

JNOTES

Generative Multinomial Models

Modellazione di valori discreti

Quando trattiamo problemi caratterizzati da feature discrete, utilizzano una metodologia diversa rispetto alle variabili continue.

Analizziamo il caso di una singola variabile categaroria.

Una variabile categaroria (o feature discreta) è una variabile che può assumere un numero finito di valori distinti e non ordinati. Es: il colore del pelo di un gatto usata come feature categaroria per predire il genere dell'animale.



Quando abbiamo una variabile categaroria $x \in \{1, 2, \dots, m\}$ siamo dicono che x può assumere uno tra m valori distinti. A differenza delle variabili continue, questi valori non hanno un ordine intrinseco o misura di distanza naturale tra loro.

- $x \in \{\text{bianco, nero, arancione, grigio, bicolore, calico...}\}$
- Ogni valore rappresenta una categoria distinta.
- Non esiste un ordine naturale tra bianco e nero, o tra arancione e grigio.

Nel modello, consideriamo il dataset di addestramento $D = \{(x_1, c_1), (x_2, c_2), \dots, (x_n, c_n)\}$

Onde:

x_i è il colore dell'animale gatto nel campione

c_i è l'etichetta o l'etichetta corrispondente (maschio o femmina)

Assumiamo che i samples siano iid, ovvero:

il colore del pelo di un gatto non influenza il colore del pelo di un altro gatto nel dataset. Tutti i campioni provengono dalla stessa distribuzione di quantitativi sostanziale.

Vogliamo per ogni classe (maschio o femmina) stimare la probabilità di osservare ciascuno dei possibili valori della feature:

$$P(x=c|c=c) = \pi_{c,c}$$

Per c è la probabilità di osservare il valore x della feature, dato che il campione appartiene alla classe c.

Inoltre per ogni classe c, i campioni $x_1, \dots, (x_{n_c}, \dots, x_{m_c})$ devono soddisfare il vincolo: $\sum_{i=1}^{m_c} x_{ci} = 1$.

es: $P(x=\text{bianco}/\text{maschino}), P(\text{x=arancione}/\text{maschino})$

$P(x=\text{grigio}/\text{maschino}), P(x=\text{nero}/\text{maschino})$

Se ho solo uomo e donna, io sarebbe per dire P

$P(x=\text{maschio})$ e $P(x=\text{femmina})$. I cose che uso sono quelle

Tuttavia, per avere un solo AND (AND) perché il colore calico è nato (spesso nelle femmine)

Note: (con un numero elevato di possibili valori categoriali, potremo avere situazioni per cui non osserviamo esempi nel nostro training set. Questo può portare a modelli di stima $P=0$.

Per il calcolo di queste probabilità $P(x_i = x_i | C=c)$, possiamo adottare un approccio frequentista per stimare i parametri del modello. In particolare utilizzando la (maxim likelihood - ML) stima della massima verosimiglianza.

La likelihood per l'intero dataset considerando le i-esime celle (x_i, c_i) che si confrontano nei valori reali è pari a:

$$L(\Pi) = \prod_{i=1}^n f_{X_i|C_i}(\Pi)(x_i, c_i | \Pi) = \prod_{i=1}^n P(x_i = x_i | C_i = c_i) P(C_i = c_i)$$

Dove:

$\Pi = (\pi_1, \pi_2, \dots, \pi_K)$ rappresenta tutti i parametri del modello per le K classi

$c_i \in \{1, 2, \dots, h\}$ è la classe di cui l'esempio appartiene

$P(x_i = x_i | C_i = c_i) = \pi_{c_i, x_i}$ è la modalità di osservare il valore x_i condizionato alla classe c_i

$P(C_i = c_i)$ è un probabilità a priori della classe c_i .

