Statistica e analisi dei dati

Classificatori naive Bayes

Kevin Muka

a.a 2024-2025

Classificatori naive Bayes

Un classificatore è un meccanismo che, dati degli oggetti su cui si vuole effettuare una distinzione, associa a ciascun oggetto una classe tra quelle disponibili. Nel nostro contesto, gli oggetti sono individui di una popolazione, e le classi indicano gruppi o categorie di interesse. Ad esempio, potremmo voler distinguere fra individui positivi o negativi a una certa condizione.

In particolare, un classificatore bayesiano si basa sul teorema di Bayes per "invertire" le probabilità condizionate: si parte da un insieme di proprietà osservate (ad esempio la presenza di una determinata caratteristica), e si stima la probabilità che l'individuo appartenga a una certa classe sulla base delle proprietà riscontrate.

Tuttavia, se si osservano più proprietà congiuntamente, la stima delle probabilità congiunte può risultare molto onerosa (in termini di numero di stime da calcolare). Per aggirare questa difficoltà, il classificatore naive Bayes assume l'ipotesi di indipendenza condizionale: conoscendo la classe, le diverse proprietà osservate sono approssimativamente indipendenti fra loro. Questa semplificazione, per quanto ingenua (da cui il termine naive), riduce drasticamente il numero di stime da effettuare, rendendo il classificatore computazionalmente efficiente e spesso sorprendentemente efficace in applicazioni reali.

Nei successivi paragrafi, si esaminerà il caso di una singola proprietà binaria e verrà mostrato come il teorema di Bayes possa essere impiegato per stimare la probabilità di appartenere a una classe. In seguito, si estenderà il discorso al caso in cui le proprietà osservate siano due, illustrando l'ipotesi di indipendenza condizionale, per poi presentare la costruzione completa del classificatore naive Bayes con più proprietà e classi.

Singola proprietà Si consideri una popolazione in cui, per ogni individuo i, è osservata una variabile binaria $x_i \in \mathcal{X} = \{0, 1\}$ che descrive la presenza di una certa caratteristica, e un'etichetta $y_i \in \mathcal{Y}$ che ne indica la classe di appartenenza. Per semplicità, si supponga che $\mathcal{Y} = \{0, 1\}$.

- Si osserva che il valore 1 indica la presenza della proprietà o della classe, e 0 invece l'assenza.

Sia N l'evento che un individuo scelto a caso presenti la caratteristica osservata, ossia $N = \{x{=}1\} = \{\omega \in \Omega \mid x(\omega) = 1\}$, e sia M l'evento che l'individuo appartenga alla classe presa in considerazione, ossia $M = \{y{=}1\} = \{\omega \in \Omega \mid y(\omega) = 1\}$. Applicando il teorema di Bayes si ha:

$$P(M|N) = \frac{P(N|M)\,P(M)}{P(N)}$$

Se a partire dal campione osservato fosse possibile ottenere una stima delle probabilità coinvolte nel teorema di Bayes, sarebbe immediato calcolare una stima di $P(M|N)$, che rappresenta la probabilità che un individuo con la caratteristica osservata, ossia $x = 1$, appartenga alla classe definita dall'evento M, ossia $y = 1$.

Se il valore stimato di $P(M|N)$ fosse sufficientemente alto, allora si potrebbe decidere di classificare gli individui con tale caratteristica come appartenenti alla classe $y = 1$. Al contrario, se il valore fosse particolarmente basso, si potrebbe decidere che la caratteristica osservata indichi l'appartenenza alla classe $y = 0$.

- La possibilità di azzardare una classificazione per un individuo potrebbe sembrare poco utile in un campione dove già si conoscono le etichette di classe. Tuttavia, questa capacità si rivela importante quando in questo campione vi sono valori mancanti, oppure quando in futuro si hanno a disposizione nuovi dati per i quali questa informazione non è disponibile.

Va notato che la scelta di quale evento sia associato alla caratteristica osservata e quale alla classe è arbitraria. In modo analogo, si potrebbe invertire la relazione tra gli eventi M e N senza alterare la validità dell'approccio.

