

Data Mining - Appunti

Federico Calò

Contents

| | | |
|----------|--|-----------|
| 1 | KDD process | 4 |
| 1.1 | Steps del KDD | 6 |
| 1.1.1 | Business Understanding | 7 |
| 1.1.2 | Data Understanding | 9 |
| 1.1.3 | Data Preparation | 10 |
| 1.1.4 | Modeling | 18 |
| 1.1.5 | Valutazione | 20 |
| 1.1.6 | Deployment | 21 |
| 2 | Learning sets of rules | 22 |
| 2.1 | Classificatore basato su regole | 22 |
| 3 | K-Means | 31 |
| 4 | Decision Trees | 33 |
| 4.1 | Affrontare i valori mancanti (missing values) | 34 |
| 4.2 | Svantaggi | 35 |
| 4.3 | Training set vs. Validation set | 36 |
| 4.4 | Metodi di selezione dei cicli | 36 |
| 4.4.1 | Reduced error pruning (REP) | 36 |
| 4.4.2 | Minimal Error Pruning (MEP) | 36 |
| 4.4.3 | Pessimistic Error Pruning (PEP) | 37 |
| 4.4.4 | Error-Based Pruning (EBP) | 37 |
| 5 | Naive Bayes Classifier - Classificatore Bayesiano | 38 |
| 5.1 | Underflow Prevention | 43 |
| 6 | I modelli di regressione | 47 |
| 6.1 | Notazione matriciale | 53 |

| | | |
|-------|---|----|
| 6.2 | Regressione con funzioni a gradini | 53 |
| 6.2.1 | Funzioni costanti a tratti per la regressione semplice | 54 |
| 6.2.2 | Alberi di regressione | 54 |
| 6.2.3 | Individuazione della sequenza di suddivisioni ottimali (crescita dell'albero) | 55 |
| 6.2.4 | Estensione al caso di due o più regressori continui | 56 |
| 6.2.5 | X ordinale | 56 |
| 6.2.6 | X nominale | 56 |

1 KDD process

L'automazione delle attività economiche produce un incremento dello stream di dati perchè anche singole transazioni (una chiamata telefonica, il credito di una carta, un test medico) sono tipicamente registrate in un computer. Le basi di dati scientifiche e governative sono in rapida crescita. C'è un divario crescente tra la generazione di dati e la loro comprensione. Risulta quindi necessario l'utilizzo di computer per analizzare i dati, ma questo non è sufficiente.

Necessitiamo di una metodologia matura che spieghi come grandi strutture di dati possono essere analizzate. Questa metodologia è stata studiata in un'area di ricerca conosciuta come KDD. Lo scopo è investigare come tecniche di Machine Learning possono essere applicate a estratti di "conoscenza" di una grande massa di dati disponibili. All'inizio vi era una certa confusione sull'area di interesse ricoperta dal Machine learning, Data Mining e Knowledge Discovery. Ora, si è giunti alla conclusione che il KDD denota l'intero processo di estrazione della conoscenza, dalla raccolta dei dati per poi effettuare una pre-elaborazione degli stessi fino all'interpretazione dei risultati.

Il Data Mining è lo step, all'interno del processo KDD, nel quale le informazioni sono estratte dai dati applicando ad essi opportuni algoritmi. Questi algoritmi sono il più delle volte quelli di Machine Learning. In un contesto aziendale il termine Data Mining è ancora utilizzato per denotare il processo di knowledge discovery e questo causa qualche confusione. Sempre sulla terminologia: KDD è ancora utilizzato anche se la conoscenza non è rigorosamente estratta dai "database".

Si potrebbe definire il Knowledge Discovery come un processo nel seguente modo: "Il Knowledge Discovery è l'estrazione non banale di informazioni implicite, precedentemente sconosciute e potenzialmente utili dai dati."

- **Non banale:** si intende che nel processo è coinvolta qualche ricerca o inferenza statistica, quindi non è un semplice calcolo di quantità predefinite;
- **Implicita:** ci si riferisce al fatto che l'informazione è implicita nel dato e non formalmente esplicita, l'informazione esplicita è estratta attraverso altre tecniche;
- **Sconosciuta:** l'informazione deve essere nuova, la novità dipende dal quadro di riferimento assunto;
- **Utile:** l'informazione deve essere utile a raggiungere lo scopo del sistema o dell'utente. I pattern completamente estranei agli obiettivi dati sono di scarsa utilità e non costituiscono conoscenza all'interno della situazione data.

Possiamo formalmente definire il KDD nel seguente modo:

Dati:

- un insieme di fatti (data) F ,
- una rappresentazione in un linguaggio L ,
- una certa misura di certezza C ,

possiamo definire un pattern come una dichiarazione S in L che descrive le relazioni tra un sotto insieme F_S di F con una certezza c , tale che S è più semplice (in un certo senso) dell'enumerazione di tutti i fatti in F_S .

Un pattern è considerato conoscenza se è interessante e abbastanza certo (o valido). Nel KDD siamo interessati in pattern che sono espressi in un linguaggio di alto livello, come:

Se Età < 25 e Corso-di-Educazione = no

Allora Incidente = si

Con probabilità = 0.2 a 0.3

In questo modo alcuni pattern possono essere capiti e usati direttamente dalle persone o possono essere input ad altri programmatori. Definiamo ora il termine **certezza** come quel livello sufficiente di certezza senza il quale i modelli diventano ingiustificati e non riescono a diventare conoscenza. La certezza coinvolge diversi fattori, quali:

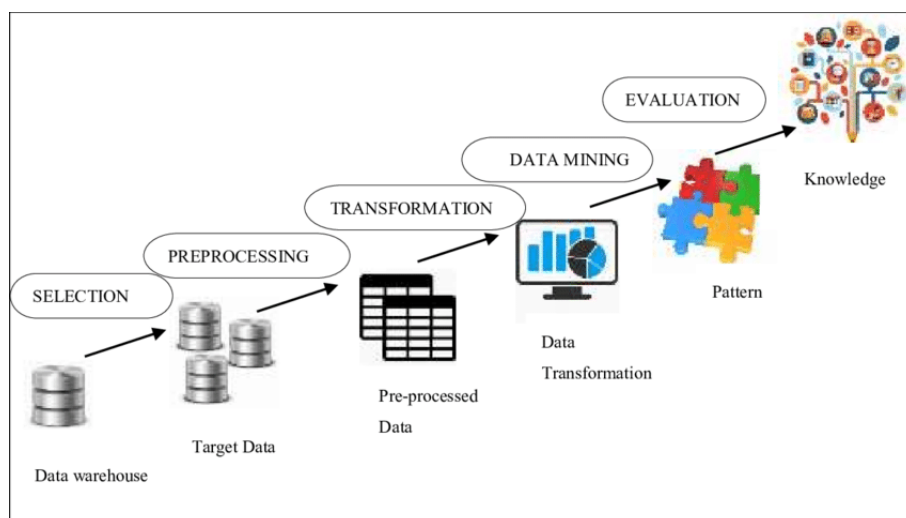
- Integrità dei dati;
- Dimensione del campione su cui è stata effettuata la scoperta;
- Il grado di supporto dalla conoscenza del dominio disponibile.

Un pattern è definito **interessante** quando è:

- Nuovo;
- Utile;
- Non banale da calcolare.

Vi sono principalmente due differenti interpretazioni dei pattern e dei modelli. La *prima interpretazione* che possiamo dare consiste nel definire un **modello** come una sintesi globale del data-set, mentre il **pattern** è una caratteristica locale del data-set, limitato a un sub-set di osservazioni e/o attributi. La *seconda interpretazione* esplica come il data mining implica l'adattamento o la determinazione di pattern da dati osservati. In questo caso il pattern è visto come una istanza del modello. I modelli adattati svolgono il ruolo di conoscenza dedotta.

Il KDD è un processo iterativo e interattivo, costituito da molti passaggi che includono molte decisioni prese dall'utente.



In una prima fase si sviluppa una **comprensione del dominio** dell'applicazione, delle relative conoscenze di base e degli obiettivi dell'utente finale. Si prosegue creando un **data set target**, selezionando un dataset o ci si focalizza su un sotto

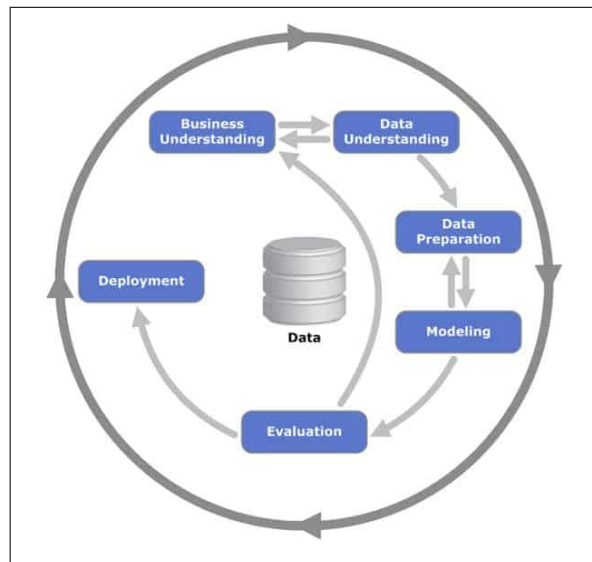
insieme di variabili o di esempi di dati sui quali deve essere eseguita la scoperta. Una volta selezionati questi dati si passa alla fase di **pulizia dei dati e preprocessing** nella quale si effettuano svariate operazioni volte a rimuovere il rumore o dei valori anomali, collezionando le informazioni necessarie per modellarle o tenere conto del rumore, decidere strategie per gestire i campi di dati mancanti o tenere conto delle informazioni sulla sequenza temporale e delle modifiche note. Successivamente si entra in una fase di **riduzione e proiezione dei dati**, nella quale si trovano funzioni utili per rappresentare i dati a seconda dell'obiettivo dell'attività. Vengono utilizzati metodi di riduzione o trasformazione multidimensionale per ridurre l'effettivo numero di variabili da considerare o per trovare rappresentazioni di dati invarianti. Conclusa questa attività, si passa alla **selezione del task di data mining**, decidendo qual è l'obiettivo del processo KDD, scegliendo tra classificazione, regressione, clustering.

Definito l'obiettivo si passa alla **scelta degli algoritmi** di data mining. Vengono selezionati i metodi che verranno utilizzati per ricercare i pattern frequenti all'interno dei dati. Vengono definiti i modelli e i parametri più appropriati e si cerca di far corrispondere particolari metodi di data mining con i criteri globali dei processi di KDD. Quindi si entra nella fase di **Data Mining**, nella quale si ricercano modelli di interesse in una particolare forma rappresentativa. L'utente può contribuire significativamente in questa fase eseguendo correttamente i passaggi precedenti. Alla fine del processo di data mining, vi è la fase di **interpretazione dei pattern minati**, nella quale si effettuano le valutazioni sui risultati ottenuti ed eventualmente si ritorna ad iterare su uno degli step precedenti. Infine si **consolida la conoscenza scoperta**, incorporando la conoscenza ottenuta nel sistema per migliorarne le prestazioni o semplicemente si documenta per essere segnalata ad altre parti di interesse. Questa fase include anche il controllo e la risoluzione di particolari conflitti con conoscenze precedentemente estratte o scoperte.

1.1 Steps del KDD

La maggior parte del lavoro nel KDD è focalizzata sullo step del data mining. Anche se gli altri step sono considerati importanti per il successo dell'applicazione del KDD nella pratica. La necessità per una standardizzazione del processo della scoperta della conoscenza ha portato alla definizione dello standard industriale CRISP-DM. L'obiettivo di questo standard è quello di sviluppare un processo neutrale per condurre il KD, e di definire i compiti, i loro outputs, la terminologia e i problemi tipici di caratterizzazione.

Il modello di processo CRISP-DM per la scoperta di conoscenza consiste in sei fasi.



La sequenza delle fasi non è rigida. È possibile spostarsi avanti e indietro tra le diverse fasi a seconda dell'esito di ciascuna fase. Le frecce indicano le dipendenze più importanti/frequenti. Il cerchio esterno simboleggia la natura ciclica di un processo KDD che può continuare dopo l'implementazione di una soluzione. Ogni fase contiene un numero di task che produce specifici output.

1.1.1 Business Understanding

La *prima fase* di questo modello viene definita **Business Understanding**, il cui primo step è quello di determinare gli obiettivi commerciali. I requisiti minimi sono:

- un problema o un'opportunità commerciale percepita
- un certo livello di sponsorizzazione esecutiva

Sviluppare una definizione chiara e comprensibile dei bisogni aziendali non è un compito semplice. E' richiesta la collaborazione dei business analyst e dei data analyst. Questo passaggio del processo è anche il momento in cui iniziare a definire le aspettative. Alla fine di questo step vengono prodotti:

- **Il background**, che descrive le informazioni note sulla situazione aziendale all'inizio del processo;
- **Gli obiettivi di business**, che descrivono i principali obiettivi dal punto di vista del business;
- **I criteri economici di successo**, che definiscono le misure per risultati di alta qualità del progetto dal punto di vista del business.

Dopo aver definito gli obiettivi commerciali, si passa alla **valutazione della situazione**, con lo scopo di raccogliere informazioni sulle risorse, sui vincoli e sulle assunzioni. Alla fine di questo sotto processo verrà prodotto un inventario contenente:

- le risorse: personale, dati, calcoli;

- vincoli: schede di compilazione, questioni legali, comprensibilità;
- assunzioni: disponibilità dei dati;

Inoltre non dimentichiamo che vengono prodotti anche **un glossario di termini**, che copre la terminologia di business e di data mining, e **un'analisi costo-beneficio**, cioè un documento contenente le spese del progetto dovrebbero essere confrontate con i potenziali guadagni.

Successivamente si passa alla **determinazione degli obiettivi del Data Mining**. In questa fase si trasformano gli obiettivi commerciali in obiettivi del processo di Data Mining e si costruiscono i relativi criteri di successo. Possiamo classificare gli obiettivi del Data Mining in:

- Classification
- Estimation (produrre una stima, Estimation)
- Prediction
- Affinity grouping
- Clustering
- Description e Profiling

I due obiettivi primari del Data Mining tendono ad essere:

- **Predizione**, che include l'uso di alcune variabili indipendenti per predire valori sconosciuti o futuri che dipendono da altre variabili;
- **Descrizione**, nella quale non si fa una distinzione tra variabili dipendenti e indipendenti e si concentra sulla ricerca di modelli interpretabili dall'uomo che descrivono i dati.

L'obiettivo della classificazione è apprendere una funzione che mappa, o classifica, un dato in una predefinita classe. Questa tecnica è molto utilizzata per i database. Invece quando si parla di regressione, si ha come obiettivo apprendere una funzione che mappa un dato in una variabile di previsione reale. Quando si cerca di creare un procedimento per trovare le associazioni tra gruppi di variabili, si utilizza il metodo Affinity Grouping. Il clustering è un'attività descrittiva in cui si cerca di identificare un insieme finito di categorie o cluster per descrivere i dati. Le categorie possono essere mutuamente esclusive ed esaustive, oppure consistere in una rappresentazione più ricca come una gerarchia o categorie sovrapposte. La clasterizzazione viene utilizzata spesso per scoprire sotto-popolazioni omogenee, identificare sotto categorie, o analisi di dati. Strettamente correlato al clustering è il compito della stima della densità di probabilità, che consiste in tecniche per stimare dai dati la funzione di densità di probabilità multivariata congiunta di tutte le variabili/campi nel database. Per **Summarization** (riassunto o descrizione o profilazione), si intendono tutti quei metodi per trovare una descrizione compatta per un sottoinsieme di dati. Invece il task **Dependency modeling** (modellare le dipendenze), consiste nel trovare un modello che descrive dipendenze significative tra variabili. Esistono due tipi di modelli dipendenti:

- il livello strutturale di specifici modelli le cui variabili sono localmente dipendenti tra loro

- il livello quantitativo il cui modello specifica quanto le variabili sono dipendenti usando una scala numerica.

Il task di rilevamento di modifiche e deviazioni si focalizza sulla scoperta delle modifiche più significanti nei dati rispetto a valori misurati o normativi in precedenza. Consiste nel trovare un modello che descrive significanti dipendenze tra variabili.

Dopo aver definito il task del Data Mining si passa a **produrre un piano progettuale** per il raggiungimento degli obiettivi di data mining e quindi il raggiungimento degli obiettivi di business. L'output di questa fase è ovviamente un piano progettuale, che specifica l'insieme degli step per la restante parte del progetto, la sua durata, le risorse richieste, gli input, gli output e le dipendenze.

1.1.2 Data Understanding

Dopo aver contestualizzato il business in cui si andrà a progettare il sistema, si passa a una fase di **Data Understanding**, cioè di comprensione dei dati, che inizia con una fase di raccolta iniziale dei dati. Durante questa fase si accede ai dati rilevanti nell'inventario delle risorse e si produce un report iniziale dei dati raccolti, nel quale si elenca la posizione dei dati, i metodi usati per acquisirli e i problemi incontrati. Successivamente si passa a descrivere i dati esaminando le loro proprietà e creando un report nel quale si descrive il formato dei dati, i potenziali valori, la quantità, l'identificatore dei campi e tutte le caratteristiche che vengono scoperte.

Ci sono due metodi principali per descrivere le variabili:

- **variabili categoriche**: i possibili valori finiti e i differenti tipi che una variabile può assumere. Questa categoria si può suddividere in:
 - **variabili nominali** che denominano il tipo di oggetto a cui si riferiscono, ma non esiste un ordine tra i valori possibili. (stato del materiale, genere, livello di educazione)
 - **variabili ordinali** che assumono un ordine tra i possibili valori. (valutazione del cliente)
- **variabili quantitative**: sui quali sono consentite operazioni aritmetiche, e si suddividono a loro volta in:
 - **variabili discrete**, i cui valori sono interi
 - **variabili continue**, i cui valori sono numeri reali.

Dopo aver descritto la tipologia di dati si passa alla verifica della loro qualità ispezionando e affrontando diverse caratteristiche quali:

- **Accuratezza**: conformità del valore memorizzato rispetto a quello effettivo
- **Completezza**: nessun valore mancante
- **Consistenza**: rappresentazione uniforme
- **Attualità**: i dati storicizzati non sono obsoleti.

Scarsa qualità dei dati e scarsa integrità dei dati sono i maggiori problemi all'interno dei progetti KDD. La maggior parte dei dati operativi non è mai stata acquisita o modellata per scopi di data mining. I dati selezionati vengono generalmente

raccolti da numerosi sistemi operativi, incoerenti e scarsamente documentati. È importante comprendere la **sensibilità temporale** dei dati. Lo specialista della gestione dei dati è responsabile della raccolta e dell'integrazione dei dati nell'ambiente informativo.

Al termine della verifica della qualità dei dati viene generato un report di qualità dei dati che riporta i risultati della verifica e se vi sono dei problemi, sarà possibile discutere di eventuali soluzioni.

A valle della verifica della qualità dei dati, è possibile avviare la fase di **esplorazione dei dati**, alla fine della quale seguirà un relativo report dei risultati ottenuti. Per le *variabili categoriche*, le distribuzioni della frequenza dei dati sono il metodo migliore per capire il contenuto dei dati. Istogrammi e grafici a torta aiutano a identificare gli schemi della distribuzione e i valori mancanti o non validi. Mentre quando lavoriamo con *variabili quantitative*, l'analista dei dati è interessato a misure come il massimo e il minimo, la media, moda, mediana e misure statistiche. Se combinate, queste misure offrono un modo efficace per determinare la presenza di dati non validi e distorti.

1.1.3 Data Preparation

Lo step successivo al Data Understanding vi è lo step di **Data Preparation**, la cui prima fase consiste nel selezionare i dati, nella quale il problema maggiore consiste nel selezionare dati da tuple di database relazionali. Si possono comunque seguire alcuni principi che includono:

- la rilevanza del dato rispetto all'obiettivo principale.
- vincoli tecnici e qualitativi,
- limiti al volume dei dati o ai tipi di dati

In questa fase si produce un report di inclusione/esclusione dei dati. La selezione dei dati può essere eseguita manualmente o automaticamente (campionamento e selezione delle caratteristiche).