Esempio: Se il nostro dataset è formato dai valori

- (black, male), (orange, male), (black, female), (orange, male),
- (white, male), (white, female), (white, male), (white, female),
- (black, female), (calico, female)

La likelihood è il prodotto

$$\begin{aligned} L(\Pi) &= P(X_1 = black | C_1 = male)P(C_1 = male) \times \\ &\quad P(X_2 = orange | C_2 = male)P(C_2 = male) \times \\ &\quad P(X_3 = black | C_3 = female)P(C_3 = female) \times \\ &\quad \vdots \\ &\quad P(X_{10} = calico | C_{10} = female)P(C_{10} = female) \end{aligned}$$

Assumendo di aver associato ogni colonna ad un valore intero $\{1, \dots, m\}$, e le spese nelle sono associate ai valori $\{0, 1\}$

Per semplificare i calcoli lavoriamo con la log likelihood

$$l(\Pi) = \sum_{i=1}^n \log P(x_i = x_i | C_i = c_i) + \epsilon \quad , \quad \{\text{immiscita, termici costanti derivante dalle probabilità a priori } P(c_i = w)\}$$

$$l(\Pi) = \sum_{i=1}^n \log \Pi_{c_i, x_i} + \epsilon$$

Calcolappiamo i termici per classe

$$l(\pi) = \sum_{c=2}^n \sum_{i|x_i=c} \log \pi_c x_i + \epsilon$$

CONSIDERANDO $l_c(\pi_c)$ È UNA COMPONENTE DELLA LOG-LIKELIHOOD CHE DIPENDE DA PARAMETRI DELLA CLASSE C

$$l(\pi) = \sum_{c=2}^n l_c(\pi_c) + \epsilon$$

$$l_c(\pi_c) = \sum_{i|x_i=c} \log \pi_c x_i$$

Possiamo quindi pensare di ottimizzare i parametri di ciascuna classe indipendentemente dalle altre, individuando il maximum-likelihood per ogni classe

$$l_c(\pi_c) = \sum_{i|x_i=c} \log \pi_c x_i = \sum_{j=1}^{N_{c,j}} \log \pi_c x_j$$

Dove

$N_{c,j}$ è il numero di volte in cui addiamo l'osservazione $x_i=j$ NEL dataset PER LA CLASSE C
RIPONDENDO L'ESEMPIO DEL DATASET ROSTRATO PRIMA

$$N_{\text{female,black}} = 2 \quad N_{\text{female,orange}} = 0 \quad N_{\text{female,white}} = 2 \quad N_{\text{female,calico}} = 1$$

$$N_{\text{male,black}} = 1 \quad N_{\text{male,orange}} = 2 \quad N_{\text{male,white}} = 2 \quad N_{\text{male,calico}} = 0$$

Dopo aver espresso la log-likelihood in termini di frequenza, dobbiamo trovare i parametri che la massimizzano. Questo è più complesso rispetto al caso gaussiano poiché dobbiamo considerare un vincolo importante, i parametri π_c , devono fornire una distribuzione di probabilità valida, quindi la loro somma deve essere 1.

Per scomporre il tutto, eliminando temporaneamente il profilo (relativo alla classe) e risolvendo il problema per una classe ciascuna:

$$\text{Vogliamo quindi massimizzare: } f(\pi) = \sum_{j=1}^{N_c} N_j \log \pi_j$$

$$\text{Considerando il vincolo } \sum_{j=1}^{N_c} \pi_j = 1$$

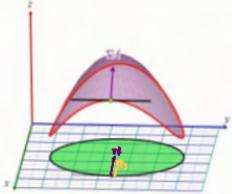
π_j è la probabilità di osservare il valore j nella classe c considerata

N_j è il numero di volte che abbiamo osservato il valore j; IN QUESTA CLASSE m è il numero totale di possibili valori discetti che la feature può assumere

Per trovare il massimo si usa la tecnica dei moltiplicazioni di fattori.

