Recupero e Riconoscimento dell'Informazione per Bioinformatica

Domande in stile esame

Chiara Solito - Bioinformatica A.A. 2021/22

Domande stile tema d'esame - Teoria

Esercizio 1

Si descriva la differenza tra approcci generativi con stima parametrica e approcci generativi con stima non parametrica, citando un esempio per ognuna delle due categorie.

Svolgimento

Gli approcci generativi sono un tipo di approccio alla classificazione tramite la teoria della deisione di Bayes: l'obiettivo è formulare un classificatore, assegnado un oggetto alla classe la cui $g(x)$ (funzione discriminante) è masssima. Nell'approccio generativo si mira a modellare tutte le classi del problema, stimando le funzioni $p(\omega_1|x)$ e $p(\omega_2|x)$, utili a calcolare sia $g_1(x)$ che $g_2(x)$, per assegnare l'oggetto alla classe che risponde più fortemente alla funzione.
Questo tipo di approccio quindi mira al calcolo della probabilità a posteriori, utile per l'apprendimento da esempi nel training di un classificatore. A sua volta può essere diviso in approcci generativi con stima parametrica e non parametrica. La stima parametrica consiste nello stimare i parametri, al fine di stimare quale "forma" descrive al meglio la mia classe, stimo quindi la media e la varianza. Due esempi sono la Stima Maximum-likelihood (ML) e la Stima di Bayes.
Gli approcci generativi non parametrici invece non fanno assunzioni sulla forma della distribuzione della probabilità. Si basano su un teorema: "Dati N punti generati cn la probabilitòà $p(x)$, se K punti cadono all'interno della regione allora $\frac{K}{N}$ è uno **stimatore corretto e consistente** di p. Da questo arrivo a:

$$p(x_0) = \frac{K}{NV}$$

dove V è il volume della regione R.
Ho due strade per calcolare un punto:

- Definisco la regione R e conto quanti punti nel mio training set cascano nlla regione. Fisso quindi R, calcolo K e ottengo $\frac{K}{NV}$. Un esempio di questo approccio è il calassificatore Parzen-Windows.
- Fisso K: vado a vedere i K oggetti più simili all'oggetto che sto cercando di classificare. Più grande è la regione che devo considerare per trovare i K punti, più bassa è la probabilità. Il classificatore più noto che fa quest'operazione è K-Nearest Neighbor.

Esercizio 2

Si descriva l'algoritmo di clustering K-means evidenziando vantaggi e svantaggi.

Svolgimento

L'algoritmo K-means è un algoritmo di clustering partizionale center-based: ogni cluster è rappresentato dalla sua media e ottimizza una funzione di errore. Descriviamone il funzionamento generale:
In ingresso decido quanti cluster voglio ottenere (il parametro K). Si parte da una clusterizzazione iniziale (casuale) e ad ogni iterazione si calcolano le medie dei clusters del passo precedente, per poi rideterminare la clusterizzazione assegnando ogni pattern alla media più vicina. Le iterazioni terminano quando c'è convergenza.
L'algoritmo si divide in una parte di Inizializzazione (in cui si genera la partizione casuale) e un ciclo di iterazioni (in cui vengono calcolate le medie e per ogni punto la distanza da esse per poi effettuare il riassegnamento alle classi più vicine agli oggetti). La condizione di stop tipicamente prevede che si faccia terminare l'algoritmo quando non ci sono cambiamenti tra l'iterazione $t-1$ e l'iterazione t (ovvero quando non cambiano nè i cluster ($y_i^{(t)}$) nè le medie ($\mu_j^{(t)}$)).
Tra i vantaggi annoveriamo la sua semplicità e il fatto che è intuitivo e quindi molto utilizzato, inoltre è molto efficiente nel clusterizzare dataset grandi, perché la sua complessità computazionale è linearmente dipendente dalla dimensione del data set.
Tra gli svantaggi però abbiamo il fatto che il numero di cluster deve essere fissato a priori, che l'ottimizzazione spesso porta ad un ottimo "locale" L'inizializzazione è inoltre cruciale: una cattiva inizializzazione porta ad un clustering pessimo. In più l'algoritmo è limitato a cluster con forma convessa e a dati vettoriali, poichè deve calcolare la media. Infine non funziona bene su dati altamente dimensionali (soffre del problema della curse of dimensionality). Parte dei problemi è risolta dalle sue varianti (K-Means ++, PAM, DPAM, ecc.)