Il più semplice tipo di campionamento è il **Simple random sampling** (campionamento semplice casuale): ogni gruppo di oggetti della dimensione richiesta ha la stessa probabilità di essere il campione selezionato. È possibile ottenere un campione molto atipico, tuttavia le leggi della probabilità impongono che più ampio è un campione, più è probabile che sia rappresentativo della popolazione da cui proviene.

Il metodo tradizionale di scelta di un campione casuale inizia con una numerazione dei membri della popolazione target. L'ordine di numerazione è irrilevante. Una volta che ogni membro della popolazione ha un numero, il campionario consulta una tabella di numeri casuali per selezionare gli indici dei membri da includere nel campione. Ovviamente il presupposto principale è quello di essere in grado di numerare i membri della popolazione target.

Vi sono due tipi di campionamento semplice casuale:

- con sostituzione
- senza sostituzione

Se la popolazione è grande rispetto al campione, c'è una probabilità molto piccola che qualsiasi membro venga scelto più di una volta e le due tecniche sono essenzialmente le stesse.

Quando la popolazione è divisa in strati o gruppi, è utile invece usare un **campionamento casuale stratificato**. Viene selezionato un campione casuale semplice da ciascuno strato separatamente e la loro unione produce un campione stratificato. Un vantaggio di questo campionamento consiste nel fatto che l'analista può controllare il numero di osservazioni all'interno di ogni gruppo o strato e può garantire che particolari gruppi all'interno della popolazione sono adeguatamente rappresentati nel campione. Quando uno strato ha un'appartenenza molto più piccola degli altri, il semplice campionamento casuale può produrre campioni senza rappresentativi di quello strato. La dimensione del campione è solitamente proporzionale alla dimensione relativa degli strati. Tuttavia, questa non è una regola.

Se i membri all'interno degli strati sono più simili tra loro rispetto ai membri di strati diversi, le stime specifiche per strato saranno più precise di quelle dell'intero campione. Attenzione però, è importante adeguarsi alla sovra rappresentazione quando le inferenze si riferiscono alla popolazione target nel suo insieme.

Alcune regole per una buona rappresentazione di progettazione per un campionamento stratificato, sono:

- gli strati devono essere scelti per:
 - avere dei mezzi che differiscono sostanzialmente tra loro
 - minimizzare la varianza all'interno di uno strato e massimizzarla tra i vari strati
- Le dimensioni del campione devono essere proporzionali alla deviazione standard dello strato.

Un'altra tecnica di campionamento è il **cluster sampling** o campionamento clusterizzato, nel quale i membri della popolazione arrivano naturalmente da cluster, ciò rende possibile campionare i cluster. In questo caso tutti i membri di ogni cluster sono considerati. Nel contesto di grandi basi di dati, un'applicazione comune del cluster sampling è di rendere casuale la scelta di blocchi di dati, e successivamente usare tutti i dati nei blocchi. La motivazione dietro questo approccio è che per recuperare un record di database da un blocco particolare, l'intero blocco deve essere letto in memoria. Questo tipo di campionamento è anche chiamato **block sampling**. I vantaggi del cluster sampling riguardano la riduzione dei costi richiesti per accedere ai campioni, e al tempo stesso aumenta la variabilità delle stime campionarie al di sopra di quella del semplice campionamento casuale, a seconda di quanto i cluster differiscono tra loro, rispetto alla variazione all'interno del cluster.

Se i membri di un cluster sono più simili dei membri di cluster diversi, gli approcci statistici che presuppongono che i dati siano indipendenti porteranno a inferenze distorte. La modellazione gerarchica può modellare esplicitamente la struttura indotta dall'amplificazione nei dati.

Il **two-stage sampling** (campionamento a due stadi) combina due idee principali: la scelta casuale dei cluster e il campionamento all'interno di ogni cluster.

Quando è possibile numerare in qualsiasi modo gli individui di una popolazione, si può effettuare il **systematic sampling** (campionamento sistematico), conosciuto anche come **every k-th sampling**. Questo tipo di campionamento si sviluppa scegliendo un membro in maniera casuale da quelli numerati tra 1 e k , successivamente include ogni k -th membro dopo il campione. Il vantaggio maggiore di questo tipo di campionamento è la sua facile implementazione, anche nel caso in cui la dimensione della popolazione è inizialmente sconosciuta o il conteggio dei membri della popolazione è computazionalmente costoso. Al contempo bisogna prestare attenzione, dato che la selezione non è casuale, campioni sistematici possono non essere rappresentativi della popolazione e quindi devono essere utilizzati con attenzione. Questo metodo è particolarmente vulnerabile alle periodicità nell'elenco dei membri della popolazione. Se la periodicità è presente e il periodo è un multiplo di k , risulterà una distorsione.

Quando vogliamo organizzare il nostro campionamento basato sul valore di una o più variabili, ma non conosciamo l'intervallo o la distribuzione di queste variabili nella popolazione target, possiamo avvalerci del **two-phase sampling** o campionamento in due fasi. Un campione iniziale può facilitare prendere decisioni più consapevoli sulle strategie di campionamento da utilizzare. Un campione iniziale può aiutare a determinare la dimensione del campione. I calcoli delle dimensioni del campione spesso richiedono stime di determinati parametri della popolazione, come la forza della relazione tra due variabili. In assenza di conoscenze e/o dati pregressi, un campione iniziale fornirebbe stime per queste quantità e queste stime determinerebbero la dimensione del campione per il campione della seconda fase.

Una domanda che ci si pone spesso è quanto un campione deve essere grande. In statistica vi sono semplici meccanismi per stimare la dimensione del campione necessaria per avere una certa probabilità di rilevare un effetto di una dimensione pre-specificata o superiore. Questi meccanismi, che raggruppati prendono il nome di *analisi di potenza*, si basano su stime delle medie e varianze delle variabili nella popolazione da campionare. E' importante capire il sistema di cause che determinano popolazione e garantire che tutte le fonti di variazione siano prese in considerazione. Un gran numero di osservazioni non ha alcun valore se le principali fonti di variazione vengono trascurate nello studio.

Un'alternativa ad impostare una dimensione del campione predeterminata consiste nel lasciare che i dati "scelgano" la dimensione del campione. L'idea di base è continuare ad aumentare la dimensione del campione fino a quando i risultati o i riepiloghi non cambiano più molto (dove "molto" è impostato in anticipo). Le varianti di questa idea sono note come campionamento progressivo, campionamento adattivo o campionamento sequenziale.

Bisogna però far attenzione al fatto che molti degli strumenti comuni per l'inferenza statistica, inclusi i t-test, presuppongono che i dati comprendano un semplice campione casuale di una popolazione e che i singoli punti dati siano quindi statisticamente indipendenti. Molte delle tecniche di campionamento violeranno questa ipotesi. Esistono tecniche statistiche specializzate per trattare i dati generati da un numero qualsiasi di campionamento casuale da un database fornisce stime imparziali delle caratteristiche del database. Tuttavia, se il database stesso rappresenta un campione casuale o sistematicamente distorto dalla popolazione reale di interesse, nessuna tecnica statistica può salvare le inferenze risultanti.

I database in tempo reale contengono degli attributi (chiamati anche **features** o caratteristiche). Il problema della selezione delle caratteristiche sorge perchè la complessità di ricerca nello spazio delle ipotesi deve essere ridotta per ragioni pratiche, e caratteristiche ridondanti o irrilevanti possono avere effetti significativi sulla qualità dei risultati del metodo di analisi (**maledizione della dimensionalità**). L'idea alla base della maledizione della dimensionalità è che dati di grande dimensione sono difficili da processare, per una serie di motivi: il numero di esempi cresce esponenzialmente con il numero di variabili e non ci sono abbastanza osservazioni per ottenere buone stime.

La **feature selection** è un processo che scegli un subset ottimo di features seguendo alcuni criteri. Vi sono sostanzialmente 3 approcci:

- **wrapper models**
- **filter models**
- **embedded methods**

Il *wrapper model* si basa su un algoritmo di data mining per determinare se un sottoinsieme di funzionalità è valido. L'algoritmo viene utilizzato come parte della funzione di valutazione e anche per indurre i patterns o il modello finale.

L'algoritmo DM può risolvere task predittivi o task descrittivi. Se vogliamo avere un buon insieme di funzionalità per miglio-

rare l'accuratezza di un classificatore, possiamo utilizzare proprio questa misura per basare le evoluzioni del classificatore. Però sorgono diversi problemi, il principale consiste nel determinare veramente l'accuratezza predittiva evitando l'overfitting. Altri problemi consistono nel fatto che un classificatore richiede tempo per apprendere i dati, oppure i dati sono troppo grandi per eseguire un algoritmo di apprendimento, quindi è necessario ridurre la dimensionalità. Vi è quindi la necessità di definire alcuni criteri di stop per garantire che il processo di valutazione termini e che non entri in un loop infinito.

Il filter models è indipendente dall'algoritmo di data mining che sarà utilizzato sul subset di feature. I subset sono valutati utilizzando proprietà intrinseche dei dati quali le misure informative e di distanza. Se consideriamo l'accuratezza stimata da un classificatore come un'altra misura, possiamo unificare i modelli di filtering e wrapper in un modello generale. Ogni componente può avere diverse scelte. Le varie combinazioni di queste scelte sono alla base di molti algoritmi di selezione delle caratteristiche esistenti.

Alcune strategie per la selezione delle caratteristiche dei subset sono:

- **enumerazione** di tutte i possibili subset e selezione dei migliori
- **generazione casuale** dei subset e selezione del migliore.
- **generazione sequenziale** dei subsets.

Possiamo inoltre selezionare i subset attraverso la **forward selection**, iniziando con un subset vuoto e gradualmente aggiungere una caratteristica alla volta, oppure attraverso la **backward selection**, nel quale si inizia con un insieme completo e si rimuove una caratteristica alla volta. Queste strategie sono basate sulla **strategia di ricerca greedy** in uno spazio di subset grande 2^N , dove N è il numero di caratteristiche.

Indipendentemente dal metodo di generazione dei sottoinsiemi di funzionalità adottato, è necessaria una misura per decidere quale funzionalità deve essere aggiunta o rimossa, oppure quale sottoinsieme deve essere mantenuto. Questo è possibile grazie alla **misura dell'informazione**, data una funzione di incertezza **U** e le probabilità delle classi precedenti **P**(c_i), l'informazione ottenuta da una caratteristica X, **IG(X)**, è definita come la differenza tra la precedente incertezza e l'incertezza posteriore attesa usando X. In formula

$$IG(X) = \sum_i U(P(c_i)) - E[\sum_i U(P(c_i|X))] \quad (1)$$

Una regola di valutazione delle caratteristiche derivata dal concetto di guadagno di informazioni afferma che la caratteristica X è preferita alla caratteristica Y se $IG(X) > IG(Y)$. Cioè, una caratteristica dovrebbe essere selezionata se può ridurre più incertezza. Se $U(x) = -x * \log(x)$ allora $\sum_i U(P(c_i))$ è una misura di entropia.

Misure della distanza. Se l'obiettivo è la classificazione, una misura alternativa può essere la distanza tra le funzioni di densità classe-condizione. Se $P(X|C)$ è la funzione di densità condizionata dalla classe della caratteristica X, nei due casi della classe ($C=c_1$ o $C=c_2$), $P(X|C)$ è definita da $P(X|c_1)$ e $P(X|c_2)$. Se $D(X)$ è la distanza tra $P(X|c_1)$ e $P(X|c_2)$, una regola di valutazione basata su P afferma che X è preferito a Y se $D(X) > D(Y)$. Questo perché stiamo cercando una feature che può separare due classi il più possibile. Maggiore è la distanza, più facile separare le due classi.

La divergenza Kullback-Leibler è una misura della distanza tra distribuzioni probabili, P e Q su un dominio V. Questa divergenza è definita come:

$$m_{KL}(P, Q) = \sum_{v \in V} q(v) \log\left(\frac{q(v)}{p(v)}\right) \quad (2)$$

Questa divergenza misura in quale misura la distribuzione P è un'approssimazione della distribuzione Q o, più precisamente, la perdita dell'informazione se noi prendessimo P invece di Q. In sintesi si misura quanto P diverge da Q. LE proprietà di questa misura sono:

- Assimetrica, ovvero $m_{KL}(P, Q) \neq m_{KL}(Q, P)$
- non definita quando $p(v)=0$
- nel caso specifico di $q(v)=0$, $q(v)\log(q(v)/p(v))=0$
- il range dei valori non è limitato. Quindi, possiamo utilizzare questo valore per stabilire quale delle due distribuzioni Q e Q' è una migliore approssimazione di P, ma non ci permette di determinare in termini assoluti se Q è una buona approssimazione di P osservando $m_{KL}(P, Q)$

La divergenza del χ^2 è definita come:

$$m_{\chi^2}(P, Q) = \sum_{y \in Y} \frac{|p(y) - q(y)|^2}{p(y)} \quad (3)$$

è rigorosamente topologicamente più forte della divergenza KL data la disuguaglianza $m_{KL}(P, Q) \leq m_{\chi^2}(P, Q)$, la convergenza nella funzione di divergenza χ^2 implica la convergenza nella divergenza KL, ma non il contrario. Inoltre è asimmetrica e non definita quando $p(y)=0$.

Un ultimo tipo di misura della distanza è la distanza di variazione, data dalla formula $m_1(P, Q) = \sum_{y \in Y} |p(y) - q(y)|$, detta anche distanza di Manhattan per funzioni di probabilità $p(y)$ e $q(y)$ e coincide con la distanza di Hamming quando tutte le features sono binari. Similarmente si può utilizzare la distanza Euclidea data da $m_2(P, Q) = \sum_{y \in Y} |p(y) - q(y)|^2$. Queste due metriche soddisfano la proprietà di simmetria e le proprietà metriche.

Dipendenza dalle misure. Si verifica con quanta forza una caratteristica è associata alla classe. Denotando attraverso $R(X)$ una misura di dipendenza tra la feature X e la classe C, preferiamo la feature X alla feature Y se $R(X) > R(Y)$. Un problema con le tre misure precedenti è che non possono rompere i legami tra due features ugualmente buone. Pertanto queste funzionalità non possono rilevare se una di esse è ridondante.

Inconsistenza delle misure. Si tenta di trovare un numero minimo di funzionalità che separano le classi in modo coerente come può fare l'intero set di funzionalità. In altre parole, le misure di incoerenza mirano a raggiungere $P(C|FullSet) = P(C|SubSet)$. Le regole di valutazione delle funzionalità derivate dalle misure di incoerenza affermano che è necessario selezionare il sottoinsieme minimo di funzionalità in grado di mantenere la coerenza dei dati mantenuti dall'insieme completo di funzionalità.

Per **selezionare le feature** ci sono 3 componenti necessari: un generatore di subset, un valutatore e un criterio di stop. Vi sono diversi metodi.

Approcci completi ed esaustivi: il *focus* è applicato su una misura di consistenza e valuta esaustivamente tutti i subset partendo da una feature, mentre il *branch and bound* consiste in una enumerazione sistematica di tutte le soluzioni, dove

ampi sottoinsiemi di candidati infruttuosi sono scartati in massa utilizzando dei limiti superiori e inferiori della quantità da ottimizzare. Inizia con un insieme completo di feature e valuta l'accuratezza stimata.

Approcci Euristici: *SFS* (sequential forward search) e *SBS* (sequential backward search) possono essere applicati a ciascuna delle misure, *DTM* è la più semplice versione di una modalità di wrapper - impara un classificatore una volta e usa qualsiasi caratteristica trovata nel classificatore.

Approcci non deterministici: *LVF* (Las Vegas Filter) e *LVW* (Las Vegas wrapper), generano subset di feature casualmente e li testano in maniera differente, LVF applica una misura inconsistente, LVW usa una stima accurata attraverso un classificatore; *algoritmi generici* e *ricottura simulata* sono anche usati nella selezione di feature. Il primo può produrre più sottoinsiemi, il secondo produce un singolo sottoinsieme.

Approcci basati sulle istanze: *Relief*, molti piccoli campioni di dati sono campioni provenienti dai dati. Le funzionalità vengono ponderate in base ai loro ruoli nella differenziazione di istanze di classi diverse per un campione di dati. È possibile selezionare funzioni con pesi maggiori.

Per le attività di data mining non di classificazione (nessuna etichetta di classe disponibile), dovrebbero essere presi in considerazione metodi alternativi. Ad esempio, una misura di entropia può essere introdotta per classificare in sequenza le caratteristiche. L'idea di base è che le caratteristiche sono rilevanti se possono descrivere le istanze in termini di cluster relativamente chiaramente definiti.

La **scalabilità** è un altro problema. In LVS la parte più dispendiosa in termini di tempo di un processo di selezione delle funzionalità viene identificata e ritardata fino a quando non è necessario. LVS è un'estensione di LVF che utilizza una misura di incoerenza (IC) con una complessità di runtime di controllo $O(n)$, dove n è il numero di istanze. Se n è enorme, è costoso calcolare IC molte volte. Si noti inoltre che il componente di generazione di sottoinsiemi di funzionalità genererà sempre più sottoinsiemi non validi che non soddisfano IC man mano che la cardinalità di un sottoinsieme valido diminuisce.

Pertanto, ha senso separare il calcolo di IC come cardinalità per tutti i dati dalla generazione di sottoinsiemi di funzionalità. Ma abbiamo bisogno di dati per generare sottoinsiemi di funzionalità. Il compromesso è che invece di utilizzare l'intero set di dati, ne utilizziamo solo una parte per la generazione di sottoinsiemi di funzionalità. Quando un sottoinsieme viene testato come valido sulla porzione di dati, viene calcolato l'IC per l'intero dato. Successivamente se IC è stato soddisfatto, allora la selezione della feature è completata, altrimenti se IC non è stato soddisfatto, le istanze inconsistenti vengono aggiunte alla porzione di dati e viene eseguito un altro ciclo di generazione di sottoinsiemi di funzionalità sulla porzione di dati ingrandita. Questo metodo è particolarmente efficace solo quando n è sufficientemente largo a causa del sovraccarico nel LVS.

I **wrapper models** cercano di risolvere uno specifico problema, quindi il criterio può realmente essere specificato. Invece consuma molto tempo se bisogna valutare uno schema a ogni iterata. I **filter models** sono molto più veloci ma non incorporano algoritmi di data mining usati per la generazione di un modello o pattern, quindi il modello può essere subottimale.

A differenza dei precedenti, nei **metodi embedded** la parte di selezione delle caratteristiche non può essere separata dall'algoritmo di data mining. È parte integrante dell'algoritmo. Per esempio, nell'algoritmo di decision tree, la selezione delle caratteristiche che contribuiscono alla creazione dell'albero finale è parte della costruzione dell'algoritmo di decision tree. Poiché siamo interessati a selezionare le funzionalità nella fase di trasformazione dei dati, non consideriamo gli embedded methods.

La **pulizia dei dati** aumenta la qualità dei dati al livello richiesto. Ciò comporta la selezione di sottoinsiemi di dati puliti, l'inserimento di impostazioni predefinite adeguate o tecniche più ambiziose come la stima della modellazione dei dati

mancanti. Alla fine di questa fase vi è un rapporto sulla pulizia dei dati, che descrive le decisioni e le azioni per affrontare i problemi di qualità dei dati ed elenca le trasformazioni dei dati per la pulizia e i possibili impatti sull'analisi dei risultati. I due problemi più comuni sono la mancanza di valori e dati di rumore.

I **noisy data**, ovvero tutti quei dati che generano rumore, consistono in variabili che hanno valori non conformi con quelli che ci aspettiamo dalle variabili. Le osservazioni in cui si verificano questi noisy data sono chiamate valori anomali, ovvero (*outliers*). Differenti tipi di outliers possono essere trattati in differenti modi.

