Naive bayes

Course of Machine Learning
Master Degree in Computer Science
University of Rome “Tor Vergata”

Giorgio Gambosi

a.a. 2018-2019

Features numerose

In presenza di numerose features, definire un modello generativo può risultare inaccettabilmente costoso.

La dimensione del training set cresce esponenzialmente con il numero di features, per permettere di considerare tutte le possibili combinazioni di valori delle features stesse

In presenza di d features, ad esempio binarie, in un classificatore binario, sarà necessario stimare approssimativamente 2^{d+1} parametri: le probabilità $p_{ij} = p(x_i|C_j)$.

Esempio: antispam per email

Esempio di *text classifier*: cerca di individure e-mail spam. Nell'ipotesi di avere un training set (messaggi marcati come spam o non-spam), le features da associare ad un messaggio saranno le parole che esso contiene (o non contiene).

Un messaggio è rappresentato da un vettore di features $\mathbf{x}$ di dimensione pari al numero di termini di un dizionario: se il messaggio contiene l'i-esimo termine del dizionario allora $x_i = 1$ (o x_i è pari al numero di occorrenze del termine), altrimenti $x_i = 0$.

Esempio: antispam per email

Se il dizionario contiene 10.000 termini e si rappresenta la sola occorrenza o meno di ogni termine (per cui il vettore è binario), allora $N = 2^{10000}$. Questo comporta la necessità di stimare $2(2^{10000} - 1)$ parametri per $p(\mathbf{x}|C_1)$ e altrettanti per $p(\mathbf{x}|C_2)$.

Le features x_i sono *condizionalmente indipendenti* rispetto alle classi C_1, C_2. Per ogni coppia i, j con $1 \leq i, j \leq d$, $i \neq j$ e per $k \in \{1, 2\}$

$$p(x_i|x_j, C_k) = p(x_i|C_k)$$

Sotto questa ipotesi, abbiamo che

$$p(x_i, x_j|C_k) = p(x_i|x_j, C_k)p(x_j|C_k) = p(x_i|C_k)p(x_j|C_k)$$

e quindi

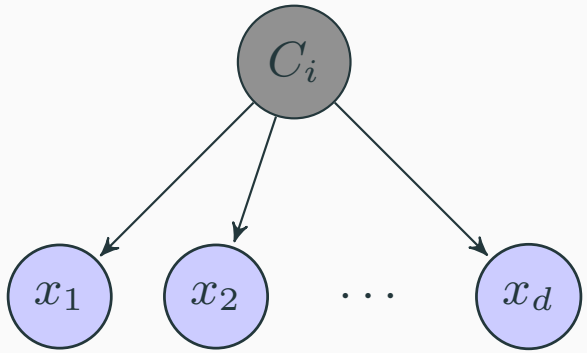
$$p(\mathbf{x}|C_k) = p(x_1, \ldots, x_d|C_k) = \prod_{i=1}^{d} p(x_i|C_k)$$

e i parametri da stimare sono $\phi_{ik} = p(x_i = 1|C_k)$ per $i = 1, \ldots, d$ e $k = 1, 2$, per numero totale di parametri pari a $2d$. Inoltre, sarà necessario stimare $\pi = p(C_1)$.

Classificatori Naive Bayes

Rappresentazione grafica (reti bayesiane)

- Nodi: variabili casuali
- Archi: relazioni di dipendenza
- Ombreggiatura: variabili visibili



Dato un training set $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, possiamo scrivere la corrispondente verosimiglianza come

$$
\begin{aligned}
L &= \prod_{i=1}^{n} p(\mathbf{x}_i, y_i) \\
&= \prod_{i=1}^{n} p(\mathbf{x}_i | y_i) p(y_i) \\
&= \prod_{i=1}^{n} p(y_i) \prod_{j=1}^{d} p(x_{ij} | y_i) \\
&= \prod_{i=1}^{n} \pi^{y_i} (1-\pi)^{1-y_i} \prod_{j=1}^{d} \phi_{j1}^{x_{ij} y_i} (1-\phi_{j1})^{(1-x_{ij}) y_i} \phi_{j2}^{x_{ij}(1-y_i)} (1-\phi_{j2})^{(1-x_{ij})(1-y_i)}
\end{aligned}
$$

E la log-verosimiglianza quindi come

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log p(y_i) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log p(x_{ij}|y_i) \\
&= \sum_{i=1}^{n} y_i \log \pi + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi) \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} y_i \log \phi_{j1} + \sum_{i=1}^{n} \sum_{j=1}^{d} (1 - x_{ij}) y_i \log(1 - \phi_{j1}) \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} (1 - y_i) \log \phi_{j2} + \sum_{i=1}^{n} \sum_{j=1}^{d} (1 - x_{ij})(1 - y_i) \log(1 - \phi_{j2})
\end{aligned}
$$

