Lezione 5

Corso	Riconoscimento e recupero dell'informazione per Bioinformatica
Data	@November 8, 2021 1:30 PM
Rifacimento	☑
Status	Completed
Tipo	Lezione

Teoria della decisione di Bayes

Rappresentare tutto il problema di rappresentzione in termini probabilistici. La regola di decisione di Bayes si basa su alcune probabilità, in particolare su quella posteriore $p(\omega_j|x)$.

Vado a vedere la classe la cui posteriore è massima, e le assegno x_0 - *argomento massimo*.

es. max[10, 20, 30] argmax = 3 (la posizione dell'argomento massimo.

$$\frac{p(x|\omega_j)p(\omega_j)}{p(x)}$$

$p(\omega_j)$ ho una probabilità a priori: quanto è probabile osservare in natura una determinata classe

$p(x|\omega_j)$mi dà l'informazione della distribuzione degli oggetti per la classe.

Contesto generale:

funzioni discriminanti lineari - posso formulare un classificatore, e assegno un oggetto alla classe la cui $g(x)$è massima, dove $g(x)$ è una funzione discriminante.

Passiamo in questo contesto così da poterlo interpretare in maniera delle geometrica: supponendo di avere uno spazio delle feature, posso isolare una regione dello spazio dove ad esempio la prima funzione $g_1(x)$ è più grande della seconda $g_2(x)$, e altre regioni in cui succede il contrario.

Posso creare una linea intorno alla quale $g_1(x) = g_2(x)$.

Questo ci permette di capire che posso affrontare il problema in due modi:

- andando a calcolare $g_1(x)$ e $g_2(x)$

  Un modo generico di formulare un classificatore: calcolo g_{1}(x) e g_{2}(x) e assegno l'oggetto alla classe che risponde più fortemente alla funzione.

- andando a stimare la funzione $g_1(x) = g_2(x)$, ovvero $g_1(x) - g_2(x)$: in maniera più intelligente

$$g(x) = g_1(x) - g_2(x)$$

Ragionando dal punto di vista di Bayes:

posso definire g_{1}(x) come voglio, ma se in particolare la descrivo come posterior:

- $g_1(x)$ è la mia Posterior $(p(\omega_1|x)$: calcolare le funzioni equivale ad applicare la regola di decisione di Bayes:

  Assegno x ad ω_1 se $p(\omega_1|x) \geq p(\omega_2|x)$

  Due strade che corrispondono a due percorsi diversi.

  - posso cercare di stimare le due funzioni: p(\omega_{1}|x) e p(\omega_{2}|x) → **APPROCCIO GENERATIVO**
  - posso cercare di stimare direttamente $g(x)$, il "confine" → **APPROCCIO DISCRIMINATIVO**

Approccio Generativo

Devo stimare $p(\omega_1|x)$ e $p(\omega_2|x)$: dato che è difficile uso:

$$\frac{p(\omega_1|x) = p(x|\omega_1)p(\omega_1)}{p(x)}$$

e

$$\frac{p(\omega_2|x) = p(x|\omega_2)p(\omega_2)}{p(x)}$$

che è più facile da stimare.

Devo stimare quindi: $p(\omega_1)$ **e** $p(x|\omega_1)$

Come?

Apprendimento da esempi.

- stimare la probabilità a priori: il modo più semplice è vedere quanto rappresentata è la classe all'interno del training set, uso la cardinalità delle classi del training set
- stimare la probabilità di x dato ω_1: come si distribuiscono gli oggetti all'interno delle classi?

  Due possibilità:

  - **approccio parametrico**: conosco la forma, il problema è stimare i parametri. Devo stimare quale specifica delle forme descrive meglio la mia classe. Devo stimare la media e la varianza della funzione.

    Assumo una determinata forma (es. gaussiana), e poi ne stimo i parametri.

    Problema: se sbaglio la forma della distribuzione, la stima che faccio è pessima! In molti problemi di riconoscimento non è possibile assumere la forma.

    In alcuni casi posso provare a stimare la forma.

  - **approccio non parametrico:** non faccio nessuna assunzione sulla forma della distribuzione della probabilità.

    Se voglio stimare la probabilità di un punto in una regione dello spazio, vado a vedere cosa succede nell'intorno: analizzo le singole regioni dello spazio. Es. se nell'intorno non ho punti di training set, la probabilità sarà bassissima. Vado a ragionare per regione nell'intorno del punto e nella quantità

Passo 1:

Sia x_{0} un punto generato con la probabilità p(x). Si definisce P come la probabilità che il punto appartenga alla regione.