Vi possono essere errori umani, i quali possono essere corretti o cancellati dall'analisi, oppure le distribuzioni simmetriche spesso indicano valori anomali. La mancanza di dati, invece, include valori che non sono presenti nei dati selezionati e necessitano di essere eliminati durante il rilevamento del rumore. Questo caso si verifica se viene commesso un errore durante l'inserimento dei dati, oppure se l'informazione non era disponibile al momento dell'inserimento oppure se i dati selezionati all'interno di risorse eterogenee hanno creato dei mismatch.

Per correggere questo tipo di errore vengono eseguite diversi tipi di azioni, l'inserimento di un valore predefinito come il termine "none" è l'ideale. Si potrebbe cancellare le righe che presentano valori mancanti, però questa tecnica, per quanto facile da implementare, può generare la perdita di dati che possono essere valutati.

Un'altra tecnica consiste nell'eliminazione della variabile dall'analisi se presenta un significativo numero di osservazioni con valori mancanti per la stessa variabile. Infine, un'ultima tecnica, consiste nel rimpiazzare il valore mancante con un altro valore, che nel caso di variabili quantitative può consistere nella media o nella mediana, mentre per le variabili categoriche può essere rappresentato dalla moda o dal valore "sconosciuto". Si potrebbe anche pensare di attuare un approccio più sofisticato è quello di predire il valore più probabile delle variabili all'interno delle osservazioni.

Successivamente si passa alla fase di **costruzione dei dati**, la quale include operazioni di preparazione dei dati, come la generazione di variabili derivate, inserimento di nuovi record o trasformazione di variabili esistenti. I dati possono essere trasformati in *una singola variabile*, per essere ulteriormente perfezionati per soddisfare i requisiti del formato di input dei particolari algoritmi di data mining da utilizzare.

Esempi sono la conversione delle variabili di tipo data dal formato US a quello Europeo, oppure il calcolo dell'età data la data di nascita, l'aggregazione dei valori all'interno di un conto corrente per gli ultimi 3,6,12 mesi. Inoltre si potrebbe effettuare un ridimensionamento o una normalizzazione dei dati. In questo caso si parla di normalizzazione dei dati quando colonne numeriche sono trasformate usando funzioni matematiche in dei range. Questo processo è importante perchè tutte le variabili all'interno di una colonna devono essere trattate in maniera uguale e non devono influenzarsi a vicenda, oppure perchè alcuni dati possono ricevere solo alcuni valori all'interno di un range.

Un altro tipo di normalizzazione è la normalizzazione min-max che performa una trasformazione lineare sui dati originali. Supponendo che min_A e max_A sono i valori di minimo e di massimo di un attributo A, questa normalizzazione mappa un valore v di A in v' nel range $[newMin_A; newMax_A]$ attraverso la formula:

$$v' = \frac{v - min_A}{max_A - min_A} (newMax_A - newMin_A) + newMin_A \quad (4)$$

Le relazioni tra i valori dei dati originali vengono preservate. Se un caso di input futuro per la normalizzazione non rientra nell'intervallo di dati originale per A, si verifica un errore "out of bounds" (fuori dal limite).

Nella normalizzazione definita z-score, anche chiamata a media zero, i valori per un attributo A sono normalizzati sulla base della media o della deviazione standard di A. Un valore v di A è normalizzato in v' attraverso la formula:

$$v' = \frac{v - mean_A}{standDev_A} \quad (5)$$

dove $mean_A$ e $standDev_A$ sono la media e la deviazione standard dell'attributo A. Questo metodo di normalizzazione è utile quando il minimo e il massimo dell'attributo A sono sconosciuti, o quando ci sono gli outliers che dominano la normalizzazione min-max.

La normalizzazione attraverso il decimal scaling trasforma i valori dell'attributo A in punti decimali, per assicurarsi che il range dell'intervallo in cui essi sono compresi sia $\{-1, +1\}$. Il numero di punti decimali dipende dal massimo valore assoluto di A. Un valore v di A è normalizzato in v' attraverso la funzione:

$$v' = \frac{v}{10^j} \quad (6)$$

dove j è il più piccolo intero tale che $\max(|v'|) < 1$

Si parla di **discretizzazione delle variabili** quando convertiamo variabili quantitative in variabili categoriche dividendo il valore di input in intervalli. Due metodi di discretizzazione sono l'**Eguale Width** e l'**Equal depth**. In entrambi questi metodi le informazioni sulla classe non vengono utilizzate nel caso in cui le osservazioni siano preclassificate, inoltre la discretizzazione viene applicata a ogni attributo indipendentemente dagli altri.

Il partizionamento attraverso l'equal width divide il range in N intervalli di uguale lunghezza, se A e B sono il più piccolo e il più grande valore all'interno dell'attributo, l'intervallo di width sarà:

$$W = \frac{B - A}{N} \quad (7)$$

Questo metodo è il più diretto, però ha anche degli aspetti negativi perché non gestisce bene i dati distorti ed è sensibile ai valori anomali.

Il partizionamento attraverso l'equal-depth divide il range in N intervalli, ognuno contenente approssimativamente lo stesso numero di esempi. Questo tipo di partizionamento ha una buona scalabilità, gestisce gli attributi categorici attraverso alcuni trucchetti, minimizza le informazioni perse durante il processo di partizionamento.

In entrambi i casi il numero di elementi tralasciati è definito dall'utente. Nel caso di classificazione dei dati, è buona pratica che questo numero non sia minore del numero di classi che vogliamo riconoscere, oppure determinarlo attraverso la formula:

$$N_{bins} = \frac{M}{3 * C} \quad (8)$$

dove M è il numero di esempi di training e C il numero di classi.

Procedure complesse di conversione di dati hanno l'obiettivo di costruire un piccolo insieme di indici da un vasto numero di variabili, in modo tale che solo alcune informazioni vengano perse e il numero totale di funzioni venga ridotto. **L'analisi fattoriale** e **l'analisi del componente principale** sono due tecniche di analisi multivariate per la riduzione dei dati.

L'analisi fattoriale affronta il problema dell'analisi della struttura delle interrelazioni (correlazioni) tra un grande numero di variabili (ad es. punteggi dei test, elementi del test, risposte al questionario) definendo un insieme di dimensioni sottostanti

comuni, note come fattori. Questa non è una tecnica che usa dipendenze, ovvero non vi è nessuna variabile considerata come criterio o variabile dipendente da cui tutte le altre dipendono e sono le variabili predittore o indipendenti. E' una tecnica interdipendente in cui ogni variabile è considerata simultaneamente e ognuna è in relazione alle altre.

Oltre alla costruzione dei dati, vi è l'integrazione dei dati, che ha come obiettivo la combinazione di informazioni provenienti da tabelle multiple o record per creare nuovi record o valori. L'output di questa fase è un insieme di dati uniti, che si riferisce all'unione di due o più tabelle unite che contengono differenti informazioni sugli stessi oggetti, o dati aggregati, sui quali verranno effettuate delle operazioni per produrre nuovi dati. Il risultato è il modello dei dati analitici, che rappresenta una ristrutturazione consolidata, integrata e dipendente dal tempo dei dati selezionati e pre-elaborati dalle varie fonti operative ed esterne.

Una decisione importante presa in questo compito riguarda l'unità di analisi. Nella ricerca nelle scienze sociali, le unità di analisi più tipiche sono le persone individuali. Altre unità di analisi possono essere gruppi, manufatti (libri, foto, giornali), unità geografiche (città, censimento, stato), interazioni sociali (relazioni diadiche, divorzi, arresti). L'unità di analisi non deve essere confusa con le unità di osservazione, che è l'unità su cui vengono raccolti i dati, cioè vengono effettuate osservazioni sistematiche. Ad esempio, uno studio può avere un'unità di osservazione a livello individuale (ad es. alunni), ma può avere l'unità di analisi a livello di gruppo (ad es. una classe). L'unità di osservazione è la stessa dell'unità di analisi quando le generalizzazioni ricavate da un'analisi statistica sono attribuite all'unità di osservazione (cioè gli oggetti su cui i dati sono stati raccolti e organizzati per l'analisi statistica).

Un'altra procedura consiste nel formattare la sintassi di alcuni tipi, non andando ad alterare i significati, perchè questi formati sono richiesti da altri modelli.

1.1.4 Modeling

In questa fase viene scelta la tecnica di modellazione tra le tecniche e gli strumenti preselezionati, come output si hanno le tecniche e le assunzioni modellate dalle tecniche prese in input. Ci sono diversi fattori su cui basare la scelta dell'algoritmo o degli algoritmi per modellare i dati, tra cui la natura dell'obiettivo, la capacità di gestire determinati tipi di dati, la capacità di gestire multiple relazioni e generare pattern relazionali, la scalabilità o il livello di familiarità.

Si possono identificare tre componenti primari in qualsiasi algoritmo di data mining:

- **modello di rappresentazione:** relazionale o proposizionale, qualitativo o quantitativo, adeguatezza per un utente.
- **modello di valutazione:** l'incertezza del modello si basa su stime probabilistiche, test statistici ecc.
- **ricerca**

Il *modello di rappresentazione* consiste nel linguaggio L usato per descrivere patterns che possono essere scoperti. Dipende dal tipo di rappresentazione, la capacità descrittiva, l'adeguatezza per un dato utente. La conoscenza scoperta può essere categorizzata dal tipo di pattern dei dati. Una scoperta quantitativa mette in relazione valori di campi numerici usando equazioni matematiche, mentre una relazione qualitativa mette in relazione logica diversi campi. Spesso si possono trovare queste due relazioni combinate. Un altro aspetto che bisogna considerare nella creazione del modello è la sua rappresentazione che deve essere appropriata per l'utente previsto.

La maggior parte degli utenti per i quali bisogna creare una rappresentazione sono gli umani, programmi per computer e sistemi di scoperta. Se si parla di utenti umani bisogna utilizzare un linguaggio naturale, molto utile ma non conveniente per

la manipolazione da parte di algoritmi di scoperta, oppure si potrebbe utilizzare un linguaggio logico, molto più utile per la computazione e se necessario può essere anche traslato in un linguaggio normale. Le informazioni sulla forma e la densità dei gruppi di record sono un altro tipo di conoscenza che si presenta al meglio visivamente per mezzo di grafici bidimensionali o tridimensionali, quindi attraverso opportune rappresentazioni visive.

Nel caso in cui gli utenti finali siano dei programmi, la rappresentazione include linguaggi di programmazione e formalismi dichiarativi. Quando invece l'utente è un sistema il cui compito è scoprire conoscenza, le nuove scoperte vengono reinserite nel sistema come conoscenza di dominio, la conoscenza di dominio e la conoscenza scoperta devono condividere una rappresentazione comune. Se la rappresentazione è limitata, nessun addestramento o nessun tipo di esempio produrrà un modello accurato per i dati. È importante che un analista di dati comprenda appieno i presupposti rappresentativi che possono essere inerenti a un particolare metodo.

Alcuni patterns sono spesso associati a un grado di incertezza, rappresentata attraverso la probabilità, la deviazione standard, misure di credenza o fuzzy sets. In maniera visiva l'informazione incerta può essere convertita in dimensione, densità e ombreggiatura.

Quando i database sono veramente grandi, con milioni di record, un'analisi completa di tutti i dati è infattibile. Gli algoritmi di rilevamento devono quindi basarsi su una qualche forma di campionamento, per cui viene considerata una parte dei dati. Le scoperte risultanti in questi casi sono necessariamente incerte. Le tecniche statistiche possono misurare il grado di incertezza. Possono anche essere utilizzati per determinare la quantità di campionamento aggiuntivo necessaria per ottenere il livello di fiducia desiderato nei risultati.

Il modello di valutazione stima l'incertezza di un particolare pattern. Ci sono diversi metodi per valutare questa incertezza. La **cross validation** è una di queste procedure che consiste nel mescolare i dati del dataset, partizionarli in k sottoinsiemi di uguale lunghezza n , per ogni sottoinsieme chiama i -esimo subset di n oggetti come insieme test e lo mette da parte, addestra il sistema sui rimanenti sottoinsiemi e testa il sistema sull'insieme di test e memorizza le performances, successivamente pulisce la memoria dimenticando ciò che ha imparato durante l'addestramento, calcola le performance medie di ogni insieme di test.

La scelta dell'algoritmo di data mining è fondamentale per conoscere se la ricerca dovrebbe essere performata nello spazio dei parametri o nello spazio dei modelli o entrambi. Se non è possibile una soluzione, si possono utilizzare metodi iterativi greedy.

La ricerca del modello avviene come un ciclo sul metodo di ricerca dei parametri, la cui rappresentazione viene modificata in modo da considerare una famiglia di modelli. Per ogni rappresentazione specifica del modello viene istanziata la modalità di ricerca dei parametri per valutare la qualità di quel particolare modello. Le implementazioni dei metodi di ricerca del modello tendono a utilizzare tecniche di ricerca euristica poiché la dimensione dello spazio dei possibili modelli vieta la ricerca esauriente e le soluzioni in forma chiusa non sono facilmente ottenibili.

Prima di costruire i modelli, è importante generare una procedura per testare la loro qualità. Successivamente la tecnica di modellazione è applicata ai dati e alla fine della quale vengono descritte le varie motivazioni delle scelte effettuate, le impostazioni dei parametri e l'output model con accuratezza aspettata, la robustezza e le possibili carenze.

L'output dell'algoritmo di data mining può essere espresso seguendo alcuni standard industriali. Il **Predictive Model Markup Language** (PMML) è uno standard basato su XML, il cui obiettivo è quello di definire e condividere modelli predittivi usando uno standard aperto. Un complesso mosaico di applicazioni software generano e consumano conoscenza, quindi si ha la necessità di una rappresentazione indipendente dal fornitore di output di data mining.

Alcuni benefici di questo standard consistono nell'eliminare i problemi e le incompatibilità di software proprietari per lo scambio di modelli tra applicazioni, oltre alla facilitazione nello sviluppo di modelli utilizzando qualsiasi fornitore e maggiore facilità nella distribuzione dell'applicazione.

Alla fine della creazione del modello, bisogna valutarlo. I risultati sono interpretati in base ai criteri di successo dell'algoritmo di data mining. Alcuni tipici criteri sono l'accuratezza del modello, quanto esso riesce a descrivere i dati osservati, quanta confidenza può essere riposta nella predizione. Alla fine di questa fase viene effettuata una valutazione dei modelli generati e revisionati i parametri delle impostazioni.

I modelli predittivi vengono valutati in base alla loro accuratezza su dati non visti in precedenza. La valutazione del modello può avvenire a livello dell'intero modello o a livello di singole previsioni. Due modelli con la stessa accuratezza complessiva possono avere livelli di varianza abbastanza diversi tra le singole previsioni. La valutazione del modello dovrebbe basarsi su un set di test indipendente dal set di formazione.

Per i modelli di classificazione, la stima dell'accuratezza è il numero complessivo di classificazioni corrette, diviso per il numero totale di campioni nel test set. Il tasso di errata classificazione (o errore) è il complemento della stima dell'accuratezza. La precisione potrebbe non essere appropriata quando gli errori possono avere costi (e conseguenze) diversi. In tal caso, un costo di errore è una misura migliore di un errore di classificazione errata. Per i modelli di regressione o per i modelli di stima dei parametri, l'accuratezza è stimata come la somma media degli errori quadrati (differenze tra valori previsti ed effettivi) (errore quadratico medio).

1.1.5 Valutazione

Mentre la valutazione del modello si occupa di fattori quali l'accuratezza e la generalità dei modelli, questa attività valuta i modelli in base agli obiettivi di business originali e ai criteri di successo. La sfida consiste nel presentare le nuove scoperte in modo convincente e orientato al business. Questa attività può essere eseguita correttamente da un analista di dati esperto che lavora con un analista aziendale.

Come output di questa operazione abbiamo una valutazione complessiva del data mining rispetto ai criteri di successo aziendale, inclusa una dichiarazione finale sul fatto che il progetto soddisfi già gli obiettivi aziendali iniziali. Questo passaggio rientra nel dominio dell'analista aziendale ed è focalizzato sullo sponsor esecutivo. L'analista aziendale esperto sarà in grado di formulare i risultati in un modo che si collega direttamente agli obiettivi aziendali che erano stati fissati per il progetto all'inizio. Inoltre, un analista di dati con una vasta conoscenza del settore e/o del settore può essere una grande risorsa in questa fase.

Il più grande vantaggio del progetto potrebbe essere un riconoscimento o una rinnovata attenzione all'importanza del patrimonio di dati aziendali e alla potenza delle soluzioni basate sui dati. Questo a sua volta può generare una cultura all'interno dell'organizzazione in grado di promuovere iniziative più ampie, a lungo termine e orientate ai dati, come il data warehousing e il data marting.

Oltre alla valutazione del modello, viene effettuato un processo di revisione nel quale viene considerato l'intero processo al fine di determinare se vi è qualche fattore o compito importante che viene trascurato o per identificare una procedura generica per generare modelli simili in futuro.

1.1.6 Deployment

Per distribuire i risultati del data mining nell'azienda, questa attività sviluppa una strategia per il deployment. In questa fase si pianifica il piano per il deployment, considerando anche l'infrastruttura tecnica che deve essere impiegata per il corretto funzionamento dell'applicativo sviluppato. Successivamente a questa fase viene creato un piano per il monitoraggio e il mantenimento dell'applicazione. Viene redatto un report del prodotto finale e in fase di revisione si valuta cosa è andato bene, cosa è andato male e cosa bisogna migliorare.

2 Learning sets of rules

2.1 Classificatore basato su regole

Un classificatore è espresso in uno stile logico costituito da un insieme di regole di decisione (if-then), che rappresentano la conoscenza nel miglior modo possibile in termini di espressione e comprensione. Vi sono due tipologie di regole: le **regole per l'apprendimento di attributi** e le **regole per la descrizione relazionale**.

Nel **concept learning**, apprendimento dei concetti, si hanno due punti di vista opposti: *estensionale*, quando si ha un insieme di oggetti fisici o astratti, o *intenzionale*, quando si hanno un insieme di condizioni sufficienti e necessarie. In entrambi i casi abbiamo un insieme di condizioni sufficienti da un insieme di esempi di concetti positivi e negativi.

Un'ipotesi h è una congiunzione di vincoli sugli attributi di istanza. Ogni costrutto può essere uno specifico valore, senza un valore attribuito o non importante. Una definizione formale di ipotesi può essere:

Data un'istanza X , un insieme di esempi $D = \langle x_i, c(x_i) \rangle$, un concetto target c , delle ipotesi H espresse come congiunzione di costrutti sugli attributi, determina un'ipotesi h in H come $h(x) = c(x)$ per ogni x in X . Il task di apprendimento è volto a determinare un'ipotesi h identica al concetto di target c sull'intero insieme di istanze X .

Ogni ipotesi cerca di approssimare la funzione target su un insieme sufficientemente largo di esempi di training che sarà a sua volta l'approssimazione della funzione target sugli esempi non osservati. L'apprendimento di concetti può essere visto con un task di ricerca su un largo spazio di ipotesi H . L'obiettivo della ricerca è di trovare l'ipotesi che misura l'insieme di training. Lo spazio di ipotesi è implicitamente definito dalla rappresentazione dell'ipotesi.

Date due ipotesi h_k e h_j , h_j è più generale o uguale a h_k quando $h_j \geq h_k$ se e solo se ogni istanza che soddisfa h_k soddisfa anche h_j . Inoltre possiamo definire che h_j è strettamente più generale rispetto a h_k quando $h_j > h_k$ e se e solo se $h_j \geq h_k$ e $h_k \not\geq h_j$.