Un modo semplice per stimare le probabilità nel teorema di Bayes è osservare la frequenza con cui si verificano gli eventi nel campione. In particolare:

- La probabilità $P(N|M)$ può essere approssimata dalla frequenza relativa con cui un individuo della classe $y = 1$ presenta la caratteristica osservata, ossia $x = 1$.
- La probabilità $P(M)$ può essere approssimata dalla frequenza relativa con cui un individuo appartiene alla classe $y = 1$.
- La probabilità $P(N)$ può essere approssimata dalla frequenza relativa con cui un individuo presenta la caratteristica osservata.

Queste frequenze tendono a convergere alle probabilità vere man mano che il numero di osservazioni nel campione cresce.

Due proprietà L'utilizzo di una singola proprietà non fornisce però un classificatore accurato. Ha senso quindi valutare l'uso di diverse proprietà osservate congiuntamente per ottenere un classificatore più preciso.

Sia $\{O{=}o_i\}$ l'evento "la caratteristica O assume valore o_i" dove $o_i \in \{o_1, \ldots, o_n\}$, e sia $\{C{=}c_j\}$ l'evento "la caratteristica C assume valore c_j" con $c_j \in \{c_1, \ldots, c_m\}$. L'evento congiunto $\{O = o_i\} \cap \{C = c_j\}$ indica che entrambe le caratteristiche assumono i rispettivi valori contemporaneamente.

- Si osserva che O e C sono variabili aleatorie, e che quelli definiti sopra sono i rispettivi eventi.
- *D'ora in avanti useremo la notazione abbreviata* $P(O{=}o_i)$ *per* $P(\{O{=}o_i\})$ *per semplicità. Inoltre si indicherà l'intersezione tra eventi* $P(O{=}o_i \cap C{=}c_j)$ *come* $P(O{=}o_i, C{=}c_j)$.

Il teorema di Bayes permette di scrivere la probabilità condizionata che un individuo appartenga alla classe definita dall'evento M (ossia $y = 1$) dato che presenta entrambe le caratteristiche osservate:

$$P(M|O{=}o_i, C{=}c_j) = \frac{P(O{=}o_i, C{=}c_j|M)\, P(M)}{P(O{=}o_i, C{=}c_j)}$$

Per stimare questa probabilità, è necessario calcolare tutte le probabilità condizionate e marginali coinvolte. Tuttavia, stimare la probabilità condizionata $P(O = o_i \cap C = c_j|M)$ risulta complesso poiché richiederebbe di considerare tutte le possibili coppie (o_i, c_j): si dovrebbero quindi fare $m \cdot n$ stime. I classificatori *naive Bayes* semplificano il problema facendo l'ipotesi che le caratteristiche osservate O e C siano indipendenti condizionatamente alla classe M, cioè che

$$P(O{=}o_i, C{=}c_j|M) \approx P(O{=}o_i|M)\, P(C{=}c_j|M)$$

Con questa semplificazione, nel numeratore della frazione si calcolano al più $m + n$ probabilità, al posto delle $m \cdot n$ necessarie senza l'ipotesi di indipendenza.

L'ipotesi di indipendenza condizionale implica che, conoscendo la classe M, la conoscenza del valore di O non fornisce informazioni aggiuntive su C e viceversa. Sebbene questo spesso non sia verificata nella pratica, semplifica enormemente il calcolo delle stime.

Si può quindi concludere che

$$P(M|O{=}o_i, C{=}c_j) \approx \frac{P(O{=}o_i|M)\, P(C{=}c_j|M)\, P(M)}{P(O{=}o_i, C{=}c_j)}$$

Al numeratore è necessario svolgere $n + m$ stime, mentre al denominatore bisogna farne comunque $n \cdot m$ dato che si tratta di una probabilità congiunta. Si osserva però che $P(M|O{=}o_i, C{=}c_j)$ è proporzionale al numeratore $P(O{=}o_i|M)\, P(C{=}c_j|M)\, P(M)$, e che il denominatore $\mathbb{P}(O{=}o_i, C{=}c_j)$ funge solo da costante di normalizzazione: esso non dipende dalla classe presa in considerazione, ma solo dal fatto che le proprietà osservate siano $O = o_i$ e $C = c_j$. Pertanto, quando si confrontano due classi diverse in un problema di classificazione, questo termine risulta identico e può essere trascurato ai fini della decisione finale, poiché non influenza l'ordine di grandezza della probabilità a posteriori per le diverse classi.

