in medicina, il carattere d'interesse è sempre rappresentato dalla malattia o dall'infezione

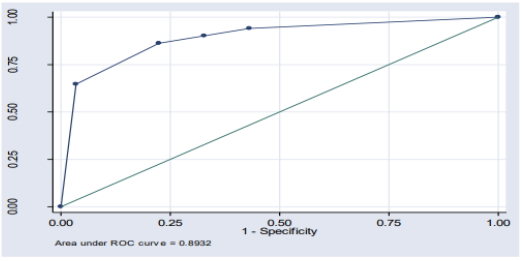
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

La sensibilità non è l'unica qualità desiderabile di un test; a noi non serve che il 100% dei positivi sia identificato come positivo, ci serve che **solo** quelli che sono positivi vengano identificati come positivi.

ROC: receiver operating characteristic. Le curve ROC sono approccio per valutare la capacità discriminatoria di un Test. Traccia la probabilità di un risultato vero positivo (sensibilità) in funzione della probabilità di un risultato falso positivo per una serie di punti di cut-off





Detailed report of Sensitivity and Specificity

Cut point	Sensitivity	Specificity	Correctly Classified
(>= 1)	100.00%	0.00%	46.79%
(>= 2)	94.12%	56.90%	74.31%
(>= 3)	90.20%	67.24%	77.98%
(>= 4)	86.27%	77.59%	81.65%
(>= 5)	64.71%	96.55%	81.65%
(> 5)	0.00%	100.00%	53.21%

TEXT MINING

CLUSTERING

Supervised approaches

- nel training sono presenti sia input che output
- esistono degli esempi con tutte le informazioni necessarie per stimare (allenare) un modello di classificazione
- la costruzione di training, validation e test set è cruciale
- sono metodi veloci e accurati
- abilità di generalizzare

Unsupervised approaches

- il modello non ha a disposizione gli output
- usato per raggruppare dati sulla base delle loro caratteristiche statistiche (misure di distanza)
- può essere usato per etichettare gruppi di osservazione

Guardiamo i metodi di clustering nel contesto del text mining, per individuare dei metodi ad esempio che ci consentano di etichettare un testo senza aver letto quel testo.

Supervised Learning	**Unsupervised Learning**
Campione di training con output noti. Noto il numero di classi.	Campione di training senza etichette. Non è noto il numero di classi (gruppi).

Il clustering è **soggettivo** → ci sono molti modi diversi di creare dei gruppi a partire dalle osservazioni. Vogliamo trovare un modo per raggruppare osservazioni che condividono caratteristiche simili sulla base di precisi criteri. Il clustering è un metodo di unsupervised learning e data exploration (uno strumento per individuare patterns o strutture di interesse tra i dati)

Vogliamo massimizzare la distanza tra i gruppi e minimizzare la distanza entro i gruppi. è sempre buona cosa, se possibile, individuare raggruppamenti naturali tra le osservazioni.

Possibili problemi:
- numero di osservazioni troppo elevata (la matrice di distanze diventerebbe troppo grande)
- la capacità di un algoritmo di separare i gruppi dipende dalla definizione di similarità (**distanza**)
- i risultati di un algoritmo di clustering possono essere interpretati in maniera diversa

Definizione: consideriamo O_1 e O_2 due oggetti presi da un universo di possibili oggetti. La distanza (similarità o dissimilarità) tra O_1 e O_2 è un numero reale denominato $D(O_1, O_2)$

Una misura di distanza deve rispettare le proprietà di:
- Non negatività: $D(x, y) \geq 0$
- Riflessività: $D(x, y) = 0$ se e solo se $x = y$
- Simmetria: $D(x, y) = D(y, x)$
- Disuguaglianza triangolare: $D(x, z) + D(z, y) \geq D(x, y)$

Se la seconda proprietà non è soddisfatta, D(x, y) è chiamata pseudo metrica

Spesso, invece di parlare di similarità, ci si riferisce a misure di dissimilarità. Una misura di dissimilarità è una funzione f(x, y) tale per cui f(x, y) > f(w, z) se e solo se x è meno simile di y che w di z.
La distanza è una misura pair wise (analizza elementi presi a coppie)

La distanza più utilizzata è la distanza Euclidea

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Un metodo che può essere usato per valutare una misura di distanza è la correlazione lineare di Pearson (che, a differenza della distanza euclidea, considera anche l'andamento delle curve e non solo la distanza)