Esercizio 3

Si descrivano in breve gli Hidden Markov Models.

Svolgimento

Gli Hidden Markov Models sono modelli probabilistici per dati sequenziali (temporali e non). Sono stati utilizzati molto a partire dal rionoscimento del parlato fino ad arrrivare ad una serie di applicazioni, in cui gli stati sequenziali potevano non essere così evidenti.
Si introducono a partire dai Modelli di Markov, definiti tramite 5 assunzioni.

1. Il sistema evolve in passi discreti.
2. Il sistema è in uno stato ad ogni istante di tempo.
3. Markovianità del primo ordine: il sistema non ha memoria, lo stato successivo dipende solo da quello corrente.
4. Modellazione probabilistica, ovvero la transizione tra gli stati è descritta in modo probabilistico.
5. Tutti gli stati sono osservabili.

La transizione tra stati viene definita tramite una matrice e le probabilità iniziali mi dicono come rimango negli stati o come passo da uno stato all'altro. La caratteristica principale però è che li stati siano osservabili e questo alle volte è limitante: per passare a un modello a stati nascosti bisogna rimuovere l'ulitma assunzione.
Negli Hidden Markov Models quindi intuisco lo stato dalle condizioni che contornano la situazione, dato che quello che osservo dipende dallo stato in cui mi trovo. Questo aggiunge un livello di incertezza: aggiungo quindi la probabilità che determinate osservazioni accadano in determinati stati.
Tecnicamente: in un modello di Markov se il sistema entra in uno stato si ha l'emissione di un solo simbolo; in un HMM se il sistema entra in uno stato si ha una distribuzione di probabilità che descrive la probabilità di osservare un determinato simbolo.
Questo mi permette di effettuare tre operazioni:

- **Valutazioni**: data una sequenza O e un modello λ, posso calcolare $P(O|\lambda)$ tramite una procedura Forward-Backward, calcolo la probabilità di avere una certa sequenza di simboli.
- **Decodifica**: data una sequenza O e un modello λ, posso calcolare la sequenza ottimale di stati generati, la sequenza di stati più probabile che il sistema segue.
- **Addestramento**: trovo il modello e i parametri ottimali del modello rispetto certe sequenze. Dato un insieme di sequenze O, determino il miglior modello λ tramite un algoritmo che si chiama Braum-Welch (EM).

Ci sono vari problemi aperti negli HMM, tra cui la scelta del numero di stati e la topologia.

Esercizio 4

Si descrivano le principali differenze tra gli algoritmi gerarchici e gli algoritmi partizionali, evidenziando in quali occasioni è più opportuno usare una classe e in quali l'altra.

Svolgimento

Algoritmi gerarchici e partizionali sono algoritmi di clustering, ovvero organizzano un insieme di patterns (entità/oggetti) in gruppi sulla base della similarità. Pattern che appertengono allo stesso gruppo sono tutti simili tra loro e gruppi diversi invece sono differenti tra loro. Questi gruppi si chiamano clusters.
Nonostante sia difficile fare una divisione, la più accettata è quella tra metodi partizionali e gerarchici.

- I metodi partizionali hanno come risultato una singola partizione del dataset (insieme di cluster disgiunti la cui unione ritorna il dataset iniziale). Tipicamente il numero di cluster è dato in ingresso. Ha come vantaggio quello di fornire un ottimo riassunto dei dati, identificando i gruppi naturali, con una rappresentazione compatta. È ideale per dataset grandi, in quanto è anche molto veloce.
  Uno svantaggio è che richiede che i dati siano rappresentati in maniera partizionale, scegliendo in anticipo il numero di cluster, che può essere un problema, estraendo solo cluster complessi.
- I metodi gerarchici invece forniscono come risultato una serie di partizioni innestate. Danno molte più informazioni rispetto al clustering partizionale, e non partono da rappresentazioni vettoriali ma da distanze, facendo si che io possa usare gli stessi algoritmi anche per altre cose. Inoltre non è necessario dare in ingresso il numero di cluster.
  Tra i vantaggi annoveriamo quello di evidenziare le relazioni tra i vari pattern, richiedendo una matrice di prossimità.
  Tra gli svantaggi c'è la lentezza e il fatto di essere greedy (quindi subottimali). Vengono usati per dataset molto piccoli, perchè già per 300 elementi la visualizzazione risulta difficile.

Esercizio 5

Si descriva il concetto di tipo di pattern nel contesto della rappresentazione dei dati. Si descrivano alcuni possibili tipi di pattern, producendo, se possibile, alcuni esempi di carattere biologico.