Un'ipotesi h **copre** un esempio positivo se è correttamente classifica l'esempio come positivo. Quindi si avrà la relazione $h(x) = c(x) = 1$. Un esempio $\langle x, c(x) \rangle$ **soddisfa** le ipotesi h quando $h(x) = 1$ indipendentemente dal fatto che x sia un esempio negativo o positivo del concetto di destinazione. Mentre definiamo che un'ipotesi h è **consistente** con un esempio $\langle x, c(x) \rangle$ quando $h(x) = c(x)$. Infine definiamo un'ipotesi h **consistente con un insieme di dati** D se è consistente con ogni esempio $x \in D$.

Per sfruttare *l'ordinamento dal generale allo specifico*, si potrebbe iniziare con l'ipotesi più specifica in H , per poi generalizzare questa ipotesi ogni volta che essa non riesce a coprire un esempio di addestramento positivo osservato (**bottom-up search**).

Per aiutarci in questo arduo compito descriviamo il procedimento dell'algoritmo **FIND-S**. Inizializziamo h all'ipotesi più specifica in H . Successivamente definiamo due cicli, quello più esterno per ogni istanza positiva x dell'insieme di training esegue il ciclo su ogni attributo della condizione a_i in h . Se il vincolo a_i è soddisfatto per x , allora non si fa nulla, altrimenti si rimpiazza a_i in h con il prossimo vincolo più generale soddisfatto da x . Alla fine dell'algoritmo viene restituita l'ipotesi h .

In questo algoritmo non vi è nessuna revisione in caso di esempio negativo perchè si assume che il target concept c si trova in H e non vi sono errori nei dati di training. L'ipotesi h è l'ipotesi più specifica in H tale per cui $c \geq_g h$, ma c non sarà mai soddisfatta da un esempio negativo.

L'algoritmo Find-S ha anche degli aspetti negativi. Non ci dice se l'apprendimento ha coperto il corretto target concept, se c'è solo un'ipotesi consistente in H con i dati o ci sono più di un'ipotesi. Inoltre non possiamo rilevare quando i dati di training sono inconsistenti, in quanto l'incoerenza negli esempi di addestramento può fuorviare Find-S, poiché ignora gli esempi negativi. Potrebbero esserci diverse ipotesi coerenti massimamente specifiche. L'algoritmo dovrebbe fare marcia indietro sulle sue scelte per esplorare un ramo diverso dell'ordinamento parziale rispetto al ramo che ha selezionato.

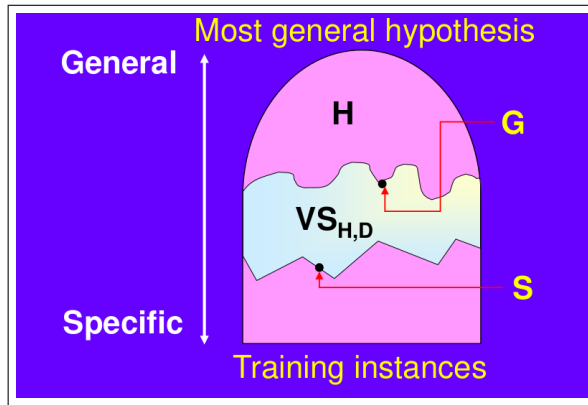
Da qui nasce l'idea del **version space**. Si decide di restituire uno spazio delle versioni invece di una singola ipotesi. Definiamo formalmente questo version space $VS_{H,D}$ come uno spazio contenente le rispettive ipotesi dello spazio H a dei dati di esempio D . In questo modo esso risulta essere un sottoinsieme di H consistente con gli esempi di training in D . Ovvero:

$$VS_{H,D} \equiv \{h \in H \mid \text{Consistent}(h, D)\}$$

Dalla definizione di questo spazio nasce l'algoritmo **List-Then-Eliminate**, nel quale in un primo passo si assume che $VS_{H,D}$ è una lista contenente ogni singola ipotesi in H . Successivamente per ogni esempio $\langle x, c(x) \rangle$ rimuove da $VS_{H,D}$ ogni ipotesi h per la quale $h(x) \neq c(x)$. Infine si restituisce lo spazio rimanente.

Anche questo algoritmo ha i suoi pro e i suoi contro. Certamente garantisce come output tutte le ipotesi consistenti con i dati di training e può trovare l'inconsistenza all'interno dei dati. Si potrebbe anche effettuare una enumerazione di tutte le ipotesi possibili solo per spazi finiti di H o per grandi spazi irrealistici di H .

Il version space può essere rappresentato dai suoi membri più generali e meno generali come in figura.



Il confine generale, o **general boundary** G , rispetto allo spazio delle ipotesi H e ai dati di addestramento D , è l'insieme dei membri massimamente generali di H consistente con D . $G \equiv \{g \in H \mid \text{Consistent}(g, D) \wedge (\neg \exists g' \in H[(g' >_g g) \wedge \text{Consistent}(g', D)])\}$

Il confine specifico, o **specific boundary**, S , rispetto allo spazio delle ipotesi H e all'insieme dei dati D , è l'insieme dei membri generali di H consistenti in D . $S \equiv \{s \in H \mid \text{Consistent}(s, D) \wedge (\neg \exists s' \in H[(s >_g s') \wedge \text{Consistent}(s', D)])\}$

Una volta ottenuti questi due insiemi, possiamo definire un algoritmo per eliminare le ipotesi che non sono generali. Questo algoritmo prende il nome di **Candidate-Elimination**, nel quale in un primo passo si assegna a una variabile G le ipotesi generali in H e in una variabile S si assegnano le ipotesi specifiche in H . Successivamente per ogni esempio di training d , (**controllo1**) si controlla se quest'ultimo è un esempio positivo e nel caso si rimuove da G ogni ipotesi inconsistente con d , per poi avviare una funzione UPDATE-S.

La funzione UPDATE-S ha il compito di controllare ogni ipotesi s in S che non è consistente con d , eliminando s da S , per poi aggiungere a S tutte le generalizzazioni minimali h di s tali che h è consistente con d e qualche membro di G è più generale

di h . Infine rimuove da S ogni ipotesi che è più generale di un'altra ipotesi in S .

Si ritorna al **controllo1** e se questo risultasse negativo, si rimuove da S ogni ipotesi inconsistente con d e si avvia una procedura UPDATE-G per aggiornare l'insieme G . Quest'ultima funzione controlla ogni ipotesi g in G che non è consistente con d e in questo caso rimuove g da G , aggiunge a G tutte le specializzazioni minimali h di g tali che h è consistente con d e qualche membro di S è più specifico di h . Infine rimuove da G ogni ipotesi che è meno generale di altre ipotesi in G .

L'algoritmo *candidate-elimination* convergerà verso il target concept a condizione che non ci sono errori negli esempi di training ed è presente qualche ipotesi in HS che descrive correttamente il target concept. Infatti il target concept è appreso quando il confine di S e G converge a un'ipotesi singola e identica.

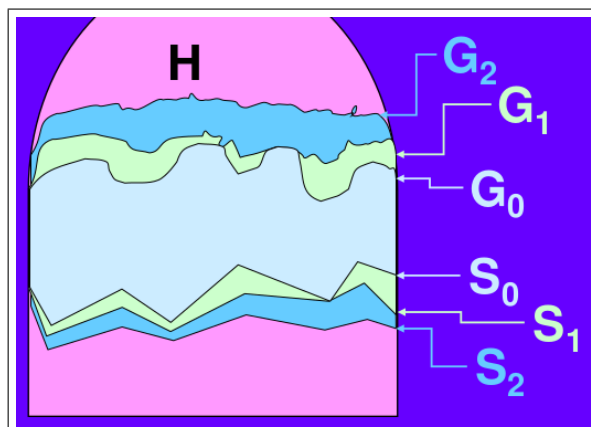
L'algoritmo genera un version space vuoto quando l'insieme dei dati di training contiene errori e il target concept non può essere descritto attraverso ipotesi rappresentative (*hypothesis representation*). Inoltre possiamo dire che l'algoritmo di Candidate-Elimination performa una ricerca bidirezionale, in quanto G e S possono crescere esponenzialmente nel numero di esempi di training.

Quando ci troviamo di fronte a concetti appresi parzialmente, abbiamo una minore confidenza nella classificazione in casi ambigui, ma potremmo attuare un approccio basato sul voto maggiore. In questo approccio assumiamo che tutte le ipotesi in H sono uguali a priori, successivamente potremmo decidere in base:

- al maggior numero di voti, i quali forniscono la classificazione più probabile per la nuova istanza,
- alla percentuale di ipotesi che votano positivamente, interpretando questo come la probabilità che l'istanza sia positiva dato il dato di training.

In questo ci viene in aiuto un algoritmo di *interactive learning* che sceglie la nuova istanza (**query**) e restituisce la corretta classificazione mediante un oracolo esterno. Se l'algoritmo sceglie sempre una query che è soddisfatta solo da metà delle ipotesi in VS , allora il target concept può essere trovato in $\log_2 |VS|$ steps.

Un problema da affrontare in questi tipi di problemi consiste nella gestione del rumore nelle istanze. Per poterlo gestire si potrebbero rendere più blande le condizioni che descrivono la coerenza dei concetti con tutte le istanze di addestramento. Inoltre se si hanno un numero limitato e predeterminato di esempi classificati in modo errato, si potrebbero mantenere diversi insiemi G e S di consistenza variabile.



L'insieme S_i è coerente con tutti tranne gli i esempi di allenamento positivi. L'insieme G_i è coerente con tutti tranne gli

i esempi di allenamento negativi. Quando G_0 incrocia S_0 l'algoritmo può concludere che nessun concept nello spazio delle regole è coerente con tutte le istanze di assestamento. L'algoritmo può recuperare e provare a trovare un concetto coerente con tutti gli esempi di addestramento tranne uno.

Una fondamentale proprietà dell'inferenza deduttiva afferma che: "un algoritmo di apprendimento che non crea nessuna assunzione a priori riguardante l'identità del target concept, non ha basi relazionali per qualsiasi istanza invisibile". Questa proprietà prende il nome di **inductive bias** che può essere definita formalmente come segue.

Definizione formale di inductive bias. Consideriamo:

- un algoritmo L di concept learning
- un insieme X di tutte le istanze
- un target concept c
- degli esempi di training $D_c = \{ \langle x, c(x) \rangle \}$
- denotiamo con $L(x_i, D_c)$ la classificazione assegnata all'istanza x_i da L dopo l'allenamento sui dati D_c

L'inductive bias di L è qualsiasi insieme minimo di asserzioni B tale che per qualsiasi target concept c e corrispondenti esempi di addestramento D_c : $(\forall x_i \in X)[(B \wedge D_c \wedge x_i) \rightarrow L(x_i, D_c)]$. Alcuni sistemi induttivi di modellazione possono essere espressi mediante sistemi deduttivi equivalenti. Il cosiddetto inductive bias è un input esplicito degli algoritmi dei sistemi deduttivi.

Analizziamo l'introduzione dell'inductive bias all'interno dell'algoritmo del candidate-elimination. Il target concept c è contenuto nello spazio delle ipotesi H . Da $c \in H$ segue deduttivamente $c \in VS_{H,D}$. In questo algoritmo L restituisce la classificazione $L(x_i, D_c)$ se e solo se ogni ipotesi in $VS_{H,D}$ produce anche questa classificazione, includendo anche le ipotesi $c \in VS_{H,D}$ (inductive bias). Quindi $c(x_i) = L(x_i, D_c)$.

Il bias induttivo è un mezzo non procedurale per caratterizzare la politica degli algoritmi di apprendimento per generalizzare oltre i dati osservati. Confrontiamo diversi algoritmi di apprendimento in relazione all'inductive bias.

Nell'algoritmo **Rote learning**, si memorizzano gli esempi, viene classificato x se corrisponde a un esempio osservato in precedenza. In questo tipo di algoritmo non c'è nessun inductive bias. Per quanto riguarda l'algoritmo **candidate elimination**, l'inductive bias consiste nel target concept rappresentato dal suo spazio di ipotesi. Mentre per quanto riguarda l'algoritmo **Find-S**, l'inductive bias è il target concept rappresentato dal suo spazio di ipotesi e da tutte le istanze negative a meno che il contrario non sia implicato dalla sua altra conoscenza (una sorta di ragionamento predefinito o non monotono).

Per apprendere concetti disgiunti si può utilizzare l'algoritmo Candidate-Elimination, il quale è un algoritmo con il minimo impegno, che generalizza solo quando è forzato a farlo. Però la disgiunzione fornisce un modo per evitare qualsiasi generalizzazione, per cui l'algoritmo non è mai costretto a farlo. Per apprendere i concetti disgiuntivi, bisogna trovare un metodo per controllare l'introduzione di disgiunzioni, in modo da prevenire disgiunzioni banali. La copertura sequenziale, o **sequential covering**, è un approccio diffuso all'apprendimento di serie di regole.

Vi sono inoltre particolari algoritmi di copertura sequenziale che performano ripetutamente l'algoritmo di candidate-elimination per la ricerca di diverse descrizioni congiuntive che insieme coprono tutte le istanze di training. Ad ogni iterata, si trova un concetto congiuntivo che è coerente con alcuni degli esempi di training positivi e tutti gli esempi di training negativi. I casi positivi che sono stati presi in considerazione vengono rimossi da ulteriori considerazioni e il processo viene ripetuto finché tutti gli esempi positivi non sono stati coperti.

Uno pseudo-codice dell'algoritmo di copertura sequenziale è il seguente:

Sequential-Covering(Target-attribute,Attributes,Examples,Threshold)

Learned-rules $\leftarrow \emptyset$

Rule \leftarrow LEARN-ONE-RULE (Target-attribute,Attributes,Examples)

while PERFORMANCE(Rule, Examples) > Threshold do:

{ Learned-rules \rightarrow Learned-rules $\cup \{Rule\}$ Examples \rightarrow Examples - { examples correttamente coperti da Rule } Rule \rightarrow

LEARN-ONE-RULE(Target-attribute,Attributes, Examples) } Learned-rules \rightarrow sort Learned-rules in accordo con PER-

FORMANCE over Examples return Learned-rules

All'interno dell'algoritmo di copertura sequenziale viene utilizzato un particolare algoritmo chiamato **LEARN-ONE-RULE**. Questo particolare algoritmo accetta un insieme di esempi di training sia positivi che negativi e restituisce una singola regola che copre tutti gli esempi. Questo algoritmo ha il pregio di essere molto accurato, ma non ha un'alta copertura. Questo algoritmo viene implementato attraverso l'algoritmo Candidate-Elimination.

Si inizia inizializzando S, il quale conterrà un esempio positivo di training. Per ogni istanza di training negativa si applica la routine Update-G a G. Scegliere una descrizione g da G come una congiunzione per l'insieme delle soluzioni. Poiché Update-G è stato applicato utilizzando tutti gli esempi negativi, g non copre esempi negativi. Tuttavia, g può coprire molti degli esempi positivi. A questo punto si rimuove da ulteriori considerazioni tutti gli esempi positivi coperti da g. Infine si ripete tutti questi procedimenti fino a quando tutti gli esempi sono coperti.

Un approccio diverso per implementare l'algoritmo LEARN-ONE-RULE è l'**algoritmo di ricerca General-to-Specific**, mediante il quale si organizza lo spazio di ricerca mediante un ordinamento dal generico allo specifico e si esplora lo spazio attraverso una **greedy fashion**. Nello specifico si utilizza una ricerca greedy top-down o general-to-specific.

La ricerca dello spazio per le precondizioni delle regole iniziando considerando la precondizione della regola più generale possibile. Ad ogni passaggio, viene aggiunto il test che migliora maggiormente le prestazioni della regola. L'aggiunta di un nuovo test restringe la copertura della regola. Il test "migliore" è quello che include il maggior numero possibile di esempi positivi ed esclude il maggior numero possibile di esempi negativi. Supponiamo che una regola copra un totale di t istanze, di cui p sono esempi positivi e t-p sono esempi negativi. Poi scegli la regola che massimizza il rapporto di purezza p/t.

Un esempio di algoritmo separate and conquer è il seguente:

```

procedure SIMPLESEPARATEANDCONQUER(Examples)

  Theory =  $\emptyset$ 
  while POSITIVE(Examples)  $\neq \emptyset$ 
    BestRule = {true}
    Rule = BestRule
    while NEGATIVE(Cover)  $\neq \emptyset$  Cover is the set of examples covered by Rule
      for Condition  $\in$  Conditions
        Refinement = Rule  $\cup$  Condition
        if PURITY(Refinement, Examples) > PURITY(BestRule, Examples)
          BestRule = Refinement
        Rule = BestRule
      Theory = Theory  $\cup$  Rule
      Examples = Examples - Cover
  return(Theory)

```

Tutti gli algoritmi di separate and conquer condividono la struttura base. Mentre, molti task di apprendimento richiedono delle modifiche a questo procedimento. Per esempio, la procedura di overfitting se c'è un po' di noisy all'interno dei dati. Pertanto molti algoritmi allentano questo vincolo e utilizzano criteri di arresto o metodi di post-elaborazione per essere in grado di apprendere teorie più semplici che non sono complete e coerenti, ma sono più predittive su dati invisibili.

Altri algoritmi sostituiscono la ricerca dall'alto verso il basso del whileloop interno con una ricerca dal basso verso l'alto, in cui le regole vengono successivamente generalizzate a partire da una regola più specifica (ad esempio costituita da uno degli esempi positivi stessi). Tuttavia, altri algoritmi non utilizzano hillclimbing, ma impiegano algoritmi di ricerca meno miopi come la beam search o la bestfirst search.

La **Beam Search** è una ricerca dal generale allo specifico viene implementata come una ricerca greedy depth-first senza backtracking. In questo algoritmo vi è il pericolo di una scelta subottimale in ogni fase. A volte per evitare questo rischio, l'algoritmo viene performato attraverso una lista di k candidati migliori a ogni step e successivamente viene scelto il migliore.

Un altro problema che si incontra nell'algoritmo di separate and conquer è che il tipo delle condizioni usate nella formulazione delle ipotesi. Questo basilare algoritmo può formulare solo regole in logica proposizionale, ma la programmazione logica induttiva ha sviluppato algoritmi che possono imparare le regole nella logica del primo ordine. Un generico algoritmo di apprendimento delle regole separate and conquer che chiama varie subroutine che possono essere utilizzate per istanziare l'algoritmo generico in algoritmi specifici noti dalla letteratura.

```

procedure SEPARATEANDCONQUER(Examples)

  Theory =  $\emptyset$ 
  while POSITIVE(Examples)  $\neq \emptyset$ 
    Rule = FINDBESTRULE(Examples)
    Covered = COVER(Rule, Examples)
    if RULESTOPPINGCRITERION(Theory, Rule, Examples)
      exit while
    Examples = Examples \ Covered
    Theory = Theory  $\cup$  Rule
  Theory = POSTPROCESS(Theory)
return(Theory)

```

SeparateAndConquer inizia con una teoria vuota. Se ci sono esempi positivi nel set di addestramento chiama la subroutine FindBestRule per apprendere una regola che coprirà un sottoinsieme degli esempi positivi. Tutti gli esempi coperti vengono quindi separati dal set di addestramento, la regola appresa viene aggiunta alla teoria e un'altra regola viene appresa dagli esempi rimanenti. Le regole vengono apprese in questo modo fino a quando non rimangono esempi positivi o fino a quando RuleStoppingCriterion non viene attivato. Spesso la teoria risultante viene sottoposta a PostProcessing.

```

procedure FINDBESTRULE(Examples)

  InitRule = INITIALIZERULE(Examples)
  InitVal = EVALUATERULE(InitRule)
  BestRule =  $\langle$  InitVal, InitRule  $\rangle$ 
  Rules = { BestRule }
  while Rules  $\neq \emptyset$ 
    Candidates = SELECTCANDIDATES(Rules, Examples)
    Rules = Rules \ Candidates
    for Candidate  $\in$  Candidates
      Refinements = REFINERULE(Candidate, Examples)
      for Refinement  $\in$  Refinements
        Evaluation = EVALUATERULE(Refinement, Examples)
        unless STOPPINGCRITERION(Refinement, Evaluation, Examples)
          NewRule =  $\langle$  Evaluation, Refinement  $\rangle$ 
          Rules = INSERTSORT(NewRule, Rules)
          if NewRule > BestRule
            BestRule = NewRule
    Rules = FILTERRULES(Rules, Examples)
  return(BestRule)

```

La procedura FindBestRule cerca nello spazio delle ipotesi una regola che ottimizzi un determinato criterio di qualità definito in EvaluateRule. Il valore di questa funzione euristica di solito è "gli esempi più alti, più positivi e meno negativi sono coperti dalla regola del candidato". FindBestRule gestisce Rules, una lista ordinata di regole candidate, che viene inizializzato dalla

procedura `InitializeRule`. Le nuove regole verranno inserite in punti appropriati (`InsertSort`), in modo che le regole siano sempre ordinate in ordine decrescente dalle valutazioni euristiche delle regole.