Massimizziamo la log-verosimiglianza annullando la relativa derivata. Per π questo dà

$$\frac{\partial l}{\partial \pi} = \sum_{i=1}^{n} \frac{y_i}{\pi} - \sum_{i=1}^{n} \frac{1 - y_i}{\pi} = \frac{n_1}{\pi} - \frac{n_2}{1 - \pi}$$

ponendo la derivata pari a 0 otteniamo quindi

$$\pi = \frac{n_1}{n_1 + n_2} = \frac{n_1}{n}$$

Per ϕ_{j1} abbiamo:

$$\frac{\partial l}{\partial \phi_{j1}} = \sum_{i=1}^{n}\sum_{j=1}^{d}\frac{x_{ij}y_i}{\phi_{j1}} - \sum_{i=1}^{n}\sum_{j=1}^{d}\frac{(1-x_{ij})y_i}{1-\phi_{j1}} = \frac{n_{j1}}{\phi_{j1}} - \frac{n_1 - n_{j1}}{1-\phi_{j1}}$$

dove n_{jk} è il numero di elementi del training set nella classe C_k aventi la j-esima componente $x_j = 1$ e $n_k - n_{jk}$ è il numero di elementi aventi la j-esima componente $x_j = 0$.

Ponendo la derivata pari a 0 otteniamo

$$\phi_{j1} = \frac{n_{j1}}{n_{j1} + (1 - n_{j1})} = \frac{n_{j1}}{n_1}$$

Allo stesso modo, otteniamo per ϕ_{j2}

$$\phi_{j2} = \frac{n_{j2}}{n_2}$$

Per classificare $\mathbf{x}$ si calcolano $p(C_1|\mathbf{x})$ e $p(C_2|\mathbf{x})$, assegnando $\mathbf{x}$ alla classe di maggiore probabilità. Dato che

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x})} = \frac{\prod_{i=1}^{d} p(x_i|C_1)p(C_1)}{p(\mathbf{x})}$$

e

$$p(C_2|\mathbf{x}) = \frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x})} = \frac{\prod_{i=1}^{d} p(x_i|C_2)p(C_2)}{p(\mathbf{x})}$$

per classificare $\mathbf{x}$ è sufficiente confrontare

$$\prod_{i=1}^{d} p(x_i|C_1)p(C_1) = \pi \prod_{i=1}^{d} \phi_{i1}^{x_i}(1-\phi_{i1})^{1-x_i} = \frac{n_1}{n} \prod_{i=1}^{d} \left(\frac{n_{i1}}{n_1}\right)^{x_i} \left(\frac{1-n_{i1}}{n_1}\right)^{1-x_i}$$

e

$$\prod_{i=1}^{d} p(x_i|C_2)p(C_2) = (1-\pi) \prod_{i=1}^{d} \phi_{i2}^{x_i}(1-\phi_{i2})^{1-x_i} = \frac{n_2}{n} \prod_{i=1}^{d} \left(\frac{n_{i2}}{n_2}\right)^{x_i} \left(\frac{1-n_{i2}}{n_2}\right)^{1-x_i}$$

In definitiva, è sufficiente confrontare

$$\frac{1}{n_1^{d-1}} \prod_{i=1}^{d} n_{i1}^{x_i} (1 - n_{i1})^{1-x_i}$$

e

$$\frac{1}{n_2^{d-1}} \prod_{i=1}^{d} n_{i2}^{x_i} (1 - n_{i2})^{1-x_i}$$

Esempio

	docID	termini	in c = Cina?
training set	1	Cinese Pechino Cinese	yes
	2	Cinese Cinese Shanghai	yes
	3	Cinese Macao	yes
	4	Tokyo Giappone Cinese	no
test set	5	Cinese Cinese Cinese Tokyo Giappone	?

Esempio: stima parametri

Priors: $P(c) = \pi = 3/4$ e $1 - \pi = 1/4$

Probabilità condizionate:

$$
\begin{aligned}
\hat{P}(\text{Cinese}|c) &= 3/3 = 1 \\
\hat{P}(\text{Shanghai}|c) = \hat{P}(\text{Macao}|c) &= 1/3 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Giappone}|c) &= 0 \\
\hat{P}(\text{Cinese}|\overline{c}) &= 1 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Giappone}|\overline{c}) &= 1 \\
\hat{P}(\text{Shanghai}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) &= 0
\end{aligned}
$$

Esempio

Classificazione

$$\begin{aligned}\hat{P}(c|d_5) &\propto 1 \cdot 0 \cdot 0 \cdot 1(-1/3) \cdot (1 - 1/3) = 0 \\ \hat{P}(\overline{c}|d_5) &\propto 1 \cdot 1 \cdot 1 \cdot (1 - 0) \cdot (1 - 0) = 1\end{aligned}$$

Classificato come ”non Cina”

Ma se avessi “Cinese Cinese Cinese Tokyo Macao”?

