Classificatore Naive di Bayes

Una delle tecniche più semplici e popolari

Quando usarlo:

- insiemi di dati di dimensione abbastanza grande
- gli attributi che descrivono le istanze sono condizionalmente indipendenti data la classificazione

Applicazioni su cui ha avuto successo:

- Diagnosi
- Classificazione di documenti testuali

Classificatore Naive di Bayes

Funzione target $f : X \to V$, con istanze x descritte da attributi $\langle a_1, a_2 \ldots a_n \rangle$.

Il valore pi`u probabile di $f(x)$è:

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} P(v_j | a_1, a_2, \ldots, a_n) \\ &= \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)} \\ &= \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n | v_j) P(v_j) \end{aligned}$$

assunzione Naive di Bayes:

$$P(a_1, a_2, \ldots, a_n | v_j) = \prod_i P(a_i | v_j)$$

che defi nisce il

Classificatore Naive di Bayes: $v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_j | v_j)$

Algoritmo Naive di Bayes

Naive_Bayes_Learn($esempi$)

For each valore target v_j

$\hat{P}(v_j) \leftarrow$ stima $P(v_j)$

For each valore di attributo a_i di ogni attributo a

$\hat{P}(a_j|v_j) \leftarrow$ stima $P(a_j|v_j)$

Classify_New_Instance(x)

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_j|v_j)$$

Consideriamo il problema *Giocare a Tennis*, e la nuova istanza

$(Outlk = sun, Temp = cool, Humid = high, Wind = strong)$

Vogliamo calcolare:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_j|v_j)$$

Consideriamo il problema *Giocare a Tennis*, e la nuova istanza

$$\langle Outlk = sun, Temp = cool, Humid = high, Wind = strong \rangle$$

Vogliamo calcolare:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_j | v_j)$$

Giocare a Tennis!!

E' la giornata ideale per giocare a Tennis ?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Consideriamo il problema *Giocare a Tennis*, e la nuova istanza

$$\langle Outlk = sun, Temp = cool, Humid = high, Wind = strong \rangle$$

Vogliamo calcolare:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_j|v_j)$$

$$P(y)P(sun|y)P(cool|y)P(high|y)P(strong|y) = 0.005$$

$$P(n)P(sun|n)P(cool|n)P(high|n)P(strong|n) = 0.021$$

$$\rightarrow v_{NB} = n$$

Naive Bayes: Considerazioni Aggiuntive

1. L'assunzione di indipendenza condizionale è spesso violata

$$P(a_1, a_2, \ldots, a_n | v_j) = \prod_i P(a_i | v_j)$$

- ...ma sembra funzionare comunque. Notare che non è necessario stimare correttamente la probabilità a posteriori $\hat{P}(v_j|x)$; è sufficiente che

$$\arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_j | v_j) = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_j | v_j)$$

- la probabilità a posteriori calcolata da Naive Bayes è spesso vicina a 1 o 0 anche se non dovrebbe

2. cosa succede se nessun esempio di apprendimento con valore di target v_j possiede valore di attributo a_i? In tal caso

$$\hat{P}(a_i|v_j) = 0, \text{ e... } \hat{P}(v_j) \prod_i \hat{P}(a_i\ v_j) = 0$$

Una soluzione tipica è la stima Bayesiana per $\hat{P}(a_i|v_j)$

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$$

dove

- n è il numero di esempi di apprendimento per cui $v = v_j$,
- n_c numero di esempi per cui $v = v_j$ e $a = a_i$
- p è la stima a priori per $\hat{P}(a_i|v_j)$
- m è il peso dato a priori (cioè il numero di esempi “virtuali”)

Esempio di Applicazione: Classificazione di Documenti Testuali

- apprendere quali documenti sono di interesse
- apprendere a classificare pagine web per argomento
- ...

Il classificatore Naive di Bayes costituisce una delle tecniche più utilizzate in questi contesti

Quali attributi usare per rappresentare documenti testuali ?

Concetto target $Interessante? : Documento \rightarrow \{+, -\}$

1. Rappresentare ogni documento tramite un vettore di parole
   - Un attributo per posizione di parola nel documento
2. Apprendimento: usare gli esempi di apprendimento per stimare
   - $P(+), P(-), P(doc|+), P(doc|-)$

Assunzione di indipendenza condizionale di Naive Bayes

$$P(doc|v_j) = \prod_{i=1}^{lunghezza(doc)} P(a_i = w_k|v_j)$$

dove $P(a_i = w_k|v_j)$ è la probabilità che la parola in posizione i sia w_k, dato v_j

Una assunzione addizionale: $P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \ \forall i, m$

LEARN_NAIVE_BAYES_TEXT($Esempi, V$)

1. collezionare tutte le parole ed altri token che occorrono in $Esempi$

- $Vocabolario \leftarrow$ tutte le parole distinte ed altri token in $Esempi$

2. calcolare (stimare) i termini $P(v_j)$ e $P(w_k|v_j)$

- **for each** valore di target v_j in V **do**
  - $doc_j \leftarrow$ sottoinsieme di $Esempi$ per cui il valore di target è v_j
  - $P(v_j) \leftarrow \frac{|doc_j|}{|Esempi|}$
  - $Text_j \leftarrow$ un unico documento creato concatenando tutti i documenti in doc_j
  - $n \leftarrow$ numero di parole e token totali in $Text_j$ (contando parole e token duplicati pi`u volte)
  - **for each** parola e token w_k in $Vocabolario$
    * $n_k \leftarrow$ numero di volte che w_k occorre in $Text_j$
    * $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabolario|}$