Ad ogni ciclo `SelectCandidates` seleziona un sottoinsieme di queste regole candidate, che vengono poi raffinate utilizzando `RefineRule`. Ogni perfezionamento viene valutato e inserito nell'elenco delle regole ordinate a meno che `StoppingCriterion` non lo impedisca. Se la valutazione di `NewRule` è migliore della migliore regola trovata in precedenza, `BestRule` viene impostata su `NewRule`. `FilterRules` seleziona il sottoinsieme dell'elenco di regole ordinate che verrà utilizzato in ulteriori iterazioni. Quando tutte le regole candidate sono state elaborate, verrà restituita la regola migliore.

Ad esempio, supponiamo di voler istanziare l'algoritmo generico nel semplice algoritmo separa e conquista. `SimpleSeparateAndConquer` cerca nello spazio delle ipotesi in modo dall'alto verso il basso. `Initialize Rule` restituirà quindi la regola più generale, ovvero la regola con il corpo true. `RefineRule` specializzerà una determinata regola aggiungendovi una condizione. Le regole saranno valutate dalla percentuale di esempi coperti che sono positivi, cioè `EvaluateRule` implementerà la subroutine `Purity` usata nel semplice algoritmo `separate-and-conquer`.

`FilterRules` lascerà passare solo il miglior perfezionamento per l'iterazione successiva, in modo che `SelectCandidates` abbia sempre una sola scelta. Insieme, queste due procedure implementano la ricerca in salita. A partire dall'elenco ordinato di tutti i perfezionamenti verrà utilizzato solo il primo (e migliore) elemento, questa parte del codice è equivalente alla parte corrispondente in `SimpleSeparateAndConquer`. Entrambi i criteri di arresto saranno sempre falsi e le regole apprese non verranno postelaborate.

Consideriamo adesso gli algoritmi di `simultaneous covering` con gli algoritmi di `sequential covering`. Gli algoritmi di `sequential covering` sono i più lenti. Per imparare un set di n regole, ogniuna delle quali contiene k `attribute-value test` nelle loro precondizioni, gli algoritmi di `sequential covering` performeranno $n \times k$ step primitivi di ricerca, indipendentemente dalle scelte.

Gli algoritmi di copertura simultanea sono generalmente veloci, poichè ogni test con m possibili risultati contribuisce a scegliere le precondizioni per almeno m regole. Gli algoritmi di copertura sequenziale effettuano un gran numero di scelte indipendenti, mentre gli algoritmi di copertura simultanea effettuano un basso numero di scelte dipendenti.

Alcune volte ci si chiede se conviene indurre le regole direttamente o convertire un decision tree in un insieme di regole. La risposta a questa domanda dipende in base a quanti i dati di training sono disponibili. Se i dati sono abbondanti, possono supportare il maggior numero di decisioni indipendenti richieste dagli algoritmi di copertura sequenziale. Se i dati sono scarsi, la condivisione delle decisioni relative ai presupposti di regole diverse può essere più efficace.

Inoltre bisogna prendere in considerazione anche il task di apprendimento. Se la descrizione del concetto è altamente disgiuntiva con molte condizioni indipendenti, gli algoritmi di apprendimento dell'albero decisionale funzionano male quando i dati sono scarsi.

Nell'apprendimento `single-concept`, l'elemento appreso è presentato con positive e negative istanze dello stesso concetto. Il sistema deve cercare le regole che descrivono il concetto di studio. Dato un nuovo caso x , se non soddisfa le precondizioni di ogni regola, deve essere considerata come un'istanza negativa. Altrimenti il sistema potrebbe apprendere le regole sia per le istanze positive che per quelle negative del concetto. Quest'ultimo è un caso semplice di `multiple-concept learning`, nel quale le precondizioni non sono necessariamente mutuamente esclusive.

Quando si classifica un nuovo caso senza il partizionamento dello spazio delle istanze, possiamo ritrovarci in tre casi:

- **nessuna classificazione:** l'istanza non soddisfa nessuna preconditione.
- **singola classificazione:** l'istanza soddisfa le regole di preconditione con la stessa conclusione, sia che essa sia positiva, sia che essa sia negativa.
- **classificazione multipla:** l'istanza soddisfa le regole di preconditione con differenti conclusioni.

Nell'apprendimento di concetti multipli, il sistema di apprendimento viene presentato con esempi di formazione che sono istanze di diversi concetti e deve trovare diverse descrizioni di concetti. Per ogni descrizione di concetto, c'è una regione corrispondente nello spazio dell'istanza. Quando i concetti sono indipendenti, un problema di apprendimento con più concetti può essere riformulato come una sequenza di problemi di apprendimento con un singolo concetto.

L'unione degli insiemi di regole apprese per ogni concetto è l'output dell'algoritmo di apprendimento multiple concept. Se i concetti si escludono a vicenda, la classificazione multipla può essere un risultato indesiderato. Per evitare classificazioni multiple, l'aggiunta di una nuova regola all'insieme delle regole apprese può richiedere la modifica delle precondizioni delle regole esistenti (problema di integrazione della conoscenza). Quando i concetti sottostanti si sovrappongono, la classificazione multipla potrebbe essere una caratteristica desiderabile.

Nei problemi di apprendimento di concetti multipli, gli esempi sono tipicamente descritti da vettori di caratteristiche: $\langle a_1, a_2, \dots, a_n, c \rangle$, dove c è il target attribute. Distinti valori di c rappresentano differenti concetti, i quali sono intesi come mutuamente esclusivi, cioè indipendenti.

Nel caso generale, gli esempi possono essere descritti da vettori di feature: $\langle a_1, a_2, \dots, a_n, c_1, c_2, \dots, c_m \rangle$ dove c_i sono gli attributi target. I concetti non sono necessariamente indipendenti. Le regole apprese possono includere queste dipendenze concettuali. Lo spazio dell'istanza di un problema di apprendimento a concetto singolo è definito anche da alcuni attributi target. Per decidere quali attributi dovrebbero essere considerati, bisogna scoprire le dipendenze tra gli attributi prima di iniziare il processo di apprendimento, per poi definire gli spazi delle istanze di conseguenza.

Eventuali dipendenze tra attributi possono essere rilevate mediante tecniche statistiche (misura lambda, presentata nella sezione sull'analisi delle associazioni). Le dipendenze concettuali possono essere scoperte online, cioè durante il processo di apprendimento, lavorando simultaneamente con diversi spazi di istanza per ogni problema di apprendimento. Ciò equivale a esplorare diversi spazi di ricerca per ogni concetto da apprendere.

3 K-Means

Alcuni algoritmi per eseguire il clustering sui dati e trovare tutti i gruppi di istanze simili all'interno dei dataset sono ad esempio il k-means, EM o il Cobweb. I cluster possono essere visualizzati e comparati con i veri cluster, se presenti.

Questo algoritmo prevede la definizione preventiva del numero di cluster, successivamente viene scelto casualmente un insieme di k punti chiamati **centroidi** che rappresentano i cluster iniziali. Il clustering procede iterativamente nel seguente modo:

- Assegna ogni punto al cluster il cui centroide è più vicino, generalmente si utilizza la distanza euclidea;
- Calcola i nuovi centroidi di ogni cluster come la media dei punti che vi appartengono;
- ripete i passaggi fino a quando non ci sono più cambiamenti significativi nella composizione dei cluster.

Non sempre questo algoritmo identifica un raggruppamento perfetto, infatti il procedimento viene ripetuto scegliendo le distanze iniziali casualmente ed effettuando diversi raggruppamenti da cui scegliere alla fine. Tra i diversi raggruppamenti verrà scelto quello che minimizza le varianze complessivamente, ossia quello che minimizza la total variation within cluster.

Andando ad analizzare i vantaggi e gli svantaggi. Tra i vantaggi possiamo vedere come questo algoritmo sia semplice da implementare con un'efficienza pari a $O(nkd)$ dove:

- d è la dimensionalità
- n è la cardinalità del dataset.
- k è il numero di cluster da identificare
- i è il numero di iterazioni necessarie per convergere

Mentre tra gli svantaggi possiamo notare che è necessario fissare a priori il numero di cluster k , inoltre questo algoritmo è sensibile al rumore nei dati e alla presenza di outlier, il risultato finale dipende da partizionamento scelto inizialmente. Inoltre i cluster assumono forzatamente forme sferiche, finendo sempre l'iterazione trovando i minimi locali.

Se dovessimo dividere l'algoritmo in diversi step, possiamo definirne 6:

- **Step 1:** viene scelto il numero k di cluster in cui selezionare i dati (nel nostro esempio 3)
- **Step 2:** vengono selezionate randomicamente 3 istanze che andranno a rappresentare i 3 cluster iniziali.
- **Step 3:** viene calcolata la distanza tra il primo punto e i 3 cluster iniziali
- **Step 4:** assegna ogni punto al cluster più vicino
- **Step 5:** una volta assegnati tutti i punti, viene calcolata la media di ogni cluster.
- **Step 6:** si riassegnano le istanze basandosi sulle medie in modo iterativo finché all'ultima iterazione non si noti variazioni rispetto all'iterazione precedente.

Non sempre l'algoritmo K-means identifica un raggruppamento perfetto, infatti inizialmente sceglie le istanze casualmente ed effettua diversi raggruppamenti da cui scegliere alla fine.

Ci sono due metodi principali per scegliere il numero k di cluster. Attraverso il metodo **elbow method**, nel quale si minimizza la distanza tra i punti di un singolo cluster e si massimizza la distanza tra i cluster. Per calcolare la distanza tra due cluster si utilizza la formula del WCSS, ovvero la somma dei quadrati delle differenze tra i vari punti dei cluster.

Mentre se si decide di scegliere il metodo **silhouette**, per ogni istanza i calcola la distanza media tra i e ogni altra osservazione j all'interno del cluster in cui i è stata raggruppata. In formula: $a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$. Per ogni istanza i calcola la distanza media tra i e ogni altra osservazione j all'interno degli altri cluster (cluster-wise) diversi da quello in cui risiede i e salva tale media soltanto per il cluster più vicino al cluster di i . Chiama questo valore $b(i)$. In formula: $b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)$.

Infine si definisce lo score Silhouette $s(i)$ per i come segue:

- $1 - \frac{a(i)}{b(i)}$ se $a(i) < b(i)$
- 0 se $a(i) = b(i)$
- $\frac{b(i)}{a(i)} - 1$ se $a(i) > b(i)$

Inoltre il valore $s(i)$ sarà sempre compreso tra -1 e 1.

Per superare il problema di dover specificare i k cluster, possiamo utilizzare l'algoritmo **DBSCAN**, il quale è molto robusto in relazione agli outlier. Inoltre questo algoritmo analizza la densità locale alle osservazioni per individuare raggruppamenti anche in forma diversa da quella sferica. Questo è dovuto all'utilizzo del concetto di *core-point* rispetto ad un numero minimo minPts , input utente, di punti vicini ad una determinata istanza e se tale istanza ha in un intorno almeno minPts istanze, allora è considerata core-point.

Dopo aver identificato tutti i core-point, DBSCAN ne prende uno a caso e lo assegna al primo cluster. Quindi assegna allo stesso cluster tutti i core-point nell'intorno del core-point preso casualmente. DBSCAN fa questo per ogni core-point estendendo il cluster finché non raggiunge i punti che non sono core-point.

Per ogni istanza che non è core-point, essa viene inglobata nel cluster se è vicina ad un core-point del cluster appena formato le istanze che non sono core-point consentono a DBSCAN di fermarsi poiché si è in prossimità di uno spazio delle osservazioni a minore densità rispetto allo spazio denso di core-point.

Stessa cosa viene fatta per altre istanze separate dal primo cluster per identificare il secondo cluster, il terzo e così via. Al termine, le istanze che non sono core-point che non appartengono a nessun cluster vengono considerate outlier.

Vi è un altro tipo di cluster è il **cluster gerarchico**, il cui obiettivo è quello di decomporre gerarchicamente la collezione di individui in una collezione di cluster innestati. Il risultato è rappresentato in forma di dendrogrammi all'interno del quale i nodi rappresentano i possibili cluster., i quali vengono generati bottom-up o top-down.

4 Decision Trees

Gli alberi di decisione sono usati in task di classificazione, ovvero l'assegnazione di oggetti a categorie che si escludono a vicenda. Il modello predittivo assume la forma di un decision tree, spesso quella specifica del binary tree.

Per costruire un decision tree, in letteratura ci sono diversi metodi tra i quali:

- Metodi di partizionamento ricorsivi, derivanti dalla disciplina della statistica
- Induzione di alberi decisionali, derivanti dal Machine Learning
- Costruzione di alberi di classificazione, derivanti dal pattern recognition.

Un importante decision tree è quello definito come **C4.5**, che accetta una collezione di casi di training, ognuno dei quali ha una tupla di valori da un insieme di attributi prefissati o **variabili indipendenti** e una classe attributo o **variabile dipendente**. Un attributo A_i è descritto come continuo o discreto in base se i suoi valori sono numerici o nominali.

C4.5 utilizza test di due tipi, ognuno dei quali implica un singolo attributo A_i . Se quest'ultimo è discreto, i possibili test sono:

- z risultati, uno per ogni valore di A_i
- $2 \leq w \leq z$ risultati, dove $G = \{G_1, G_2, \dots, G_w\}$ è una porzione dei valori di A_i .

Se A_i è continuo, viene suddiviso in corrispondenza di un valore di soglia derivato. Il test viene diviso nel momento $A_i \leq \theta$ (o a $A_i > \theta$) con risultati vero o falso. In questo caso per determinare quale attributo classifica meglio il dato, si utilizza l'entropia. Prendendo in considerazione:

- S come l'insieme degli esempi di training analizzati
- C_j una classe label in C_1, C_2, \dots, C_k .
- $RF(C_i, S)$, $i=1, \dots, k$ le relative frequenze di classi in S che appartengono alla classe C_i

possiamo considerare l'entropia E su S calcolata come segue: $E(S) = - \sum_{i=1 \dots k} RF(C_i, S) \log(RF(C_i, S))$. L'entropia misura il grado di incertezza contenuto in S .

Da notare come questo valore assume il valore massimo quando gli eventi sono equiprobabili. Questa è una buona cosa poiché un insieme di eventi ugualmente probabili è quanto mai incerto. Mentre assume un valore minimo quando solo uno degli eventi ha una probabilità diversa da zero.

Considerando due classi di problemi (C_1 =positivo, C_2 =negativo) dove:

- p è la porzione degli esempi positivi in S ,
- n è la porzione negativa di esempi in S ,

l'entropia di S è computata come: $E(S) = -p \log_2 p - n \log_2 n$.

Un altro valore che può essere utilizzato per la selezione dei test è il cosiddetto **information gain**, che rappresenta la riduzione prevista dell'entropia a causa della distribuzione delle istanze su S in accordo con il test t . Considerando S_1, \dots, S_z il partizionamento di S attraverso il test t sull'attributo A_i , l'information gain può essere ottenuto come: $G(S, t) = E(S) - \sum_i \frac{|S_i|}{|S|} E(S_i)$. In questo modo si sceglie il criterio che massimizza $G(S, t)$.

Un problema legato a questo parametro consiste nel favorire test con numerosi risultati. In questo caso l'information gain è massimizzato da un test su un attributo chiave che ogni S_i contiene un solo caso. Una prima soluzione a questo problema consiste nel limitare i decision trees a test binari. Però sono possibili più test sullo stesso attributo. Ciò può portare a alberi decisionali che sono ancora più incomprensibili per gli esperti umani di quanto non avvenga normalmente. Inoltre questa soluzione è computazionalmente espansiva.

Una seconda soluzione tiene conto delle potenziali informazioni dalla partizione stessa, senza considerare la classe, quindi in formula: $P(S, t) = - \sum_i \frac{|S_i|}{|S|} \log(\frac{|S_i|}{|S|})$. La maggior parte possibile delle informazioni fornite determinando il valore di un attributo dovrebbe essere utile ai fini della classificazione o, equivalentemente, il meno possibile dovrebbe essere "sprecato". Perciò scegliamo il test t che massimizza la formula $\frac{G(S, t)}{P(S, t)}$, definendo così il **criterio del gain ratio**.

Per trattare i continuous attributes all'interno dell'algoritmo C4.5 si può definire un nuovo attributo valorizzato con valori discreti per testare l'attributo continuo partizionando il valore continuo in un insieme di intervalli discreti e utilizzando $A_c = \text{true}$ se $A < c$. Il valore migliore di c è selezionato usando l'information gain.

Però testare tutti i possibili valori di c è impossibile e non necessario, poichè non ci sarebbe nessuna differenza tra due valori c e c' quando non ci sono esempi di training che assumono un valore intermedio per l'attributo A . Da questo possiamo estrapolare la seguente procedura:

- Ordinare gli esempi secondo il loro valore per A
- Per ogni coppia ordinata X_i, X_{i+1} nella lista ordinata, se le classi di X_i e X_{i+1} sono differenti, utilizzo il punto intermedio tra i loro valori come un candidato soglia.
- Selezione il miglior candidato in accordo con l'information gain

4.1 Affrontare i valori mancanti (missing values)

Se incontriamo valori mancanti di A possiamo intraprendere diverse soluzioni:

- Off-line: affronto il problema durante lo step di data transformation del processo KDD,
- On-line: modifica il processo di induzione del decision tree per gestire mancanti.

Nell'approccio on-line vi sono diversi problemi. Un primo problema è la valutazione dei test. La selezione di un test su cui partizionare il training set viene effettuata sulla base di criteri euristici quali gain e gain ratio. In questo caso non possiamo valutare la desiderabilità relativa di due test che utilizzano attributi con valori diversi.

Un secondo problema relativo all'approccio online consiste nel partizionare il training set. Una volta che il test è stato selezionato, i casi di addestramento con valori sconosciuti dell'attributo rilevante non possono essere associati a un particolare risultato del test, e quindi non possono essere assegnati a un particolare sottoinsieme T_i .

Un ulteriore problema consiste nel classificare un caso sconosciuto. In questo caso ci si chiede come si deve comportare il decision tree nel caso in cui il sistema ha un valore sconosciuto per l'attributo testato?

Analizziamo nel dettaglio i tre problemi. Nel primo il test non fornisce informazioni circa la classe di appartenenza del caso in cui il valore dell'attributo di test è sconosciuto. Il gain ratio può essere alterato considerando i casi con valori sconosciuti come un gruppo aggiuntivo.

Per il problema del partizionamento del dataset, possiamo utilizzare un approccio probabilistico. Supponiamo di avere il test t con z esiti v_1, v_2, \dots, v_z , che partiziona S in z sottoinsiemi S_1, S_2, \dots, S_z , l'idea è quella di associare ogni caso in ogni sottoinsieme S_i con un peso rappresentante la probabilità che il caso appartenga al dataset:

- se il caso ha un esito conosciuto con peso 1
- altrimenti il peso è la probabilità del risultato v_i in quel punto.

Per quanto riguarda il terzo problema, se si incontra un nodo decisionale in cui il valore dell'attributo rilevante è sconosciuto, il sistema esplora tutti i possibili risultati e combina aritmeticamente le classificazioni risultanti. Una classificazione è una distribuzione di classi piuttosto che una singola classe. La classe con la probabilità più alta viene assegnata come classe prevista.