DIFESIONE SUI MOLTIPLICATORI DI FATTORI

Questa tecnica ci permette di trovare il massimo di una funzione sulla curva che definisce il nucleo della funzione stessa. (nel caso di una funzione crescente monotona, corrisponde al massimo assoluto)



DEFINISCI UNO SCENARIO SENZIETTO IN 2 DIMENSIONI, SAPPIONO DI VOLTA TROVARE IL PUNTO PIÙ ALTO DI UNA COLLINA MA POSSIMO MUOVERSI SOLO LUNGO UN SENTIERO SPECIFICO (VINCULO). LA "COLLINA" RAPPRESENTA LA NOSTRA FUNZIONE DI log-likelihood. IL "SENTIERO" IL VINCULO (LA SOMMA DELLE PROBABILITÀ DEVE ESSERE 1). IL "PUNTO PIÙ ALTO DEL SENTIERO" È IL MASSIMO VINCULATO CHE STIAMO CERCANDO.

IL METODO DI MOLTIPLICAZIONE DI VALORI SI BASE SULL' OSSERVAZIONE CHE AL PUNTO DI MASSIMA VNL, IL GRADIENTE DELLA FUNZIONE OBIETTIVO (LA DIREZIONE DI MASSIMA CRESCITA DELLA COLLINA) DEVE ESSERE PARALELO AL VINCULO (SENTIERO), E QUINDI // AL GRADIENTE DEL VINCULO. SE NON FOSSE COSÌ POTREMO ANCHE MUOVERMI LUNGO IL SENTIERO IN UNA DIREZIONE CHE AVRIA L'ALTITUDINE ↗ ↘.

MATHEMATICAMENTE: $f(\nu) \underset{\text{VINCULO}}{\rightarrow} g(x)$

$$\nabla f(x) = \lambda \nabla g(x) \quad \lambda \text{ È DETTO AGGRIMENTO DI BALANZA, NEL MASSIMO VNL HA PIAVA IMPORTANZA}$$

DEFINISCI LA FUNZIONE DI BALANZA

$$L(\pi, \lambda) = f(\pi) - \lambda \left(\sum_{j=1}^m \pi_j - 1 \right) = \sum_{j=1}^m N_j \log \pi_j - \lambda \left(\sum_{j=1}^m \pi_j - 1 \right)$$

$$\begin{cases} \frac{\partial L}{\partial \pi_i} = 0 & \forall i = 1, \dots, m \\ \frac{\partial L}{\partial \lambda} = 0 \end{cases}$$

$$\left. \begin{aligned} \frac{\partial L}{\partial \pi_i} &= \frac{N_i}{\pi_i} - \lambda = 0 \Rightarrow \pi_i = \frac{N_i}{\lambda} \\ \frac{\partial L}{\partial \lambda} &= 1 - \sum_{j=1}^m \pi_j = 0 \end{aligned} \right\} \Rightarrow \sum_{j=1}^m \frac{N_j}{\lambda} = 1 \Rightarrow \lambda = \sum_{j=1}^m N_j = N \text{ OVVERO IL NUMERO DI ELEMENTI PER LA CLASSE CONSIDERATA.}$$

QUINDI $\pi_i = \frac{N_i}{N}$ CHE È LA FREQUENZA RELATIVA CON CUI ABBIANO OSSERVATO IL VALORE NEL DATASET

Per massimizzare la probabilità di osservare i dati che abbiano effettivamente osservato (osservazione di verosimilitudine) la distribuzione di probabilità ottimale deve rispettare esattamente la frequenza dei valori nel dataset.

Quindi la soluzione per ogni classe è:

$$\pi_{c,i} = \frac{N_{c,i}}{N_c}$$

$$P(X_c = x_c | c = c) = \pi_{c,i}^{x_c}$$

In our example, we would have

$$N_{male} = 5 \quad N_{female} = 5$$

thus

$$\begin{aligned} \pi_{female, black} &= 0.4 & \pi_{female, orange} &= 0 & \pi_{female, white} &= 0.4 & \pi_{female, zebra} &= 0.2 \\ \pi_{male, black} &= 0.2 & \pi_{male, orange} &= 0.4 & \pi_{male, white} &= 0.4 & \pi_{male, zebra} &= 0 \end{aligned}$$

Più Attributi Categorici

Se addiamo più di un attributo categorico, teoricamente potremo modellare la loro probabilità congiunta, come una singola variabile casuale categorica. I valori di questa variabile saranno le diverse combinazioni dei valori degli attributi individuali.