Caso generale Questa idea può essere estesa al caso in cui il numero di proprietà osservate non è limitato a due, ma può essere arbitrariamente grande.

Supponiamo ora di avere n proprietà osservate per ogni individuo, rappresentate da variabili aleatorie discrete $X_1, \ldots, X_n$, ciascuna in un proprio insieme di valori possibili. Analogamente, consideriamo m possibili classi $\{y_1, \ldots, y_m\}$ che possono descrivere l'oggetto in esame. Formalmente, ogni classe corrisponde a un evento $\{Y = y_k\}$ con $k \in \{1, \ldots, m\}$.

Per un individuo di cui abbiamo misurato $(x_1, \ldots, x_n)$ come realizzazioni di $X_1, \ldots, X_n$, vorremmo attribuirgli la classe $\{Y = y_k\}$ che risulta più "probabile" alla luce di tali proprietà. Il teorema di Bayes ci dice che:

$$\mathbb{P}\big(Y = y_k \mid X_1 = x_1,\ \ldots,\ X_n = x_n\big) \;=\; \frac{\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n \mid Y = y_k\big)\,\mathbb{P}(Y = y_k)}{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}$$

Per classificare l'individuo, dobbiamo scegliere la classe $\{Y = y_k\}$ che massimizza la probabilità a posteriori $\mathbb{P}(Y = y_k \mid X_1 = x_1, \ldots, X_n = x_n)$. Tuttavia, la stima diretta della probabilità congiunta $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid Y = y_k)$ può risultare molto onerosa, poiché richiede di considerare tutte le combinazioni dei valori $(x_1, \ldots, x_n)$.

Il classificatore naive Bayes semplifica tale stima assumendo che, condizionatamente alla classe $Y = y_k$, le variabili $X_1, \ldots, X_n$ siano approssimativamente indipendenti. In formule:

$$\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n \mid Y = y_k\big) \;\approx\; \prod_{i=1}^{n} \mathbb{P}\big(X_i = x_i \mid Y = y_k\big)$$

Sostituendo questa ipotesi nel teorema di Bayes, otteniamo:

$$\mathbb{P}\big(Y = y_k \mid X_1 = x_1,\ \ldots,\ X_n = x_n\big) \;\approx\; \frac{\mathbb{P}(Y = y_k) \displaystyle\prod_{i=1}^{n} \mathbb{P}\big(X_i = x_i \mid Y = y_k\big)}{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}$$

Il denominatore $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ non dipende dalla classe y_k ma solo dai valori osservati $(x_1, \ldots, x_n)$. Per la decisione di classificazione, cioè per confrontare le probabilità di classi diverse, esso funge da costante di normalizzazione, la stessa per ogni classe candidata. Di conseguenza, è sufficiente determinare la classe $\{Y = y_{k^*}\}$ che massimizza il prodotto:

$$\mathbb{P}(Y = y_k) \prod_{i\,=\,1}^{n} \mathbb{P}(X_i = x_i \mid Y = y_k)$$

In pratica, per classificare un individuo con proprietà $(x_1, \ldots, x_n)$, si calcola per ogni classe y_k il prodotto $\mathbb{P}(Y = y_k) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid Y = y_k)$ e si sceglie la classe che ne produce il valore più alto. In notazione compatta:

$$k^* = \arg \max_{k \in \{1,\ldots,m\}} \Big[\mathbb{P}(Y = y_k) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid Y = y_k) \Big]$$

Grazie all'ipotesi di indipendenza condizionale, il numero di stime richieste scende drasticamente rispetto a una modellazione pienamente congiunta, e questo rende il classificatore naive Bayes computazionalmente molto vantaggioso. Sebbene nella pratica le variabili X_i possano non essere completamente indipendenti all'interno di una stessa classe, l'approssimazione risulta spesso efficace.

Questo completa la costruzione del classificatore naive Bayes.