Svolgimento

Un pattern è come metto assieme le varie features, dove con feature intendiamo una misura che effettuo e che può essere di diversi tipi (discreta, continua, in valori binari, in valori nominali). Abbiamo diversi tipi di pattern, i più usati sono i **pattern vettoriali**: sono un insieme prefissato di features, messe in ordine, fondamentale ma arbitrario. L'oggetto viene rapresentato come un punto in uno spazio d-dimensionale, detto spazio delle features, dove d è il numero delle features (con 2 lo posso vedere, con 3 immaginare,...). La scelta del numero di features è cruciale: usando molte features, gli spazi diventano molto grandi e possiamo avere problemi, come quello della **Curse of Dimensionality**. Mentre per gli esseri umani più sono le misure che effettuo e meglio riconosco un oggetto, per un calcolatore è differente, poiché inizia ad avere problemi di stima.
Un esempio di pattern di tipo biologico sono invece le **sequenze**: si presentano in forma ordinata ed, appunto, sequenziale, in cui l'ordine è fondamentale e non arbitrario. Abbiamo inoltre i grafi, gli insiemi e varai altri tipi di pattern. Tipicamente pattern complessi sono più espressivi, ma molte tecniche si applicano meglio ai vettori.

Esercizio 6

Si descriva il problema della validazione del clustering.

Svolgimento

Validare un clustering è difficile perché, se quello che vogliamo fare è riuscire a prendere gli oggetti e, in maniera non supervisionata, trovare i gruppi all'interno del dataset, ci sono moltissime soluzioni, tutte potenzialmente valide.
Per validare abbiamo 3 possibilità:

1. **Interpretazione**
   È la più qualitativa, in cui usiamo informazioni del contesto applicativo, con l'utilizzo della conoscenza a priori sul problema (intesa anche come "interpretazione dei risultati").

2. **Validazione Quantitativa**
   È un insieme di procedure che valutano il risultato di un'analisi clustering in modo quantitativo e oggettivo.

   a. **Validazione Quantitativa Interna**

   Ottenuto un raggruppamento cerco di capire se ha senso tramite criteri interni di valutazione. Uso solo i dati disponibili, quindi è completamente non supervisionato. Scelgo un criterio di bontà e uso quello per capire se ho definito un buon clustering.

   b. **Validazione Quantitativa Esterna**

   Ottenuto un raggruppamento prendo un problema identico di cui ho già la soluzione e confronto il clustering da me ottenuto con la soluzione. Uso i criteri esterni come indici di validità:

   - Rand
   - Jaccard
   - Fowlkes
   - Mallows

   Costruisco una matrice di partizionamento: ho una funzione indicatrice $I_U(i, j)$, che vale 1 se gli oggetti i e j sono nello stesso cluster secondo il clustering 1.
   Avrò due funzioni indicatrici, quella del clustering vero e quella del mio clustering. Da qui posso calcolare la matrice di contingenza, cerco di capire quanto d'accordo vanno le due partizioni. I criteri interni sono molto difficili da stimare, devono misurare il fitting tra una partizione data e il dataset. Il problema fondamentale è stimare il numero di cluster e la baseline. Un indice usato molto spesso per stimare il numero ottimale di cluster è l'Indice ddi Davies-Bouldin.

Esercizio 7

Si descriva l'idea alla base della PCA, evidenziando vantaggi e svantaggi di tale tecnica.

Svolgimento

PCA è un approccio lineare non supervisionato per ridurre la dimensionalità delle features. L'idea è di minimizzare lo scarto quadratico medio tra i dati originali e quelli ricostruiti. Questa tecnica estrae le direzioni di massima varianza tra dati, trovando la matrice A (da $Y = A^T * X$).
L'algoritmo si basa sul fatto che l'informazione più importante da mantenere è la varianza dei dati. Per questo:

1. Estrae la direzione di massima varianza dei dati
2. Mantiene anche la maggior aderenza ai dati originali
3. Minimizza lo scarto quadratico medio.