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

In realtà, la correlazione lineare di Person è una misura di similarità, ma può essere trasformata in una misura di dissimilarità nel seguente modo:

$$d_p = \frac{1 - \rho(\mathbf{x}, \mathbf{y})}{2}$$

Ci sono due diversi metodi di clustering: Clustering gerarchico e clustering partizionale
Negli approcci gerarchici viene sempre utilizzato il **dendogramma**. La similarità tra due oggetti in un dendogramma è rappresentata dall'altezza del più basso nodo interno che condividono.
Ci sono due metodi per costruire un dendogramma

1) Bottom-up (agglomerativo)
   - ogni osservazione è considerata un cluster, allo step successivo la miglior coppia di osservazioni (minor distanza) finisce dentro un nuovo cluster
   - questo procedimento è ripetuto fino a che tutti i cluster sono fusi assieme
2) Top-down (divisivo)
   - tutte le osservazioni sono considerate come un unico cluster
   - viene individuato un metodo di separazione
   - si suddividono ricorsivamente i cluster

Il primo passo per costruire un metodo gerarchico è sempre costruire una matrice di distanza.

Misure di distanze tra i gruppi (Linkage)
Complete linkage: la distanza tra cluster è determinata dalla più grande distanza tra ogni coppia nei due differenti clusters
Single linkage: la distanza tra cluster è determinata dalla più piccola distanza tra ogni coppia nei due differenti clusters
Group average linkage: la distanza tra cluster è determinata dalla distanza media tra ogni coppia nei due differenti clusters

Un quarto metodo di linkage è quello della devianza

Riunisce, ad ogni step, i due gruppi dalla cui fusione deriva il minimo incremento possibile della devianza "intra".

$$DEV_T = \sum_{s=1}^{p}\sum_{i=1}^{n}(x_{is} - \bar{x}_s)^2 = \sum_{i=1}^{n}\sum_{s=1}^{p}(x_{is} - \bar{x}_s)^2$$

x_s è la media del gruppo. Tale devianza può essere scomposta come segue:

$$DEV_{IN} = \sum_{k=1}^{g}\sum_{s=1}^{p}\sum_{i=1}^{n_k}(x_{is} - \bar{x}_{s,k})^2$$

che è la devianza intra-gruppo riferita alle p variabili con riferimento al gruppo k, $x_{s,k}$ dove è la media della variabile s con riferimento al gruppo k

$$DEV_{OUT} = \sum_{s=1}^{p}\sum_{k=1}^{g}(\bar{x}_{s,k} - \bar{x}_s)^2 n_k$$

che è la devianza tra i gruppi.

Nel passare da k a k +1 gruppi (aggregazione) DEV_{IN} aumenta, mentre ovviamente DEV_{OUT} diminuisce. Ad ogni passo si aggregano tra loro quei gruppi per cui vi è il minor incremento della devianza intra-gruppo (metodo di Ward)

Clustering partizionale

Sono metodi non gerarchici. Ogni osservazione è posizionata esattamente in uno dei K cluster sulla base di precise regole.

Un esempio di questi metodi è il metodo delle k medie

In questo metodo una buona suddivisione delle osservazioni in gruppo è quella per cui la variazione intra-cluster, $W(C_k)$ è minima. Per risolvere questo problema di minimizzazione, dobbiamo definire questa distanza.

Una delle misure più utilizzate è la ***distanza euclidea quadratica***.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Dove $|C_K|$ è il numero di osservazioni nel cluster K.

A questo punto il problema di minimizzazione può essere scritto come:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Definito il valore di K, l'algoritmo funziona come segue:

- Assegnare casualmente un valore, da 1 a k, a ogni osservazione
- Calcolo il valore del centroide (vettore di dimensione p che contiene le medie delle variabili per le osservazioni nel cluster k-esimo)
- Assegnare ogni osservazione al cluster per il quale il centroide risulta più vicino. La vicinanza è determinata dal valore della distanza euclidea
- Ripetiamo il secondo e il terzo punto finchè non raggiungiamo la convergenza.

L'algoritmo garantisce che il valore della

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

viene minimizzato ad ogni step

Questo è vero perché vale l'uguaglianza:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

dove $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ è la media della variabile j nel cluster C_k.