4.2 Svantaggi

La tecnica di induzione del decision tree è molto efficiente in termini di tempo di elaborazione e fornisce un metodo molto intuitivo per analizzare i risultati. Questa intuitività, soprattutto se confrontata con l'output meno accessibile delle reti neurali, spiega perché gli alberi decisionali sono ancora attraenti. Nonostante tutto hanno degli svantaggi.

Gli attuali metodi di induzione dell'albero decisionale non sono ottimali. Durante la formazione di un albero decisionale, una volta che l'algoritmo prende una decisione sulla base su cui dividere il nodo, tale decisione non viene mai rivista. I metodi per la costruzione di un albero decisionale ottimo hanno una complessità esponenziale.

Gli alberi decisionali soffrono di un problema di errori che si propagano in un albero, un problema molto serio con l'aumentare del numero di classi. Poiché i regression tree funzionano in base a una serie di decisioni locali, se una di queste è sbagliata ogni decisione da quel punto in poi potrebbe essere sbagliata, inoltre non avremmo mai un percorso corretto.

Gli alberi decisionali sono limitati ai problemi che possono essere risolti dividendo lo spazio della soluzione in rettangoli successivamente più piccoli.

In generale, gli alberi decisionali soffrono di frammentazione. Quando l'albero ha molti strati di nodi, la quantità di dati che passa attraverso le foglie inferiori è così piccola che è difficile prendere una decisione precisa sulla classe associata a ciascuna foglia. Per ridurre al minimo la frammentazione, l'analista può potare o tagliare alcune delle foglie e dei nodi inferiori per far collassare efficacemente parte dell'albero. Il risultato è un modello migliorato in cui vengono rivelati i modelli reali piuttosto che il rumore (evitamento dell'overfitting) ed è più semplice da capire.

Generalmente si sceglie ipotesi brevi rispetto a ipotesi lunghe, per svariati motivi. Definiamo H come uno spazio di ipotesi, considerando un errore nell'ipotesi $h \in H$ questo si ripercuote sui:

- Training data, e per calcolare l'errore si può definire la funzione $error_{training}(h)$
- sull'intera distribuzione D di dati, il cui errore si può calcolare definendo la funzione $error_D(h)$.

L'ipotesi $h \in H$ overfitta i dati di training se c'è un'ipotesi alternativa $h' \in H$ tale che $error_{training}(h) < error_{training}(h')$ e se $error_D(h) > error_D(h')$. Si dice che un'ipotesi, h , supera l'insieme di addestramento, D , quando esiste un'ipotesi, h' , che supera h sulla distribuzione totale delle istanze in cui D è un sottoinsieme.

Questo può accadere se c'è rumore all'interno dei dati che crea un albero grande per l'ipotesi h e l'ipotesi h' non fitta. Inoltre avere un numero limitato di campioni, rende le coincidenze sempre più possibili, anche se gli attributi non correlati all'attributo di destinazione possono partizionare bene i dati di addestramento.

Per evitare l'overfitting possiamo **fermare la crescita dell'albero** quando lo split dei dati non è statisticamente significativo, in quanto non ci sono abbastanza dati per rendere le decisioni affidabili. Inoltre per evitare una crescita esponenziale, si potrebbe togliere dall'albero tutti quei nodi per i quali non si hanno sufficienti evidenze.

4.3 Training set vs. Validation set

A volte i dati sono partizionati in training set e testing set. I primi vengono utilizzati per apprendere, mentre i secondi per validare. Tipicamente si utilizza una proporzione del 70-30.

4.4 Metodi di selezione dei cicli

4.4.1 Reduced error pruning (REP)

In questo caso si utilizza un set di potatura per stimare l'accuratezza dei sottoalberi e l'accuratezza sui singoli nodi. Considerando un sub-tree con un nodo v come vertice, si definisce il **Gain della potatura a v** = # classificazioni errate in T - # classificazioni errate in v . Si pota T al nodo v solo se il gain è positivo. In questo caso bisogna rispettare una *restrizione Bottom-up*: T può essere eliminato solo se non contiene un sottoalbero con errore inferiore a T .

4.4.2 Minimal Error Pruning (MEP)

Questo algoritmo prende in considerazione per potare i rami se l'errore statistico è minore o uguale all'errore di backed-up, e l'errore del sotto albero di .

$$E(T) = \min(e(v), \sum_i p_i E(T_i))$$

Per stimare l'errore $e(v)$ si usa Laplace o la stima m di probabilità.

La **stima di Laplace** al nodo v , con N esempi, con n_c il numero di esempi della classe di maggioranza e con k il numero di classi, si ha:

$$p_c = \frac{n_c + 1}{N + k}$$

Questa formula assume che tutte le classi hanno la stessa probabilità e il grado di potatura dipende dal numero di classi.

Per risolvere questi problemi si utilizza la stima m , che prende in considerazione p_{ca} ovvero la probabilità della classe C e m , un parametro non negativo.

$$p_C = \frac{p_{Ca} \times m + n_C}{N + m}$$

Questo metodo dà importanza agli account con probabilità priori, non risente del numero delle classi, si può variare il valore di m ottenendo una serie di alberi potati differenti. La scelta di m dipende dalla confidenza con i dati

4.4.3 Pessimistic Error Pruning (PEP)

Questo algoritmo non richiede un set di potatura per la stima dell'errore, la quale avviene sui nuovi dati direttamente dall'insieme di crescita. Si effettua una stima sulla frequenza relativa con una correzione, considerando la probabilità d'errore.

$$q = \frac{N - n_c + 0.5}{N}$$

dove N rappresenta il numero di esempi e n_c il numero di esempi nella classe maggiore. Inoltre si può calcolare l'errore di un nodo se potato:

$$q(v) = \frac{N_v - n_{C,v} + 0.5}{N_v}$$

Mentre l'errore di un sotto albero può essere calcolato come:

$$q(T) = \frac{\sum_{l \in leaf(T)} N_l - N_{C,l} + 0.5}{\sum_{l \in leaf(T)} N_l}$$

4.4.4 Error-Based Pruning (EBP)

In questo caso si decide di potare un sotto albero T se $q(v) \leq q(T) + SE(q(T))$, dove SE è l'errore standard di una statistica X sopra un numero di esempi n :

$$SE(X) = \sqrt{\frac{X(1-X)}{n}}$$

5 Naive Bayes Classifier - Classificatore Bayesiano

Il **Bayesian Framework** è composto da un insieme di dati osservati D ai quali siamo interessati nel determinare le migliori ipotesi da qualche spazio H . Per un task di classificazione, le ipotesi è un classificatore. Il *teorema di Bayes* è un metodo per calcolare la probabilità di un ipotesi basata sulle probabilità a priori e sui dati osservati.

Definiamo:

- con h le ipotesi
- con $P(h)$ la probabilità a priori di h , cioè, la probabilità iniziale che valga l'ipotesi h , prima di aver osservato i dati di addestramento. Riflette alcune conoscenze di base che abbiamo sulla possibilità che h sia un'ipotesi corretta.
- con $P(D)$ la probabilità a priori di D , cioè, la probabilità che i dati di addestramento D vengano osservati non sapendo quale ipotesi sia valida.
- con $P(D|h)$ denotata probabilità di osservare i dati D dato un mondo in cui vale l'ipotesi h .
- con $P(h|D)$ probabilità a posteriori di h , ovvero la probabilità che h sia valida dopo aver visto i dati di addestramento D .

Nella formula:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h|D)$ cresce con $P(h)$ e con $P(D|h)$
- $P(h|D)$ diminuisce all'aumentare di $P(D)$, perché quanto più è probabile che D venga osservato indipendentemente da h , tanto meno prove D fornisce a sostegno di h .

L'ipotesi $h \in H$ con la massima probabilità è l'ipotesi con **massimo a posteriori (MAP)**, ovvero:

$$h_{MAP} = \operatorname{argmax}_{h \in H} P(h|D) = \operatorname{argmax}_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

Il termine $P(D)$ può essere eliminato dalla formula perchè è una costante indipendente di h . Assumendo che tutte le ipotesi sono altrettanto probabili, le equazioni possono essere ulteriormente semplificate. L'ipotesi che massimizza $P(D|h)$ è la più probabile. $P(D|h)$ è chiamata la probabilità dei dati D data h . Questa ipotesi è chiamata ipotesi di probabilità massima (ML), ovvero:

$$h_{ML} = \operatorname{argmax}_{h \in H} P(D|h)$$

A differenza di una ipotesi ML, una ipotesi MAP incorpora la conoscenza a priori.

Per applicare il teorema di Bayes ai problemi di classificazione, definiamo:

- uno spazio di istanze X

- $c : X \rightarrow \{0, 1\}$, una funzione di classificazione che deve essere appresa
- uno spazio delle ipotesi H
- una funzione booleana definita su X tale che $h \in H: h : X \rightarrow \{0, 1\}$

L'obiettivo è di trovare h tale che $h(x)=c(x)$ per tutte le istanze $x \in X$. L'informazione disponibile è una sequenza di esempi di training $\langle x_1, d_1 \rangle \cdots \langle x_m, d_m \rangle$, dove x_i è un'istanza di X e d_i è il valore target di x_i , ovvero $d_i = c(x_i)$.

Se si volesse implementare un **algoritmo di apprendimento MAP per forza bruta**, basterebbe calcolare per ogni $h \in H$ la probabilità a posteriori:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

L'output dell'ipotesi h_{MAP} con la più alta probabilità a posteriori:

$$h_{MAP} = \operatorname{argmax}_{h \in H} P(h|D)$$

In questo algoritmo si assume che l'insieme dei dati di training D è senza rumore, ovvero $d_i = c(x_i)$. Il concetto target c è contenuto nello spazio dell'ipotesi H . Inoltre si presume che non abbiamo nessuna ragione a priori di credere che un'ipotesi è più probabile di un'altra.

Per stimare $P(h)$, si assume che $P(h_i)$ è uguale per ogni $h_i \in H$, inoltre si assume che la somma di $P(h_i)$ è uguale a uno. Perciò $P(h) = \frac{1}{|H|}$ per tutte $h \in H$.

Per stimare $P(D|H)$ si assume:

$$P(D|h) = \begin{cases} 1 & \text{if } d_i = h(x_i) \forall x_i \in D \\ 0 & \text{altrimenti} \end{cases}$$

La probabilità dei dati D , data l'ipotesi h , è 1 se D è consistente con h , 0 altrimenti.

Per stimare $P(D)$ basta utilizzare il teorema della probabilità e la mutua esclusione delle ipotesi.

$$P(D) = \sum_{h_i \in H} P(D|h_i)P(h_i) = \sum_{h_i \in VS_{H,D}} 1 \times \frac{1}{|H|} + \sum_{h_i \notin VS_{H,D}} 0 \times \frac{1}{|H|} = \frac{|VS_{H,D}|}{|H|}$$

dove $VS_{H,D}$ è un sotto insieme di ipotesi da H che sono consistenti in D .

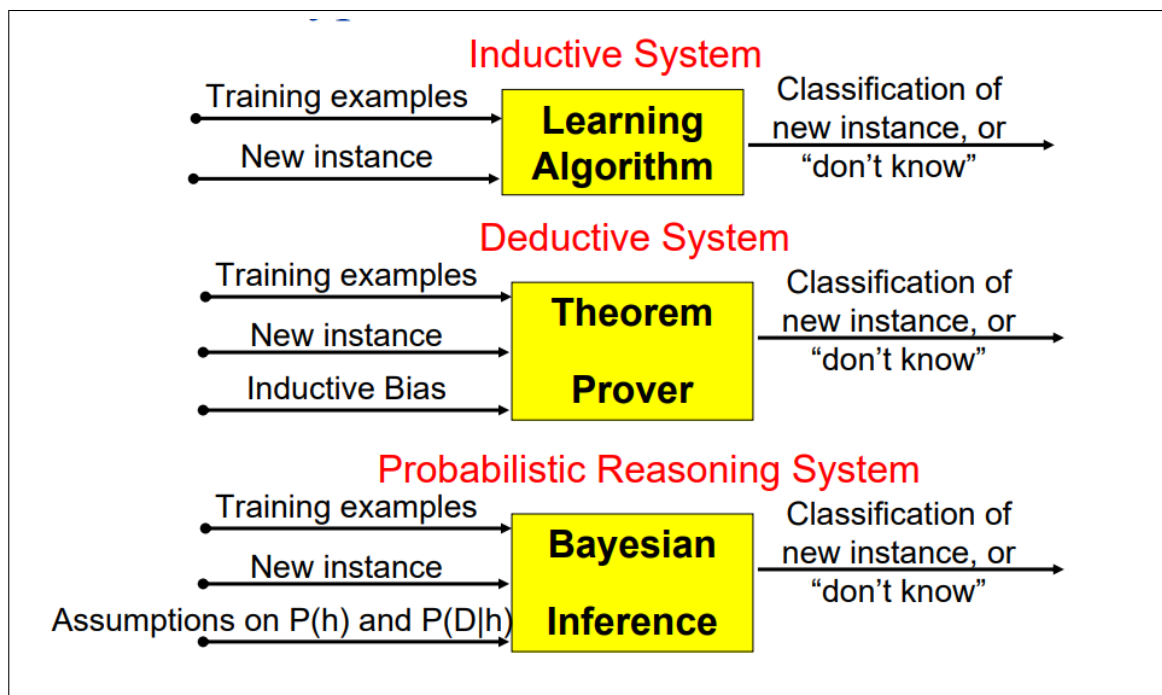
Perciò quando h è inconsistente con l'insieme dei dati D , $P(h|D) = 0 \times \frac{P(h)}{P(D)} = 0$, altrimenti:

$$P(h|D) = \frac{1 \times \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}}$$

Sotto le scelte di $P(h)$ e $P(D|h)$, ogni ipotesi consistente ha una probabilità a posteriori di $\frac{1}{|VS_{H,D}|}$ e ogni ipotesi inconsistente ha una probabilità a posteriori di 0. Perciò possiamo assumere che *ogni ipotesi consistente è un'ipotesi MAP*.

Una classe di learners consistenti è una classe di algoritmi di apprendimento che produce un'ipotesi che commette zero errori sugli esempi di training. Ogni learner consistente produce un'ipotesi MAP, se assumiamo una distribuzione di probabilità uniforme a priori su H e assumiamo che questa sia deterministica, su un insieme di dati di training senza rumore.

Il framework bayesiano consente un modo per caratterizzare il comportamento degli algoritmi di apprendimento (ad esempio, FIND-S), anche quando l'algoritmo di apprendimento non manipola esplicitamente le probabilità. Questa caratterizzazione degli algoritmi di apprendimento è simile nello spirito alla caratterizzazione del pregiudizio induttivo del learner. Invece di modellare il metodo di inferenza induttiva con un sistema deduttivo equivalente, lo modelliamo con un sistema di ragionamento probabilistico equivalente basato sul teorema di Bayes.



Riconsideriamo la definizione di h_{MAP} : $h_{MAP} = \operatorname{argmax}_{h \in H} P(D|h)P(h)$, potremmo riscriverla cercando di massimizzarla/minimizzarla in termini di \log_2 :

$$h_{MAP} = \operatorname{argmax}_{h \in H} (\log_2(P(D|h)) + \log_2(P(h)))$$

$$h_{MAP} = \operatorname{argmin}_{h \in H} (-\log_2(P(D|h)) - \log_2(P(h)))$$

Nella teoria dell'informazione vi è un problema per decodificare un algoritmo per trasmettere messaggi estratti a caso. La probabilità del messaggio i è p_i . Qual è il codice che minimizza il numero previsto di bit che dobbiamo trasmettere per codificare un messaggio estratto a caso? Ci viene in aiuto il teorema di Shannon, il cui codice ottimale assegna $-\log_2(p_i)$ bits per codificare il messaggio i . La lunghezza del messaggio atteso è: $\sum_i -p_i \log_2 p_i$

La lunghezza della descrizione del messaggio i rispetto al codice C , $L_C(i)$, è il numero di bits richiesto per codificare il messaggio i usando il codice C .

In formule:

$$h_{MAP} = \operatorname{argmin}_{h \in H} (-\log_2 P(D|h) - \log_2 P(h))$$

Che possiamo interpretare alla luce del teorema di Shannon come:

- $-\log_2 P(h)$ è la descrizione della lunghezza di h sotto la condizione ottimale per l'ipotesi di spazio H . $\rightarrow L_{C_H}(h) = -\log_2 P(h)$ dove C_H è il codice ottimale per lo spazio dell'ipotesi H .
- $-\log_2 P(D|h)$ è la lunghezza della descrizione di D data h sotto la sua codifica ottimale $\rightarrow L_{C_{D|h}} = -\log_2 P(D|h)$ dove $C_{D|h}$ è il codice ottimale per descrivere i dati D assumendo che sia chi li ha visti sia il ricevente conoscano l'ipotesi h .

Perciò l'ipotesi h_{MAP} è l'ipotesi h che minimizza la somma: $h_{MAP} = \operatorname{argmin}_{h \in H} (L_{C_H}(h) + L_{C_{D|h}}(D|h))$

Il **principio della lunghezza minima della descrizione** raccomanda di scegliere l'ipotesi che minimizza la somma delle due lunghezze relative alle descrizioni, una che rappresenta l'ipotesi e l'altra che rappresenta i dati data l'ipotesi: $h_{MDL} = \operatorname{argmin}_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$.

$L_{C_2}(D|h)$ è zero quando l'ipotesi non sbaglia a classificare i dati di training. Il principio MDL è un bias induttivo che consiglia, analogamente al rasoio di Occam, di scegliere l'ipotesi più semplice (più breve) che spieghi i dati di addestramento. Fornisce un modo per bilanciare la complessità dell'ipotesi con il numero di errori commessi dall'ipotesi. Potrebbe selezionare un'ipotesi più breve che commette alcuni errori rispetto a un'ipotesi più lunga che classifica perfettamente i dati di addestramento. Fornisce un modo per gestire il sovradattamento.

Se confrontassimo

$$h_{MDL} = \operatorname{argmin}_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$$

con

$$h_{MAP} = \operatorname{argmin}_{h \in H} L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

risulta chiaro che le ipotesi MAP e il principio MDL sono strettamente collegati. Quando $C_1 = C_H$ e $C_2 = C_{D|h}$ allora $h_{MDL} = h_{MAP}$.

Invece per garantire che $C_1 = C_H$ e $C_2 = C_{D|h}$ dobbiamo conoscere tutte le probabilità a priori $P(h)$ come anche tutte le $P(D|h)$. In pratica a volte è più semplice specificare una rappresentazione che cattura la conoscenza sulle probabilità relative delle ipotesi piuttosto che specificare completamente la probabilità di ciascuna ipotesi. Tipicamente, le applicazioni dell'MDL a problemi pratici di apprendimento spesso includono argomenti che forniscono una qualche forma di giustificazione per le codifiche scelte per C_1 e C_2 .

In generale, il classificatore più probabile per una nuova istanza è ottenuta combinando le predizioni delle ipotesi, pesate attraverso le loro probabilità a posteriori. Se l'eventuale classificazione del nuovo esempio può assumere qualsiasi valore v_j dall'insieme V , allora la probabilità $P(v_j|D)$ che la classificazione corretta per la nuova istanza è v_j è solo:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Qualsiasi sistema che classifica nuove istanze in questo modo è chiamato classificatore ottimale di Bayes o studente ottimale di Bayes. Nessun altro metodo di classificazione che utilizza lo stesso spazio di ipotesi e la stessa conoscenza preliminare può in media superare questo metodo.

Una curiosità del **classificatore bayesiano ottimale** è che le predizioni possono corrispondere a un'ipotesi non in H . Questo perché lo spazio di ricerca non è H , ma H' , il quale include ipotesi che performano confronti tra combinazioni lineari delle predizioni da ipotesi multiple in H .