La direzione migliore è quella a massima varianza, la seconda è quella ortogonale alla prima. L'algoritmo si divide in 6 passi:

1. Calcolare la media lungo ogni direzione
2. Centrare i dati
3. Trovare la matrice di covarianza
4. Trovare gli autovalori e autovettori della matrice
5. Ordinare gli autovalori in ordine decrescente
6. Ottenere la matrice A mettendo in colonna gli L autovalori più grandi

Un vantaggio è che sicuramente è la migliore tecnica lineare di riduzione della dimensionalità di un insieme di dati, in termini di errore quadratico medio. Inoltre i parametri del modello possono essere ricavati dai dati e la proiezione nello spazio è un'operazione molto veloce, per cui la tecnica è molto conosciuta e utilizzata. Tra gli svantaggi invece annoveriamo il costo dal punto di vista computazionale e non è chiaro il costo dei dati incompleti, inoltre non si tiene conto della densità di probabilità e non è detto che le direzioni a varianza maggiore siano quelle ottimali.

Esercizio 9

Si descrivano le principali problematiche e procedure relative all'analisi automatica dei dati derivanti da esperimenti di expression microarray.

Svolgimento

I microarray sono stati la prima tecnologia in grado di ritornare il livello di espressione dei singoli geni: gli array sono un substrato dove vengono immobilizzati i probes, che rappresentano i pezzi di DNA immobilizzati sull'array - i.e. il substrato immobile. Con un processo chiamato ibridazione si calcola l'espressione e i targets rappresentano le sequenze di cDNA che vengono ibridate sull'array - i.e. il substrato mobile. L'output è un'immagine con degli spot che misurano quanto è espresso l'RNA. Abbiamo alcuni problemi relativi all'analisi automatica di questi output:

- Segmentazione degli spot
  È necessario elaborare l'immagine per capire come quantificare il segnale e capire quali pixel appartengono allo sfondo e quali al gene. Bisogna eliminare il rumore. Nell'identificazione degli spot infatti l'array potrebbe essere ruotato e per array a due canali avere un disallineamento, devo comunque assegnare una cardinalità agli spot. In seguito bisogna normlaizzare e poi segmentare. Per la segmentazione possiamo avere più approcci (quali Fixed Circle, Adaptive Circle o Adaptive Shape).
- Quantificazione del segnale
  Dobbiamo passare dall'immagine ad un numero che esprime quanto è espresso un gene. Nella quantificazione del segnale dobbiamo stimare il foreground (intensità media, intensità mediana)
- Rilevamento della qualità
  La stima della qualità dell'array e dei dati in generale, potrei avere delle situazioni non buone che rovinano l'esperimento. In questo caso la oattern recognition può essere utile, addestrando un classificatore con degli esempi classificati a mano.

Esercizio 10

Si descrivano in breve le SVM.

Svolgimento

Le SVM (Support Vector Machines) sono classificatori binari, che individuano un iperpiano nello spazio delle features, dividendolo così in due regioni, nel caso di due dimensioni individuano una retta $y = ax + b$, che scritta come un vettore individua le due regioni:

- $w^T \cdot x + b > 0$
- $w^T \cdot x + b < 0$

Dove w è $[a - 1]$ e z (dopo nominato x) $= \begin{bmatrix} x \\ y \end{bmatrix}$ Devo quindi guardare il segno dell'equazione per capire in quale classe classificare il punto z (o x). Se ho dati linearmente separabili esiste una retta per cui tutti gli oggetti sono linearmente separabili, se esiste un iperpiano per cui le classi risultino essere perfettamente separate.
In questo caso posso stringere i vincoli: massimizzo un margine μ, dato dalla distanza dei punti più vicini alla retta di entrambe le classi. Ogni punto già classificato rappresenta un vincolo che utilizzo per creare la retta, che rafforzo creando una zona d'incertezza: $-1 < w^T \cdot x + b < 1$. Se un punto si trova nella zona d'incertezza non sono del tutto sicuro della mia classificazione. Ottengo infine un problema di ottimizzazione quadratica con vincoli:

$$\rho = \frac{2}{||w||}$$

Infine trovo w e b che minimizzano:

$$\frac{1}{2}||w|| \rightarrow \frac{1}{2}w^T \cdots w$$

Trasformando con i moltiplicatori di Lagrange, trovo un coefficiente α per ogni vincolo - $L(w, b, \alpha)$: alla fine trovo i coefficienti α_i per cui la funzione risulta ottimizzata ($w = \sum_{i=1}^{N} \alpha_i y_i x_i$).
Tutti gli α sono uguali a zero tranne i punti vicini al margine: i support vectors, ovvero gli unici punti necessari alla definizione dell'iperpiano.
Quando ho dati non linearmente separabili invece non esiste nessuna retta tale per cui le due classi siano separate, per cui rilasso i vincoli utilizzati. e introduco un valore ξ_i che indica la posizione del vettore rispetto all'iperpiano.

$$w^T \cdot x_i + b \geq 1 - \xi_i$$

Se $\xi_i = 0$ la classificazione è corretta, se è tra 0 e 1 allora è corretta ma nella zona al margine, altrimenti è errata. Voglio fare ovviamente in modo che gli ξ_i siano i più piccoli possibile, si chiamano Slack Variables.