Smoothing

Pericolo: valori nulli. Ad esempio, se $n_{k1} = 0$ (nel training set non compare nessun elemento di C_1 con feature $x_k = 1$), per classificare un elemento con $x_k = 1$ abbiamo

$$\frac{1}{n_1^{d-1}} \prod_{i=1}^{d} n_{i1}^{x_i} (1 - n_{i1})^{1-x_i} = 0$$

e il modello non classificherà mai tale elemento in C_1.

Se al tempo stesso $n_{k2} = 0$ allora, per classificare lo stesso elemento,

$$\frac{1}{n_2^{d-1}} \prod_{i=1}^{d} n_{i2}^{x_i} (n_2 - n_{i2})^{1-x_i} = 0$$

Quindi la classificazione è impossibile.

Smoothing

In generale, sia $z \in \mathbb{R}^k$. Dato un training set $z_1, \ldots, z_n$, la probabilità $\phi_i = p(z = i)$ stimata con ML risulta

$$\phi_j = \frac{n_j}{n}$$

dove n_j è il numero di elementi $z = j$: se $n_j = 0$ allora $\phi_j = 0$.

Questo equivale ad escludere che la caratteristica $z = j$, non osservata nel training set, possa mai comparire.

Smoothing e classificatori binari

Laplace smoothing: assegna valori non nulli a probabilità stimate pari a 0.

Applicazione di smoothing:

$$\phi_j = \frac{n_j + 1}{n + k}$$

Anche in questo caso

$$\sum_{j=1}^{d} \phi_j = 1$$

$$\phi_{j1} = \frac{n_{j1} + 1}{n_1 + 2} \qquad \phi_{j2} = \frac{n_{j2} + 1}{n_2 + 2} \qquad \pi = \frac{n_1 + 1}{n + 2}$$

Esempio

	docID	termini	in c = Cina?
training set	1	Cinese Pechino Cinese	yes
	2	Cinese Cinese Shanghai	yes
	3	Cinese Macao	yes
	4	Tokyo Giappone Cinese	no
test set	5	Cinese Cinese Cinese Tokyo Giappone	?

Esempio: stima parametri

Priors: $P(c) = \pi = 3/4$ e $1 - \pi = 1/4$

Probabilità condizionate:

$$
\begin{aligned}
\hat{P}(\text{Cinese}|c) &= (3+1)/(3+2) = 4/5 \\
\hat{P}(\text{Shanghai}|c) = \hat{P}(\text{Macao}|c) &= (1+1)/(3+2) = 2/5 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Giappone}|c) &= (0+1)/(3+2) = 1/5 \\
\hat{P}(\text{Cinese}|\overline{c}) &= (1+1)/(1+2) = 2/3 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Giappone}|\overline{c}) &= (1+1)/(1+2) = 2/3 \\
\hat{P}(\text{Shanghai}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) &= (0+1)/(1+2) = 1/3
\end{aligned}
$$

Esempio

Classificazione

$$\begin{aligned}\hat{P}(c|d_5) &\propto 4/5 \cdot 1/5 \cdot 1/5 \cdot (1-2/5) \cdot (1-2/5) \approx 0.0012 \\ \hat{P}(\overline{c}|d_5) &\propto 2/3 \cdot 2/3 \cdot 2/3 \cdot (1-1/3) \cdot (1-1/3) \approx 0.13\end{aligned}$$

Esempio: estensione

Consideriamo il numero di occorrenze

Priors: $P(c) = \pi = 3/4$ e $1 - \pi = 1/4$

Probabilità condizionate:

$$\begin{aligned}
\hat{P}(\text{Cinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Giappone}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Cinese}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Giappone}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

I denominatori sono $(8+6)$ e $(3+6)$ in quanto il numero di occorrenze di termini in c e $\overline{c}$ è 8 e 3, rispettivamente, e in quanto ci sono 6 termini nel vocabolario

Esempio

Classificazione

$$\begin{aligned}\hat{P}(c|d_5) &\propto 4/5 \cdot 1/5 \cdot 1/5 \cdot (1-2/5) \cdot (1-2/5) \approx 0.0012 \\ \hat{P}(\overline{c}|d_5) &\propto 2/3 \cdot 2/3 \cdot 2/3 \cdot (1-1/3) \cdot (1-1/3) \approx 0.13\end{aligned}$$