Recupero e Riconoscimento dell'Informazione per Bioinformatica

Domande in stile esame

Chiara Solito

Corso di Laurea in Bioinformatica
Università degli studi di Verona
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

Classificazione generativa: crea modelli per le classi, si calcola likelihood e prior Classificazione discriminativa: si calcola direttamente le posterior e il confine di decisione. Generativo parametrico: la prior è numero di oggetti in una classe / numero di tutti gli oggetti. Per la condizionale devo stimare un parametro phi (media, varianza) e calcolarlo.

Esercizio 2

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

Svolgimento

Il k-means è un algoritmo di clustering partizionale. L'idea iniziale è che ogni cluster è rappresentato dalla sua media. Nella fase di inizializzazione inizializzo due medie a caso e per ogni punto trovo quale media sia la più vicina. Nella fase seguente ripristino le medie dei cluster e ricalcolo le distanze di ogni punto dalla media e li riassegno nei cluster. Ripristino le medie e ricalcolo le distanze e riassegno i punti al cluster più vicino fino a quando le medie non cambiano. Vantaggi: ottimo per clusterizzare dataset grandi, in qunato la complessità è linearmente dipendente dalla dimensione del dataset. Svantaggi:il numero di cluster deve essere fissato a priori, l'ottimizzazione porta ad un ottimo locale, lavora solo su dati vettoriali numerici, non funziona bene su dati altamente dimensionali (soffre di curse of dim).

Esercizio 3

Si descrivano in breve gli Hidden Markov Models.

Svolgimento

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

CP il risultato è una singola partizione di dati, nel CG è una serie di partizioni innestate (dendrogramma). CP richiede che i dati siano rappresentati sotto forma di vettori, CG richiede matrice di similarità. CP ottimo per dataset grandi, CP bisogna settare numero di cluster, CG no; CG è più informativo del CP.

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 un insieme di feature relative ad uno stesso oggetto. Esistono molti diversi tipi di pattern. Quelli molto usati sono i vettori, dove ogni oggetto ha un insieme prefissato di features, messe in ordine in un vettore. Altri tipi sono le sequenze, ovvero dati che si presentano in forma ordinata e sequenziale, un esempio sono le seq di aa o di nucleotidi. Altri tipi sono i grafi (insieme di nodi collegati da archi) e gli insiemi (collezioni non ordinate di features).

Esercizio 6

Si descriva il problema della validazione del clustering.

Svolgimento

La validazione è un insieme di procedure che valutano il risultato di un'analisi di clustering in maniera oggettiva e quantitativa. In base al tipo di custering effettuato si usano diversi indici, usando anche conoscenza a priori del problema (es etichetta delle classi).

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 dati originali e quelli ricostruiti. Questa tecnica estra le direzioni di massima varianza tra dati. Proietta i dati nella prima direzione che è quella di max varianza, la seconda è max varianza e prtogonale alla rima. Si basa sul calcolo dei autovalori e autovettori della matrice di covarianza dei dati. Si iniza calcolando la media

lungo ogni direzione m, si sottrae la media dai dati, si trova la matrice di covarianza, si trovano autovalori e autovettori di C, si ordinano gli autoavlori e i primi L autovettori corrispondono alla matrice A di trasformazione. Vantaggi: miglior tecnica oer ridurre la dimensionalità dell'insime di dati, i parametri del modello possono essere ricavati dai dati, la proiezione nello spazio è un'operazione veloce. Svantaggi: alto costo computaizonale per il calcolo dei parametri, non si sa come gestisca dati incompleti, non tiene conto della densità della probabilità dello spazio considerato, non è detto che le irezioni a varianza maggiore siano le migliori.

Esercizio 9

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

Svolgimento

Per i microarray i problemi principali sono segmentazione degli spot e la rimozione del rumore, la quantificazione del segnale e il rilevamento della qualità. Nell'identificazione degli spot infatti l'array potrebbe essere ruotato e per array a due canali avere un disallineamento. Nella segmentazione dobbiamo decidere cosa è il segnale e cosa è il background. Alcune soluzioni sono Fixed circle, adaptive circle ecc. Nella quantificazione del segnale dobbiamo stimare il foreground (intensità media, intensità mediana).

Esercizio 10

Si descrivano in breve le SVM.

Svolgimento

Sono classificatori binari e suddividono lo spazio in due regioni tramite un iperpiano. Esse definiscono l'iperpiano ottimale come quello che massimizza il margine mu, che è la distanza minima tra le due classi (usando vettori di supporto, cioè esempi del trainig set). La soluzione trovata è ottimale (minimizzazione convessa) e il metodo è geometrico e non robabilistico. $y(x) = wx + b$. Quando si addestra si trova il vettore w e il parametro b ottimali (che massimizzano il margine) a partire dal training set, il margine vale $2/||w||$ Se i dati non sono linearmente separabili si introduzono le slack variable epsilon che consentono la classificazione errata di qualche punto. Il vincolo diventa $y_i(wx_i + b) \geq 1 - \epsilon_i$. In questo modo alcuni punti possono attraversare il margine pagando un prezzo ϵ_i. È necessario minimizzare in numero di errori, quindi troare w e b che minimizzino $1/2ww + C$ Sommatoria ϵ_i Il parametro C pesa gli errori. I support vector ora sono i punti sul mergine e oltre il margine

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

Svolgimento