Esercizio 11

Si descriva la regola di decisione di Bayes per la classificazione, evidenziandone vantaggi e svantaggi.

Svolgimento

Dato un problema, dopo averlo rappresentato si provvede a costruire il modello attraverso l'uso del training set. Uno dei problemi darisolvere è la classificazione (ASSEGNARE LA CLASSE A UN OGGETTO). L'obiettivo è quello di costruire un classificatore determinando una funzione $f()$ che dato un pattern in input x ritorna delle etichette $y \rightarrow y = f(x)$.
La teoria delle decisioni di Bayes è un metodo per classificare (discriminare tra le diverse classi) che fa uso di metodi probabilistici, in cui si conoscono sempre tutte le probabilità necessarie (è molto usato nella classificazione a oggetti). Siano $w_1, \ldots, w_n$ le classi disponibili,dato un oggetto:a quale classe deve essere assegnato? Come detto prima il problema della decisione è posto in termini probabilistici e sono diverse probabilità per costruire la regola di decisione:

1. **Probabilità a priori**: utilizzo solo le informazioni a priori, ovvero assegno x alla classe con maggior probabilità.Se $P(w_1) > P(w_2)$ decidi w_1, w_2 altrimentiÈ un sistema limitato poichè non si considerano i pattern.

2. **Probabilità condizionale(LIKELIHOOD)**: misura la probabilità di avere la misurazione x conoscendo lo stato di natura della classe w_j. Se $P(x|w_1) > P(x|w_2)$ decidi w_1, altrimenti decidi w_2. È migliore della regola basata sulla probabilità a priori perchè qui si considera l'ossevazione, ma si basa solo sull'osservazione.
   SOLUZIONE = REGOLA DI DECISIONE DI BAYES: mette insieme la probabilità a priori e la probabilità condizionale nella

3. **Probabilità a posteriori**:
   $$\frac{P(x|w_j)P(w_j)}{P(x)} \rightarrow posterior = \frac{likelihood \times priori}{evidenza}$$
   si moltiplicano la probabilità a priori e la probabilità condizionale divendendo per l'evidenza come fattore scala che descrive quanto frequentemente si osserva un pattern x. Non dipende da w_1 o w_2, perciò è ininfluente per la regola di decisione. Se $P(x|w_1)P(w_1) > P(x|w_2)P(w_2)$ decidi w_1, altrimenti decidi w_2.

Nella pratica le probabilità non sono note, il classificatore si costruisce dall'apprendimento da esempi. Si utilizza un training set per effettuare una stima delle probabilità, per poi applicare Bayes in un secondo momento.Vi sono diversi approcci della stima di probabilità:

- Per stime parametriche(si conosce la forma della prob. di funzione e se ne vogliono stimare i parametri) → GAUSSIANA
- Per stime non parametriche(non si conosce la forma la prob. della funzione è stimata dirattamente dai dati) → ISTOGRAMMA
- Per stime semi-parametriche(i parametri possono cambiare la forma della funzione) → NEURAL NETWORKS

VANTAGGI: stima accurata
SVANTAGGI: stimare la posterior non è sempre banale, integrare in tutto lospazio dei parametri può essere difficile

Esercizio 12

Descrivere la costruzione del modello 3 fasi.

Svolgimento

Il problema principale è quello di assegnare la classe a un oggetto. Per discriminare la classe possono essere utilizzati tre metodi: classificazione,clustering,detection. In tutti i casi va costruito il modello a partire dai dati. Per farlo e poi usarlo bisogna passare sotto tre fasi.

1. **Rappresentazione**: Obiettivo è quello di rappresentare digitalmente il problema in modo che il calcolatore possa comprendere i dati. A sua volta la rappresentazione può essere suddivisa in tre parti: campionamento, rappresentazione, pre-processing.

   Campionamento: Rappresenta la raccolta vera e propria dei dati, effettuando le misure suglio oggetti delproblema. Durante il campionamento è bene tenere presente: frequenza di campionamento,risoluzione, capacità di gestione ai cambiamenti di contorno.