Metodo dei k-medoidi

Il medoide è una osservazione per la quale la dissimilarità media con tutte le altre osservazioni nel cluster è minima

$$M^* = argmin_M \sum_i min_k \, d(x_i, m_k)$$

Steps del metodo:

- determinare un set di k medoidi
- individuare k clusters assegnando ogni punto al più vicino medoide

Ogni cluster è rappresentato da una silhouette, la quale mostra quali punti cadono nel cluster e quali invece occupano una posizione intermedia. L'intero clustering è rappresentato da tutte le silhouettes in un unico diagramma che illustra la qualità del clustering.

Costruzione della silhouette:

1. Considerare ogni osservazione i del data set, assumere A come il cluster di i
2. Calcolare la distanza media a_i di i con tutti le altre osservazioni in A

$$a_i = \frac{1}{N_A} \sum_{j \in A, j \neq i} d(i,j)$$

3. Considerare ogni cluster C differente da A e definire la distanza media $d(i, C)$ di i con le osservazioni in C

$$d(i,C) = \frac{1}{N_C} \sum_{c \in C} d(i,c)$$

4. Calcolare $d(i, C)$ per ogni clusters $C \neq A$, e selezionare quello con $b_i = min\ d(i, C), C \neq A$

La silhouette di i è definita come:

$$sil_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Osservazioni con elevata silhouette sono ben raggruppate. Il valore di k può essere determinato scegliendo il k che mi dà il valore di silhouette medio più alto

Nei metodi di clustering il fine non è la divisione in gruppi, ma l'**interpretazione** dei risultati, ovvero essere capaci di riconoscere le differenze che dividono i gruppi.

TEXT MINING

L'obiettivo finale del text mining è quello di estrarre le informazioni principali da un testo senza per forza doverlo leggere.

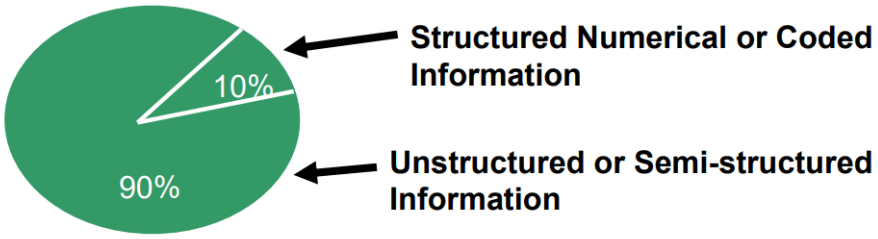
I metodi tradizionali di Data Mining utilizzano dati strutturati (matrici n x p). L'analisi di documenti e testi si riferisce invece all'analisi di dati non strutturati. Spesso dati strutturati e non strutturati sono presenti all'interno dello stesso problema (es. questionari). In questo caso si parla di dati semi strutturati

Text mining comprende metodi per l'estrazione di informazioni da testi.

Esempi di text mining: Trend analysis, emerging technology, influence of new stories, citation pattern analysis, sentiment analysis, link genes to disease

Il text mining è molto importante perchè la maggior parte dei dati sono dati non o semi strutturati

Circa il 90% dei dati è sotto forma di dati non strutturati.



Ci sono due tipi principali di text mining:

1) Analisi di testi: dato un insieme di testi estrarre il maggior numero di informazioni
2) Retrieval: dato un insieme di testi, selezionare la porzione di testo più vicina a una specifica query (es. web search)

La più comune rappresentazione di un insieme di testi nel text mining è chiamata matrice document - term

Term: tipicamente una parola singola, può essere però composto anche da brevi composizioni

Document: termine generico che indica un documento

La matrice può essere molto grande e composta da valori binari o conteggi.

Con questa trasformazione perdiamo, ovviamente, la semantica del testo

	Database	SQL	Index	Regression	Likelihood	linear
D1	24	21	9	0	0	3
D2	32	10	5	0	3	0
D3	12	16	5	0	0	0
D4	6	7	2	0	0	0
D5	43	31	20	0	3	0
D6	2	0	0	18	7	6
D7	0	0	1	32	12	0
D8	3	0	0	22	4	4
D9	1	0	0	34	27	25
D10	6	0	0	17	4	23

$$D_1 = (d_{i1}, d_{i2}, \ldots, d_{it})$$

Molte parole del vocabolario non sono utili per il text mining. Queste parole vengono chiamate **stop words.** L'eliminazione delle stop words diminuisce la dimensionalità e aumenta l'efficienza

Stemming

Tecniche utilizzate per eliminare la radice / suffisso delle parole (parte molto corposa di pre processing)

Metodi di stemming di base → rimozioni di lettere e trasformazioni di parole

Feature selection → seleziono un sotto insieme di termini che più discrimina l'obiettivo finale. Spesso le performance non peggiorano neanche diminuendo di molto il numero di termini considerati.