Ragioniamo come in un caso monodimensionale, in cui la regione è semplicemente un intervallo, un segmento di una certa lunghezza.

$$P(x_0 \in R) = \sum_{x=2, x=3, x=4} p(x) = \sum_{2 \leq x \leq 4} p(x)$$

ragionando nel continuo arrivo a:

$$P = Prob(x_0 \in R) = \int_R p(x)dx$$

> 💡 **Teorema**:
> **Dati N punti generati con la probabilità $p(x)$, se k punti cadono all'interno della regione, allora $\frac{k}{N}$ è uno stimatore corretto e consistente di P.**

Passo 2:

Se

i. $p(x)$ è continua

ii. $p(x)$ non varia molto all'interno della regione R

allora possiamo scrivere:

$$P = \int_R p(x)dx \approx p(x_0)V$$

Dove V è il volume della regione R e x_{0} è un qualsiasi punto interno alla regione. Allora:

$$\frac{K}{N} = p(x_0)V \rightarrow \forall x_0 \in R, p(x_0) = \frac{K}{NV}$$

Sto facendo due assunzioni che sono in contrasto tra di loro: la prima è meglio se la regione R è ampia, la seconda se invece la regione è piccola.

Quindi ora la regola è:

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

Ho due strade per calcolare un punto:

- Ho x_{0}: **definisco la regione R** e conto quanti punti nel mio training set cascano nella regione. Fisso quindi R, calcolo K e ottengo $\frac{K}{NV}$

Più sono i punti nel volume fissato V, più alta è la probabilità.

CLASSIFICATORE: Parzen Windows

- **Fisso K**: Qual è il senso? Se voglio almeno un tot di punti ma per farlo devo avere un volume molto grane e la probabilità sarà grossolana.

  Più grande è la regione cege devi considerare per trovare K punti, più bassa è la probabilità.

  CLASSSIFICATORE: K-Nearest Neighbor

K-Nearest Neighbor

Classificatore famoso, generativo, non parametrico, semplice, intuitivo e molto utilizzato.

Si basa sul concetto di similarità.

> 💡 *Idea*: se devo classificare un oggetto ì, vado a vedere gli oggetti che gli sono più simili.

Il **Nearest Neighbor**:

> l'oggetto più simile tra tutti a quello che sto analizzando. Classifico l'oggetto come il suo nearest neighbor. (K=1)

Posso generalizzare la cosa:
supponendo di avere un training set (insieme di oggetti della classe 1, insieme della classe 2 e il mio x_0 da classificare).

- fisso un k
- vedo quali sono i k punti più vicini all'oggetto di test nel training set
- classifico questo oggetto a seconda dei suoi vicini

Questo classificatore implementa la regola di Bayes dove le probabilità a posteriori sono stimate in modo generativo e in modo non parametrico.

> Vantaggi:
>
> - tecnica semplice e flessibile
> - estremamente intuitiva

- funziona anche per dati non vettoriali
- ragionevolmente accurata
- ci sono pochi parametri da aggiustare
- sono stati dimostrati molti risultati teorici su questa tecnica

XAI (explainable artificial intelligent) ha dato una "seconda vita" a KNN

Il classificatore

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

Notazioni:

- $X = \{(x_1, y_1), ..., (x_n, y_n)\}$
  dove x_i è l'i-esimo oggetto e y_i la sua etichetta
- c è il numero di classi
- N_j è il numero di oggetti in X che appartengono alla classe \omega_j
- $N_j = |\omega_j|$
- $\sum_{j=1}^{c} N_j = N$
- K è il parametro di KNN

Goal: classificare x_0, con la regola di Bayes

Devo calcolare $p(\omega_1|x)$ e $p(\omega_2|x)$

Quindi scompongo:

$$p(\omega_j|x) = \frac{p(x_0|\omega_j)p(\omega_j)}{p(x_0)}$$

$p(x_0)$ = considero i K punti più vicini a x_0 → regione R con volume V

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

$p(x_0|\omega_j)$ = $\frac{K_j}{N_j V}$ dove K_j è il numero di punti in $R \in \omega_j$

$p(\omega_j) = \frac{N_j}{N}$

Quindi:

$$\frac{\frac{K_j}{N_j V}\frac{N_j}{N}}{\frac{K}{NV}} \rightarrow \frac{K_j}{K}$$

Per ogni classe prendo il massimo. Dei K punti presi, K_1 appartengono alla classe 1, ... K_j appartengono alla classe j. Prendo la classe più frequente all'interno dell'insieme/regione R.

Svantaggi:

- tutti i punti del training set devono essere mantenuti in memoria
- la scelta è molto locale, prendendo quindi pochi punti nello spazio
- la scelta della misura di similarità è fondamentale: sbagliandola, sbaglio tutto
- a seconda della matrice utilizzata, potrebbe essere necessario preprocessare lo spazio
- la scelta di K è spesso cruciale:
  - legato al grado di non-linearità che voglio

Scegliere K

Vado a vedere cosa succede nella regione, o lo faccio provando diversi valori e guardando l'andamento del classificatore, oppure empiricamente la radice quadrata degli oggetti.

Genericamente si usa un K dispari.

Ridurre la dimensione del training-set

Cancello punti che non hanno effetto sul confine di decisione (condensing) - non è detto però che questi punti non siano importanti per il testing set