Il classificatore bayesiano ottimale è piuttosto costoso da applicare, in quanto calcola la probabilità per ogni ipotesi in H e dopo combina le predizioni per ogni ipotesi da classificare per ogni nuova istanza. Potremmo approssimare la formula $P(V_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$ a $P(V_j|D) = P(v_j|h_{MAP})P(h_{MAP}|D)$. Da $P(h_{MAP}|D)$ non dipende una nuova istanza, la decisione è attualmente basata sul massimo valore di $P(v_j|h_{MAP})$.

Un'approssimazione alternativa è data da una scelta randomica delle ipotesi. Consideriamo l'**algoritmo Gibbs**, il quale sceglie un'ipotesi $h \in H$ casualmente in accordo con la distribuzione a posteriori della probabilità su H , e usa h per predire la classificazione sulla istanza successiva a x .

Supponiamo che le ipotesi utilizzate per la classificazione siano estratte casualmente da H secondo la distribuzione di probabilità a priori su H . Allora l'errore di classificazione errata atteso per l'algoritmo di Gibbs è al massimo il doppio dell'errore atteso del classificatore ottimo di Bayes.

$$E[\text{error}_{Gibbs}] \leq 2E[\text{error}_{BayesOptimal}]$$

In generale, le prestazioni dell'algoritmo di Bayes forniscono uno standard naturale rispetto al quale è possibile confrontare altri algoritmi.

Il **naive Bayes classifier** è un'istanza del metodo di apprendimento derivato secondo il framework bayesiano. Si applica all'apprendimento di tasks dove ogni istanza è descritta da una tupla di attributi valori $\langle a_1, \dots, a_n \rangle$ e dove l'attributo target può assumere qualsiasi valore da un insieme finito V .

Una nuova istanza viene assegnata al valore target più probabile v_{MAP} , dati i valori degli attributi $\langle a_1, \dots, a_n \rangle$:

$$v_{MAP} = \operatorname{argmax}_{v_j \in V} P(v_j|a_1, a_2, \dots, a_n)$$

Secondo il teorema di Bayes:

$$v_{MAP} = \operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n|v_j)P(v_j)}{P(a_1, a_2, \dots, a_n)} = \operatorname{argmax}_{v_j \in V} P(a_1, a_2, \dots, a_n|v_j)P(v_j)$$

In questo caso possiamo stimare $P(v_j)$ attraverso la frequenza relativa del valore target v_j all'interno del training set. Questo approccio non è praticabile per $P(a_1, \dots, a_n|v_j)$, in quanto vi sono troppe probabilità da stimare, ma proporzionalmente pochi dati di training.

Come ipotesi di semplificazione, possiamo attribuire valori che sono condizionatamente indipendenti dato il valore target:

$$P(a_1, a_2, \dots, a_n|v_j) = \prod_{i=1}^n P(a_i|v_j)$$

Quindi si potrebbe utilizzare il **Naive (o Simple) Bayes Classifier**, all'interno del quale abbiamo:

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_{i=1}^n P(a_i|v_j)$$

In questo caso il numero di parametri distinti da stimare è uguale al numero di valori di attributo distinti moltiplicato per il numero di valori target distinti.

Lo step di apprendimento consiste nella stima dei vari $P(v_j)$ e di $P(a_i|v_j)$ basati sulla frequenza dei dati di training. L'insieme di queste stime corrisponde all'ipotesi appresa. Successivamente l'ipotesi viene usata per classificare ogni nuova istanza applicando la regola:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i=1}^n P(a_i|v_j)$$

Ogni volta che il presupposto del naive Bayes dell'indipendenza condizionale è soddisfatto, questa classificazione di Bayes v_{NB} è identica alla classificazione MAP v_{MAP} .

Il classificatore naive Bayes approssima il classificatore ottimale Bayes:

$$v_{OBC} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j|D) = \underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

attraverso:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j|h_{NB})P(h_{NB}|D) = \underset{v_j \in V}{\operatorname{argmax}} P(v_j|h_{NB}) = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i=1}^n P(a_i|v_j)$$

Questo avviene solo quando l'assunzione di indipendenza condizionale del classificatore Naive Bayes viene soddisfatta: $h_{NB} = h_{MAP}$.

5.1 Underflow Prevention

Moltiplicare numerose probabilità, le quali sono comprese tra 0 e 1, può produrre un numero float molto grande. Poiché log è una funzione crescente strettamente monotona, vale la seguente equazione:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \hat{P}(v_j) \prod_{i=1}^n \hat{P}(A_k = a_{ki}|v_j) = \underset{v_j \in V}{\operatorname{argmax}} (\log_2 2\hat{P}(v_j) + \sum_{i=1}^n \log_2 \hat{P}(A_k = a_{ki}|v_j))$$

È meglio eseguire tutti i calcoli sommando i logaritmi delle probabilità piuttosto che moltiplicando le probabilità. La classe con il punteggio finale di probabilità logaritmica non normalizzata più alto è ancora la più probabile.

Una differenza interessante tra il metodo di apprendimento naive Bayes e gli altri metodi di apprendimento considerati finora è che non esiste una ricerca esplicita nello spazio delle ipotesi possibili. Lo spazio delle ipotesi possibili H è lo spazio delle possibili dipendenze tra attributi.

L'ipotesi dei classificatori naive Bayes è chiaramente falsa in molti casi. Per questo motivo, l'apprendimento del classificatore naive Bayes è stato considerato poco affidabile ed è stato inizialmente utilizzato come base per il confronto con algoritmi più sofisticati. Nonostante questo scetticismo, negli anni passati è stato dimostrato che in molti domini l'accuratezza della previsione del classificatore naive Bayes si confronta bene con quella di altri algoritmi di apprendimento più complessi tra cui l'apprendimento dell'albero decisionale, l'apprendimento delle regole e gli algoritmi di apprendimento basati su istanze.

Per calcolare un classificatore Naive Bayes è necessario stimare due insiemi di parametri:

- $\theta_{ijk} = \hat{P}(A_k = a_{ki} | v_j)$
- $\pi_j = \hat{P}(v_j)$

Se ciascun attributo può assumere I possibili valori discreti, e $|V| = J$, allora il numero di parametri da stimare è:

- nIJ parametri di tipo θ_{ijk} , dove sono $n(I-1)J$ sono indipendenti, poichè devono soddisfare l'equazione $1 = \sum_i \theta_{ijk}$
- J parametri di tipo π_j , dove solamente $J-1$ sono indipendenti, poichè devono soddisfare l'equazione $1 = \sum_j \pi_j$

Possiamo stimare questi parametri utilizzando la stima di massima verosimiglianza (ML), la quale è basata sulle frequenze relative dei differenti eventi all'interno dei dati. Tuttavia, per piccoli set di dati di training:

- La stima ML è basata su una sottostima della probabilità
- Quando una stima di probabilità è zero, questo termine di probabilità dominerà il classificatore di Bayes a causa della moltiplicazione dei termini nella formulazione del naive Bayes

Una stima alternativa è la **m-estimate di probabilità**: $\frac{N_c + mp}{N + m}$ dove:

- $\frac{N_c}{N}$ è la frequenza relativa corrispondente alla probabilità che noi desideriamo stimare
- p è la stima a priori della probabilità che vogliamo stimare
- m è una costante che determina quanto pesare p in relazione ai dati osservati

Tipicamente, $p = (\frac{1}{k})$ se un attributo ha k possibili valori. Il parametro m è chiamato dimensione equivalente del campione, poichè può essere interpretato come un aumento delle n osservazioni effettive con ulteriori m campioni virtuali distribuiti secondo p . La stima m della probabilità è "più fluida" della stima ML. Quando $m=k$ e $p = (\frac{1}{k})$, possiamo ricavare la stima di Laplace:

$$\frac{N_c + 1}{N + k}$$

La **distribuzione multidimensionale** è una distribuzione discreta che dà la probabilità di scegliere una data raccolta di N elementi da un insieme di k elementi con ripetizioni e le probabilità di ciascuna scelta data da p_1, \dots, p_k . Questi parametri sono i parametri della distribuzione multinomiale.

La funzione di massa di probabilità della distribuzione multinomiale è:

$$f(x_1, \dots, x_k; n, p_1, \dots, p_k) = Pr(X_1 = x_1 \dots X_k = x_k) = \begin{cases} \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} & \text{quando } \sum_{i=1}^k x_i = n \\ 0 & \text{altrimenti} \end{cases}$$

per interi non negativi x_1, \dots, x_k .

La **distribuzione Dirichlet** è il coniugato a priori dei parametri della distribuzione multinomiale. E' definita come:

$$D(p|\alpha) = \frac{1}{B(\alpha)} \prod_{h=1}^k p_h^{\alpha_h-1}$$

dove:

- $p = \langle p_1, \dots, p_k \rangle$ rappresentano le variabili casuali, con $p_i \geq 0, \sum p_i = 1$
- $\alpha = \langle \alpha_1, \dots, \alpha_k \rangle$ rappresentano i parametri della distribuzione (tanti parametri quante le probabilità in p), con $\alpha_i > 0$

$B(\alpha)$ è una costante normalizzata come: $\int_p D(p|\alpha) = 1$. I parametri $\alpha_i - 1$ possono essere interpretati come "peso dell'osservazione a priori" per gli eventi disciplinati da p_i , ovvero l' i -esimo elemento sarà osservato $\alpha_i - 1$ volte. Dopo aver disegnato N elementi, se si ottengono N_h elementi di tipo h , la posteriore bayesiana è una distribuzione di Dirichlet con parametri: $(m_1 + \alpha_1, \dots, m_k + \alpha_k)$. La stima MAP è:

$$p_h = \frac{N_h + \alpha_h - 1}{N + \sum_{h=1}^k \alpha_h - k}$$

Nel caso di attributi continui A_k , la distribuzione di probabilità si presume essere Gaussiana per ogni possibile valore target v_j . Pertanto, per addestrare una NBC dobbiamo stimare la media e la deviazione standard di ciascuna di queste gaussiane:

- $\mu_{kj} = E[A_k | v_j]$
- $\sigma_{kj}^2 = E[(A_k - \mu_{kj})^2 | v_j]$

Dobbiamo stimare tutti questi parametri in maniera indipendente. Si deve anche stimare i valori target a priori: $\pi_j = \hat{P}(v_j)$. Per questi parametri possiamo utilizzare la stima ML o la stima MAP.

Lo stimatore ML per μ_{kj} è:

$$\hat{\mu}_{kj} = \frac{1}{\sum_i \delta(v_j^i)} \sum_i A_k^i \delta(v_j^i)$$

dove l'apice i si riferisce all' i -esimo esempio di addestramento, e dove

$$\delta(v_j^i) = \begin{cases} 1 & \text{se il valore target dell}'i\text{-esimo esempio di training è } v_j \\ 0 & \text{altrimenti} \end{cases}$$

Lo stimatore ML per σ_{kj}^2 è:

$$\sigma_{kj}^2 = \frac{1}{\sum_i \delta(v_j^i)} (A_k^i - \hat{\mu}_{kj})^2 \delta(v_j^i)$$

Questo stimatore ML è distorto, quindi a volte viene utilizzato lo stimatore imparziale della varianza minima.

$$\sigma_{kj}^2 = \frac{1}{(\sum_i \delta(v_j^i)) - 1} (A_k^i - \hat{\mu}_{kj})^2 \delta(v_j^i)$$

Alcune proprietà del classificatore naive bayesiano sono:

- **Incrementalità:** con ogni esempio di training, la priorità e la probabilità possono essere aggiornate dinamicamente: flessibile e resistente agli errori.
- **Combina conoscenza precedente e dati osservati:** probabilità a priori di un'ipotesi moltiplicata per la probabilità dell'ipotesi dati i dati di addestramento.
- **Ipotesi probabilistiche:** produce non solo una classificazione, ma una distribuzione di probabilità su tutte le classi.
- **Meta-classificazione:** i risultati di diversi classificatori possono essere combinati, ad esempio, moltiplicando le probabilità che tutti i classificatori prevedono per una determinata classe

Un classificatore naive bayes può essere utilizzato anche per la text categorization, all'interno del quale può essere effettuata una classificazione manuale (molto accurata se fatta da esperti e utilizzata per problemi di piccole dimensioni), oppure attraverso una classificazione automatica attraverso un documento. Quest'ultima viene utilizzata attraverso sistemi basati su regole codificate manualmente per sistemi commerciali che hanno dei complessi linguaggi di query, inoltre la precisione può essere elevata se una regola è stata attentamente perfezionata nel tempo da un esperto in materia anche esperto del linguaggio di query.

Per classificare si utilizzano metodi di apprendimento attraverso funzioni di assegnazione di documenti etichettati, i quali richiedono dati di training classificati manualmente, ma i dati possono essere accumulati (e perfezionati) da dilettanti.

Gli attributi sono posizioni di testo, i valori sono parole, quindi usando un Naive Bayes Classifiers per classificare testi si ha:

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_i P(x_i | c_j)$$

Assumendo che la classificazione è indipendente dalle posizioni delle parole, si potrebbero utilizzare gli stessi parametri per le stesse posizioni e ottenere quindi un modello bag of words.

6 I modelli di regressione

Con un modello di regressione si esprime la relazione funzionale esistente tra variabili. In particolare, tale relazione funzionale lega una variabile aleatoria quantitativa (variabile dipendente o risposta) a una (o più) variabili aleatorie quantitative e/o discrete (variabili indipendenti o regressori). Si dispone di n osservazioni per ognuna delle quali si conoscono i valori assunti dai regressori e dalla risposta (training set supervisionato).

Nei **modelli di regressione con errore additivo** si ha una variabile Y dipendente, un insieme di variabili indipendenti o regressori X_1, \dots, X_p e una funzione $Y = f(X_1, \dots, X_p) + \epsilon$, composta da una componente deterministica osservabile $f(X_1, \dots, X_p)$ e ϵ la componente stocastica non osservabile.

Nei modelli di regressione è fondamentale l'indipendenza dei regressori: $\epsilon_i \perp X_1, \dots, X_p$, inoltre indicando con:

- $E(\epsilon_i) = 0$ il valore atteso nullo (effetto non sistematico)
- $Var(\epsilon_i) = \sigma^2$ omoschedasticità
- $Cov(\epsilon_i, \epsilon_j) = 0$ incorrelazione tra gli errori

si ottiene un'ipotesi del secondo ordine: $E(Y_i | X_{1i}, \dots, X_{pi}) = f(X_{1i}, \dots, X_{pi})$ e $Var(Y_i | X_{1i}, \dots, X_{pi}) = \sigma^2$

La funzione f esprime la forma del legame funzionale tra Y e X_1, \dots, X_p . All'interno dei modelli parametrici, tale forma è nota, ma dipende da un numero finito di parametri. Mentre nei modelli non parametrici tale forma non è nota.

Si definiscono modelli di regressione semplice quando si prende in considerazione una sola variabile indipendente, ad esempio: $Y = f(X; \beta) + \epsilon$. In questo caso il regressore può essere quantitativo o qualitativo. Quando si prende in considerazione una sola variabile indipendente e f assume la forma di una retta con $\beta = (\beta_0, \beta_1)$, i modelli di regressione prendono il nome di modelli di regressione lineare semplice e assumono una forma: $Y = \beta_0 + \beta_1 X + \epsilon$.

All'interno dei modelli di regressione lineare semplice si assume che:

- $E(\epsilon_i) = 0$ il valore atteso nullo (effetto non sistematico)
- $Var(\epsilon_i) = \sigma^2$ omoschedasticità
- $Cov(\epsilon_i, \epsilon_j) = 0$ incorrelazione tra gli errori
- La variabile X è nota senza errore (non stocastica) ed è osservata per almeno due valori distinti.

quindi abbiamo: $E(Y_i | X_{1i}, \dots, X_{pi}) = \beta_0 + \beta_1 X_{1i}$.

Il parametro β_1 esprime la variazione nel valore atteso condizionato di Y indotta da una variazione unitaria nel regressore. β_1 dice quanto varia in media Y per variazione unitaria della X . Il modello è detto lineare in X perchè l'entità di questa variazione non dipende dal particolare valore assunto da X .

Per stimare i parametri si può usare il metodo dei minimi quadrati: $D(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$. Questa è una funzione quadratica che ammette un solo punto di minimo, ottenuto derivando rispetto β_0 e a β_1 e ponendo tali derivate pari a zero.

Si ottiene quindi il seguente sistema di equazioni normali:

$$\begin{cases} \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

La cui soluzione non è altro che una coppia di punti:

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

La retta dei minimi quadrati ha svariate proprietà:

- E' unica poichè è unico il minimo della funzione $D(\beta_0, \beta_1)$
- Passa per il punto di coordinate medie (\bar{x}, \bar{y}) , ottenute tramite $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- la media dei valori di y previsti coincide con la media dei valori di y osservati
- i residui $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ hanno media nulla e non sono correlati coi valori del regressore

Sotto le ipotesi classiche del modello di regressione lineare semplice, gli stimatori dei minimi quadrati $\hat{\beta}_0$ e $\hat{\beta}_1$ sono lineari e non distorti e a varianza uniformemente minima nella classe degli stimatori lineari e non distorti. Per il teorema di GGauss-Markov si ha quindi:

- $E(\hat{\beta}_0) = \beta_0$
- $E(\hat{\beta}_1) = \beta_1$
- $Var(\hat{\beta}_0) = \frac{\sigma^2}{n} \left(1 + \frac{n\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)$
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$

Si scompone la devianza :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

nelle due componenti:

- $\sum_i (y_i - \hat{y}_i)^2$ che corrisponde alla devianza di dispersione
- $\sum_i (\hat{y}_i - \bar{y})^2$ che corrisponde alla devianza di regressione

da qui si ottiene quindi:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Se alle ipotesi classiche si affianca anche l'ipotesi di normalità distributiva del termine di errore, è possibile costruire intervalli di confidenza e sottoporre a verifica ipotesi sul valore dei parametri. L'ipotesi interessante è $H_0 : \beta_1 = 0$, la quale esprime l'assenza di una relazione lineare tra Y e X. Questa ipotesi può essere valutata ricorrendo a due test equivalenti.

Il **T-test**:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}}$$

In questo test, sotto H_0 la statistica si distribuisce come una t di Student con $n - 2$ gradi di libertà.

Il **Test F**:

$$F = (n - 2) \frac{R^2}{1 - R^2}$$

Attraverso tale test l'ipotesi H_0 si distribuisce come una F di Snedecor-Fisher con 1 e $n-2$ gradi di libertà.

Ulteriori indicazioni sull'inadeguatezza del modello lineare semplice sono fornite dalle cosiddette **diagnostiche grafiche**. Queste sono molteplici e si riconducono tutte più o meno esplicitamente all'esame del comportamento dei residui $\hat{\epsilon}_i = y_i - \hat{y}_i$ che ci servono come surrogato degli errori ϵ_i che non sono osservabili.

In questi casi si potrebbe utilizzare il **diagramma di Anscombe**, ovvero il diagramma dei residui rispetto ai valori interpolati, che idealmente dovrebbe presentare una disposizione del tutto casuale dei punti, se il modello selezionato è in effetti valido.

Un altro diagramma è il cosiddetto **diagramma quantile-quantile**, il quale verifica l'assunto di normalità per la distribuzione degli ϵ_i . Sull'asse delle ordinate sono riportati i valori di $\hat{\epsilon}_i$, opportunamente standardizzati e ordinati in senso crescente, mentre sull'asse delle ascisse troviamo i corrispondenti valori attesi sotto l'ipotesi di normalità, eventualmente approssimati per semplicità di calcolo. Se l'ipotesi di normalità è valida, ci aspettiamo che i punti osservati si dispongano lungo la bisettrice del primo e terzo quadrante.

Quando ci troviamo ad utilizzare modelli di regressione lineare semplici con un regressore qualitativo, possiamo ricodificare la variabile qualitativa (non numerica) usando una o più variabili indicatrici (dummies).