Misure di distanza
Data una matrice term - document, possiamo definire il concetto di distanza tra documenti. Gli elementi della matrice possono essere 0 e 1 o frequenze di apparizione (spesso normalizzate). In questa situazione possiamo utilizzare la distanza Euclidea o distanze cosine.

La distanza basata sul coseno misura l'angolo tra i vettori rappresentanti i documenti. I documenti nella stessa direzione sono molto vicini. La misura angolare viene trasformata in una misura che varia tra 0 e 1. 1 quando la similarità è massima, 0 quando la similarità è minima.

$$d_c(D_i, D_j) = \frac{\sum_{k=1}^{T} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{T} d_{ik}^2 \sum_{k=1}^{T} d_{jk}^2}}$$

Modelli di rappresentazione
Un modello di rappresentazione definisce come un documento è rappresentato e la sua rilevanza rispetto all'obiettivo.

Boolean model
Ogni documento è trattato come una bag of words, la sequenza delle parole non è considerata.
Data una collezione di documenti D, si consideri

$V = \{t_1, t_2, \ldots, t_{|V|}\}$

come il set di parole/termini distinti della collezione considerata. V è chiamata **vocabolario**.

Vocabolario
$V = \{t_1, t_2, ..., t_{|V|}\}$

Un peso $w_{ij} > 0$ è associato con ogni t_i di un documento $\mathbf{d}_j \in D$.
Un termine che non appare nel documento $\mathbf{d}_j$, avrà peso $w_{ij} = 0$.

$\mathbf{d}_j = (w_{1j}, w_{2j}, ..., w_{|V|j})$

Vector space model
Anche in questo caso i testi sono considerati come bag of words. Ogni documento è rappresentato da un vettore. Ogni peso non è più rappresentato da 0 e 1 ma da un conteggio delle frequenze

Term Frequency (TF) Scheme: Il peso di un termine t_i in un documento $\mathbf{d}_j$ è il numero di volte che t_i appare in $\mathbf{d}_j$, rapprensetato da f_{ij}.

TF: term frequency

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|V|j}\}}$$

IDF: inverse document frequency

$$idf_i = \log \frac{N}{df_i}$$

N: numero di documenti totale

df_i: numero di documenti in cui appare t_i

$$w_{ij} = tf_{ij} \times idf_i.$$

In questo caso d_i viene confrontato con q attraverso misure di similarità

Similarità basata sul coseno

$$cosine(\mathbf{d}_j, \mathbf{q}) = \frac{\langle \mathbf{d}_j \bullet \mathbf{q} \rangle}{\| \mathbf{d}_j \| \times \| \mathbf{q} \|} = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iq}^2}}$$

Il coseno è spesso usato anche nel contesto del clustering

Analisi testuale

Data la trasformazione di un documento o insieme di documenti in una matrice possiamo applicare i metodi di data mining noti:

- Classificare documenti (supervised)
- Clustering di documenti (unsupervised)

Classificazione di documenti

- Motivation
  - Classificazione automatica di un numero elevato di documenti (Web pages, e-mails, ...)
  - Customers: richieste di informazioni, domande
- A classification problem
  - Training set: esperti generano un training data set
  - Classification: il metodo determina le regole di classificazione
  - Application: le regole di classificazione possono essere applicare per classificare nuovi documenti
- Techniques
  - Logistic regression, naïve Bayes, ecc.

Clustering di documenti

- Motivation
  - Raggruppare testi che condividono termini o parole simili (articoli, news, ...)
  - Customers: ricerche di documenti
- Clustering problem
  - Presenti un gran numero di documenti
  - Clustering: i raggruppamenti vengono determinati sulla base di misure similarità
  - Application: un gruppo di documenti può essere descritto più velocemente
- Techniques
  - K-medie, clustering gerarchico, ecc…