   Rappresentazione: Si rappresentano completamente gli oggetti, estraendo le features (le features devono essere rilevanti, discriminanti, quantificabili e interpretabili) e salvandole nella struttura dati che è il pattern (tipi di pattern: vettori, sequenze, insiemi,grafi).

   Pre-processing/scaling: con il quale si apportano delle ottimizzazioni varie in primo luogo standardizzare i dati cioè uniformare le scale. Per farlo abbiamo due vie:
   1) *Data standardization*: Produce dati senza dimensionalità, le operazioni vengono eseguite dimensione per dimensione e si producono dati che sono in formato standard; vengono perciò perse molte informazioni riguardo la scala e la locazione dei dati. Da $x, y, z \rightarrow f(x), f(y), f(z)$. Esempi : Z-score, standadizzazione della deviazione standard, DOMAIN, DOMAIN per la deviazione standard.
   2) *Data transformation*: L'obiettivo è sempre quello di migliorare la rappresentazione, le operazioni sono eseguite su tutte le dimensioni simultaneamente. Da $x, y, z \rightarrow f_1(x, y, z), f_2(x, y, z), f_3(x, y, z)$. La data transformation opera una combinazione lineare delle features. Vuole ridurre la dimensionalità dello spazio mantenendo la maggior quantità di informazione possibile, e mostrare particolari strutture. Esempi : PCA (supervisonata), LDA (non supervisionata).

2. **Costruzione del modello**: dopo aver rappresentato correttamente i dati devo trovare un modelloche spieghi i dati del training set. Il training set deve essere largo, completo, variabile. La costuzione del modello avviene proprio attraverso l'uso del training set e attraverso conoscenze a priori (per esempio classificazione e detection che avvengono in modo supervisionato).

3. **Testing**: una volta costruito il modello è possibile verificare se l'oggetto (dopo la sua rappresentazione) appartiene o meno al modello. Quindi testo l'algoritmo.

Esercizio 13

Si descriva il problema della validazione del classificatore.

Svolgimento

È importante capire se il sistema di classificazione disegnato rappresenta una buona scelta. L'obiettivo principale è quello di misurare la capacità di generalizzazione: capacità di classificare correttamente anche oggetti sconosciuti (non presenti nel training set).NB: non è detto che classificare bene gli oggetti presenti nel training set implichi una buona capacità di generalizzazione. es. Overtraining : il sistema ha imparato talmente bene i pattern del training set che non è più in grado

di generalizzare.Per testare la capacità di generalizzazione è bene avvalersi di un insieme di oggetti inerenti al problema, ma che non fanno parte del training set, questo insieme è chiamato testing set e serve per testare le capacità discriminative del classificatore costruito, contando gli errori di classificazione.
Il testing set può essere costruito partendo da:

- **Nuovo Campionamento**: riestrarre altri esempi dal problema e utilizzarli per testare, è un metodo corretto ed efficace però non sempre possibile e troppo dispendioso.
- **Cross Validation**: una divisione del training set - divido in vari modi il training set e una parte la uso per fare il testing. Vi sono per questo metodo diverse varianti:

  - HOLDOUT= l'insieme dei dati viene partizionato casualmente in due partidi egual dimensione. Una delle due parti viene usata come testing.

  - AVERAGED HOLDOUT= per rendere il risultato meno dipendente dalla partizione, si effettuano una serie di divisioni holdout e si fa una media sui risultati trovati.

  - LEAVE ONE-OUT= si esclude un solo oggetto xi dal training set. Dopo la fase di apprendimento si usa xi come oggetto di testing. Si ripete l'operazione per tutti gli oggetti xipossibili e fa una media dei risultato ottenuti. Questo metodo funziona bene per insiemi di dati ridotti.

  - LEAVE K-OUT (S-FOLD)= si partiziona l'insieme in S porzioni. Si usano le S-1 porzioni per istruire il classificatore usando l'ultima come test. Questa operazione viene eseguita S volte facendo la media finale dei risultati.

È bene inoltre, per la validazione del classificatore, tenere presente:
le LEARNING CURVES, sapendo che l'errore di classificazione diminuisce all'aumentare degli elementi dell'insieme;
le MATRICI DI CONFUSIONE, che spiegano como si comporta un classificatore rispetto alle varie classi (calcolando accuratezza,precisione,sensibilità,specificità)
la SIGNIFICATIVITÀ STATISTICA, uno strumento per comparare tra di loro classificatori.