Una **variabile indicatrice** I_E è una variabile che assume i valori 1 o 0, a seconda che una certa condizione E sia o non sia verificata. In particolare:

- $I_E = 0$ se la condizione E non è verificata

- $I_E = 1$ se la condizione E è verificata

Prendiamo in esame una variabile X qualitativa dicotomica $X \in A, B$. Si consideri la condizione $X = A$. La variabile indicatrice I_A sarà uguale a 1 se $X=A$ altrimenti sarà uguale a 0. Con una variabile indicatrice è possibile codificare numericamente una variabile dicotomica. In questo caso il modello di regressione lineare diventa semplice:

$$Y = \beta_0 + \beta_1 I_A + \epsilon$$

Da questa equazione si ottiene:

- $Y = \beta_0 + \epsilon$ se $X \neq A$ allora $E(Y_i | X_i = B) = \beta_0$
- $Y = \beta_0 + \beta_1 + \epsilon$ se $X = A$ allora $E(Y_i | X_i = A) = \beta_0 + \beta_1$

Il parametro β_1 esprime la differenza tra valore atteso condizionato di Y quando $X = A$ e il valore atteso di Y quando $X = B$. Quindi β_1 ci dice quanto varia in media Y al variare di X .

I parametri possono essere stimati attraverso il metodo dei minimi quadrati:

$$D(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 I_{Ai})^2$$

derivando rispetto a β_0 e a β_1 si ottiene il sistema di equazioni normali

$$\begin{cases} \sum_i (y_i - \beta_0 - \beta_1 I_{Ai}) = 0 \\ \sum_i (y_i - \beta_0 - \beta_1 I_{Ai}) I_{Ai} = 0 \end{cases}$$

Possiamo riscrivere la seconda equazione come:

$$\sum_{x_i=A} (y_i - \beta_0 - \beta_1 * 1) * 1 + \sum_{x_i=B} (y_i - \beta_0 - \beta_1 * 0) * 0 = \sum_{x_i=A} (y_i - \beta_0 - \beta_1)$$

Indicando con n_A il numero di unità statistiche per le quali $x_i = A$, possiamo ricavare:

$$\beta_1 = \frac{\sum_{x_i=A} y_i}{n_A} - \beta_0$$

Sostituendo nella prima equazione normale l'espressione per β_1 appena trovata e procedendo in modo analogo si ottiene:

$$\sum_f [y_i - \beta_0 - (\bar{y}_A - \beta_0) I_{Ai}] = \sum_{x_i=A} (y_i - \bar{y}_A) + \sum_{x_i=B} (y_i - \beta_0)$$

da qui si ricava $\hat{\beta}_0 = \frac{\sum_{x_i=B} y_i}{n_B} = \bar{y}_B$ e quindi $\hat{\beta}_1 = \bar{y}_A - \bar{y}_B$. Il valore previsto dal modello è dunque pari a $\hat{y}_i = \bar{y}_B + (\bar{y}_A - \bar{y}_B) I_{Ai}$, ovvero:

- $X_i = A \rightarrow \hat{y}_i = \bar{y}_A$
- $X_i = B \rightarrow \hat{y}_i = \bar{y}_B$

Anche in questo caso valgono il teorema di Gauss-Markov. Sotto le ipotesi classiche del modello di regressione lineare semplice, gli stimatori ottenuti sono lineari, non distorti e a varianza uniformemente minima nella classe degli stimatori lineari e non distorti. Si ha quindi:

- $E(\hat{\beta}_0) = \beta_0$
- $E(\hat{\beta}_1) = \beta_1$
- $Var(\hat{\beta}_0) = \frac{\sigma^2}{n_\beta}$
- $Var(\hat{\beta}_1) = \sigma^2 \left[\frac{1}{n_A} + \frac{1}{n_B} \right]$

Inoltre $E(Y_i - \hat{Y}_i) = 0$ quindi $Var(Y_i - \hat{Y}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{n \left(I_{Ai} - \frac{n_A}{n} \right)^2}{n_A n_B} \right)$

Possiamo scomporre la varianza: $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2$ in:

- Devianza entro $\sum_i (y_i - \hat{y}_i)^2 = \sum_{x_i=A} (y_i - \bar{y}_A)^2 + \sum_{x_i=B} (y_i - \bar{y}_B)^2$
- Devianza tra $\sum_i (\hat{y}_i - \bar{y})^2 = n_A (\bar{y}_A - \bar{y})^2 + n_B (\bar{y}_B - \bar{y})^2$

Da qui otteniamo quindi $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \eta^2$

Per valutare la bontà di adattamento possiamo utilizzare il coefficiente di correlazione di Fisher per due gruppi $\eta^2 = \frac{n_A (\bar{y}_A - \bar{y})^2 + n_B (\bar{y}_B - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$. In questo caso il numeratore calcola devianza fra due gruppi, mentre il denominatore calcola devianza totale.

Qualora iniziassimo a considerare gli aspetti inferenziali, accanto alle ipotesi classiche si affianca l'ipotesi di normalità distributiva del termine di errore. Anche in questo caso si possono costruire intervalli di confidenza e sottoporre a verifica ipotesi sul valore dei parametri. In particolare si considera l'ipotesi $H_0 : \beta_1 = 0$ che esprime l'assenza di dipendenza in media tra Y e X.

Questa ipotesi può essere testata attraverso due test equivalenti. Il primo è il T-test per il confronto fra due medie:

$$t = \frac{\hat{\beta}_1}{\sqrt{S^2 \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}} = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{S^2 \left[\frac{1}{n_A} + \frac{1}{n_B} \right]}}$$

Sotto l'ipotesi H_0 tale statistica si distribuisce come una t di Student con $n-2$ gradi di libertà.

Un altro tipo di test è il test F di Anova:

$$F = (n-2) \frac{R^2}{1-R^2} = \frac{n_A(\bar{y}_A - \bar{y})^2 + n_B(\bar{y}_B - \bar{y})^2}{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}}$$

Sotto H_0 il test si distribuisce come una F di Snedecor-Fisher con 1 e $n-2$ gradi di libertà.

Il modello di regressione lineare semplice può essere esteso in due modi attraverso:

- una regressione polinomiale
- una regressione lineare multipla

Nella prima si prende in considerazione una sola variabile indipendente e f assume la forma di un polinomio di grado p con $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ e quindi $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_p X^p + \epsilon$. Questo modello è lineare nei parametri ma non lineare nella variabile. Nel secondo si prendono in considerazione p variabili indipendenti X_1, X_2, \dots, X_p e f assume la forma di una equazione lineare con $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$, avremo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$. Questo modello è lineare nei parametri, additivo e lineare nei regressori.

Per quanto riguarda le ipotesi classiche per il modello di regressione polinomiale sono:

- $E(\varepsilon_i) = 0$ valore atteso nullo (effetto non sistematico dell'errore).
- $Var(\varepsilon_i) = \sigma^2$ omoschedasticità
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ incorrelazione tra gli errori
- La variabile X è nota senza errore (non stocastica) ed è osservata per almeno $p+1$ valori distinti

quindi ricaviamo che: $E(Y_i|X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_p X_i^p$. I parametri β_i esprimono la variazione nel valore atteso condizionato di Y (espressa nella stessa unità di misura della X) indotta da una variazione unitaria nel regressore. Poiché questo modello non è più lineare in X , tale variazione dipenderà dal particolare valore di X in corrispondenza del quale viene calcolata.

Per quanto riguarda il modello di regressione multipla le ipotesi classiche estese sono:

- $E(\varepsilon_i) = 0$ valore atteso nullo (effetto non sistematico dell'errore).
- $Var(\varepsilon_i) = \sigma^2$ omoschedasticità
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ incorrelazione tra gli errori
- La variabile X è nota senza errore (non stocastica) ed è osservata per almeno $p+1$ valori distinti

Il j -esimo parametro β_j esprime la variazione nel valore atteso condizionato di Y indotta da una variazione unitaria nel j -esimo regressore, fermo restando il valore degli altri regressori. β_j esprime quanto varia in media Y per variazione unitaria della X_j a parità degli altri regressori. Il modello è detto lineare nei regressori perchè l'entità di questa variazione non dipende dal particolare valore assunto dal generico regressore X_j . Il modello è detto adattivo nei regressori perchè l'entità della variazione indotta dal generico regressore X_j non dipende dai particolari valori assunti dagli altri regressori.

Un caso particolare del modello di regressione multipla consiste nel regressore qualitativo poliatomico, nel quale le variabili indicatrici possono essere decodificate numericamente. Generalmente una variabile qualitativa a k modalità può essere codificata numericamente con $k-1$ variabili indicatrici.

6.1 Notazione matriciale

Tutti i modelli di regressione lineare semplice, i modelli di regressione polinomiale e i modelli di regressione lineare multipla sono lineari nei parametri. Pertanto, ricorrendo alla notazione matriciale è possibile esprimere i tre tipi di modelli usando una formulazione comune: $Y = X^T \beta + \varepsilon$, dove:

- X è il vettore della colonna dei regressori
- β è il vettore della colonna dei coefficienti

Quindi avremo:

- Per la regressione lineare semplice: $X^T = [1, X]$ e $\beta^T = [\beta_0, \beta_1]$
- Per la regressione polinomiale: $X^T = [1, X, X^2, \dots, X^p]$ e $\beta^T = [\beta_0, \beta_1, \dots, \beta_p]$
- Per la regressione lineare multipla: $X^T = [1, X_1, X_2, \dots, X_p]$ e $\beta^T = [\beta_0, \beta_1, \dots, \beta_p]$

6.2 Regressione con funzioni a gradini

Il modo più semplice per approssimare una qualunque funzione $y=f(x)$, con $x \in \mathfrak{R}$ è quello di usare una funzione approssimante a gradini, cioè una funzione costante a tratti su intervalli. Per farlo è necessario:

- una **partizione** di \mathfrak{R} composta da J elementi $R_1, \dots, R_j, \dots, R_J$
- Un insieme di J valori reali $c_1, \dots, c_j, \dots, c_J$ ognuno dei quali associato a un elemento della partizione $R_1, \dots, R_j, \dots, R_J$.

Data la partizione $R_1, \dots, R_j, \dots, R_J$ e i valori $c_1, \dots, c_j, \dots, c_J$, ricorrendo a opportune funzioni indicatrici, una generica funzione a gradini può essere rappresentata analiticamente come:

$$\hat{f}(x) = c_1 I(x \in R_1) + \dots + c_j I(x \in R_j) + \dots + c_J I(x \in R_J) = \sum_{j=1}^J c_j I(x \in R_j)$$

dove

$$I(x \in R_j) = \begin{cases} 1 & \text{se } x \in R_j \\ 0 & \text{altrimenti} \end{cases}$$

Il numero e le dimensioni degli intervalli $R_1, \dots, R_j, \dots, R_J$ determinano la complessità della funzione costante a tratti e la qualità con cui essa approssima $f(x)$. La qualità dell'approssimazione aumenta se, a parità di numero di intervalli, si posizionano intervalli più piccoli là dove $f(x)$ presenta valori della derivata prima più elevati in valore assoluto. Inoltre la qualità dell'approssimazione aumenta se aumenta il numero di intervalli.

6.2.1 Funzioni costanti a tratti per la regressione semplice

In questo caso si ha una variabile dipendente continua Y e una variabile indipendente a valori in \mathfrak{R} X . La funzione che si vuole approssimare è rappresentata dal valore atteso condizionato: $f^*(x) = E[Y|X = x] \forall x \in \mathfrak{R}$. Dato il training set $T = y_i, x_i, i = 1, \dots, n$, f^* viene approssimata mediante una funzione \hat{f}_T scelta nella classe delle funzioni costanti a tratti, sulla base della minimizzazione dell'errore empirico di previsione: $EEP_T(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$.

Per minimizzare l'errore empirico di previsione mediante una funzione costante a tratti è necessario individuare:

- la partizione ottimale di \mathfrak{R} in termini di numero di elementi e loro dimensioni
- i valori ottimali per i parametri c_j associati agli elementi di tale partizione

Quando la partizione di \mathfrak{R} è nota, $EEP_T(f)$ deve essere minimizzato solo rispetto ai J parametri c_j . Si può dimostrare che se $f(x) = \sum_{j=1}^J c_j I(x \in R_j)$, allora indipendentemente dalla forma degli elementi della partizione in cui ciascuna sommatoria si riferisce alle sole unità che presentano valore di x appartenente al corrispondente intervallo della partizione.

Quando invece la partizione di \mathfrak{R} non è nota $EEP_T(f)$ deve essere minimizzato congiuntamente sia rispetto alla partizione che ai parametri. In questo caso non esiste una soluzione analitica e una ricerca esaustiva tra tutte le possibili partizioni risulterebbe computazionalmente troppo complessa.

6.2.2 Alberi di regressione

$EEP_T(f)$ può essere minimizzato ricorrendo a una procedura ricorsiva per approssimazioni successive:

- Passo 1: si considera la partizione degenera composta da un solo elemento, che coincide con l'intero insieme \mathfrak{R} .
- Passo 2: si individua una nuova partizione composta da J elementi, suddividendo in due parti uno degli elementi della partizione ottenuta al passo precedente in modo che: $EEP_T(f^{(J-1)}) > EEP_T(f^{(J)})$

Le funzioni costanti a tratti ottenute ricorrendo a questo tipo di suddivisioni possono essere rappresentate mediante un **albero binario**, i cui **nodi** sono affermazioni di tipo logico relative ai valori di X , ognuno associata a una suddivisione. Partendo dalla **radice**, si esamina ciascuna di queste affermazioni: se è vera si segue il ramo di sinistra, altrimenti si segue quello di destra. Ogni nodo **foglia** è associato a uno degli elementi R_j e fornisce il corrispondente valore c_j .

6.2.3 Individuazione della sequenza di suddivisioni ottimali (crescita dell'albero)

Al passo J si considera la formula: $\hat{f}_T^{(J)}(x) = \sum_{j=1}^J \hat{c}_j I(x \in R_j)$ con $EEP_T(\hat{f}_T^{(J)}) = \frac{1}{n} \sum_{j=1}^J \left\{ \sum_{x_i \in R_j} (y_i - \hat{c}_j)^2 \right\}$.

Al passo J+1 in prima fase si considera l'elemento R_l sul quale si cicla suddividendolo secondo una regola logica in due parti R'_l e R''_l , ottenendo quindi una partizione a J+1 elementi. Vengono stimati i J+1 parametri della funzione associata a questa nuova partizione:

$$\begin{cases} \hat{c}'_l = \frac{\sum_{x_i \in R'_l} y_i}{n'_l} = \bar{y}'_l \\ \hat{c}''_l = \frac{\sum_{x_i \in R''_l} y_i}{n''_l} = \bar{y}''_l \end{cases}$$

Successivamente si calcola l'errore empirico di previsione associato a questa nuova funzione pari a $\frac{1}{n} \left\{ \sum_{j \neq l} D_j + D_l^* \right\}$ con $D_l^* = \sum_{x_j \in R'_l} (y_j - \hat{c}'_l)^2 + \sum_{x_j \in R''_l} (y_j - \hat{c}''_l)^2$. D_l^* rappresenta la devianza entro i due gruppi di unità definiti dai due insiemi R'_l e R''_l . Infine si valuta il decremento nell'errore empirico associato alla suddivisione dell'elemento R_l : $\Delta(s, R_l) = D_l - D_l^* = \sum_{x_i \in R_l} (y_i - \bar{y}_l)^2 - \left[\sum_{x_i \in R'_l} (y_i - \bar{y}'_l)^2 + \sum_{x_i \in R''_l} (y_i - \bar{y}''_l)^2 \right]$. $\Delta(s, R_l) = D_l - D_l^*$ rappresenta la devianza tra i due gruppi di unità definiti dai due insiemi di R'_l e R''_l .

Queste operazioni per tutte le possibili suddivisioni di R_l , per calcolare tutti i corrispondenti decrementi nell'errore empirico di previsione. Tra tutte le possibili suddivisioni di tutti gli elementi della partizione ottenuta al passo J si sceglie quella a cui è associato il maggior decremento nell'errore empirico di previsione.

La strategia di crescita individuata è subottimale, in quanto il risultato ottenuto al passo J è condizionato da quanto avvenuto nei passi precedenti. Non è necessario che a ogni passo ci si limiti a suddividere un solo elemento della partizione. Ogni passo della procedura ricorsiva induce una riduzione nell'errore empirico di previsione (**riduzione della distorsione**). Ad ogni passo della procedura ricorsiva, si riduce il numero di unità all'interno degli elementi R_j della partizione (**aumento della varianza**). La procedura può proseguire fino a quando $J = n$, ovvero fino a quando l'errore empirico di previsione si annulla (**overfitting**).

Per scegliere il numero ottimale di elementi della partizione è necessario stabilire dei criteri di arresto. Il primo potrebbe essere quello di individuare un numero minimo di osservazioni di ciascuna foglia, non considerando tutte le suddivisioni che generano almeno un elemento che contiene un numero di osservazioni inferiore a tale soglia. Un secondo è quello di individuare una soglia minima per il decremento nell'errore empirico di previsione, non considerando tutte le suddivisioni alle quali è associata una riduzione nell'errore empirico di previsione inferiore a tale soglia.

Questi due tipi di criteri di arresto possono essere utili poichè limitano la complessità computazionale dell'algoritmo, ma non sono in grado di individuare la complessità ottimale, in quanto sono soggettivi e non tengono conto dell'andamento dell'errore di generalizzazione.

6.2.4 Estensione al caso di due o più regressori continui

Il generico elemento R_j della partizione viene suddiviso in due parti ricorrendo a regole logiche del tipo " $X_h < s_h$?" dove s_h è uno dei valori di X_h osservati sulle unità del training set che appartengono a R_j . Il numero massimo di suddivisioni possibili per il generico R_j risulta pari a $p(n_j - 1)$. Nello spazio p -dimensionale, suddivisioni di questo tipo generano partizioni di composte da ipercubi con lati paralleli agli assi coordinati.

Ci sono diversi metodi per scegliere il miglior test per ogni nodo interno. Nel metodo CART la scelta del miglior test per ogni nodo interno si basa sulla minimizzazione la devianza tra due gruppi. Nella procedura di partizione ricorsiva implementata dal metodo, la natura dello spazio dei regressori influisce solo sulla definizione delle possibili suddivisioni del generico elemento della partizione $\{R_1, \dots, R_j, \dots, R_J\}$. Una volta stabiliti i criteri con cui suddividere gli insiemi, le procedure di crescita e di potatura dell'albero viste possono essere applicate anche nel caso di uno o più regressori categorici.

6.2.5 X ordinale

Quando abbiamo un X ordinale $\{m_1 < \dots < m_k < \dots < m_K\}$, il generico elemento R_j della partizione viene suddiviso in due parti ricorrendo a regole logiche del tipo $X < m$?, dove m è una delle modalità di X osservate sulle unità del training set che appartengono a R_j . Il numero massimo di suddivisioni possibili per il generico R_j risulta pari a $K-1$.

6.2.6 X nominale

Quando abbiamo un X nominale $\{m_1, \dots, m_k, \dots, m_K\}$, il generico elemento R_j della partizione viene suddiviso in due parti ricorrendo a regole logiche del tipo $X \in M$?, dove M è un sottoinsieme delle modalità di X osservate sulle unità del training set che appartengono a R_j , quindi $M \in \{m_1, \dots, m_k, \dots, m_K\}$. Il numero massimo di suddivisioni possibili per il generico R_j risulta pari a $2^{K-1} - 1$.

Sia gli **alberi di regressione**, sia gli **alberi di modelli** sono sempre basati sul partizionamento ricorsivo dello spazio definito dalle variabili indipendenti, ma alle foglie sono associate delle funzioni di regressione lineare. La differenza consiste nel fatto che gli alberi di regressione approssimano mediante una funzione costante a tratti, mentre gli alberi di modelli approssimano per mezzo di una funzione lineare a tratti.