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

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

Esercizio 6

Si descriva il problema della validazione del clustering.

Svolgimento

Esercizio 7

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

Svolgimento

Esercizio 9

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

Svolgimento

Esercizio 10

Si descrivano in breve le SVM.

Svolgimento

Esercizio 11

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

Svolgimento

Esercizio 12

Svolgimento