ES:

Coupe del Peugeot: {Punto, Nero Avantour, Clio} → valori

Luminosità del LED: {alto, medio, bajo} → valori

Etc: {cuchillo, tenedor, cuchara} → valori

Per ogni classe dovremo stimare $(3 \times 3) = 36$ valori diversi

L'apprendo diventa immenso a vista!

ESPANSIONE CONJUNTIVA (MOLTE COMBINAZIONI)

Dati insufficienti: (tutte le combinazioni hanno $P=0$ non appaiono come nel training set)

Complessive combinazioni: molti parametri da calcolare.

Per risolvere questo problema, possiamo addurre l'approssimazione Naïve Bayes, che assume che gli attributi siano condizionalmente indipendenti, data la classe.

Cioè, assuniamo che:

$$P(X_1, X_2, \dots, X_m | C) = P(X_1|C) \times P(X_2|C) \times \dots \times P(X_m|C)$$

Dove X_i rappresenta il i -esimo attributo, C la classe

(Questi ipotesi sono solitamente una semplificazione forte, gli attributi normalmente sono completamente indipendenti tra loro)

Le cifre a Naïve Bayes, possono stimare i parametri per ciascun attributo separatamente e poi combinarli.

A attributo; e classe, calcoliamo

$$\pi_{C,i,j} = \frac{N_{C,i,j}}{N_{C,j}}, \quad N_{C,i,j} \text{ numero di volte che l'attributo } i \text{ assume il valore } j \text{ nella classe } C \\ N_{C,j} \text{ è il numero totale di campioni della classe } C \text{ per cui } i \text{ è definito l'attributo } j$$

(stessi calcoli nel caso con singolo attributo)

La probabilità predittiva di un nuovo samle con attributi x_1, x_2, \dots, x_m data la classe C è quindi

$$P(X = (x_1, x_2, \dots, x_m) | C = c) = \prod_{j=1}^m \pi_{C,j, x_{C,j}}$$

Probabilità che l'attributo j assuma il valore x_j nella classe C

Modelli multinomiali per il Conteggio di Eventi: Applicazioni al Text Mining

Fino ad ora abbiamo considerato il caso in cui ogni campione ha un singolo valore categorico, (ad esempio dimostrano un solo colore). Ora passiamo ad una situazione diversa dove:

• Addiamo conteggi di eventi più diffusi (che soprattutto categoriche).

2. Ogni campione può alesicare multiple occorrenze dello stesso evento

Per esempio un documento di testo può contenere più occorrenze delle stesse parole, e la distribuzione di queste parole può dipendere dal tema ("topic") del documento.

Modello Bag-of-Words

In questo modello

- Immaginiamo l'insieme delle parole
- Consideriamo solo quante volte una parola appare
- rappresentiamo un documento come un vettore di contagi: $x = (x_{c,1}, x_{c,2}, \dots, x_{c,n})$ dove $x_{c,j}$ rappresenta il numero di volte che la parola j appare nel documento.

Esempio: consideriamo un vocabolario di sole 3 parole: "latte", "cane" "animale". Il documento: "Il latte è un animale, così come il cane. Il latte è più indipendente del cane" potrebbe essere rappresentato come:
 $x = (3, 2, 1) \rightarrow$ 3 occorrenze di latte, 2 di cane + 1 di animale.

Per modellare questo tipo di dati usano la **distribuzione multinomiale**, che è una generalizzazione della distribuzione binomiale, a più di 2 categorie. La distribuzione multinomiale, descrive le probabilità di osservare un certo numero di occorrenze, per ciascuna categoria quando eseguiamo un certo numero di prove indipendenti.

Per ogni classe c ($\text{topic } c$), abbiamo un insieme di parafrazai $T_c = (T_{c,1}, T_{c,2}, \dots, T_{c,m})$ dove:

- $T_{c,j}$ rappresenta la probabilità di osservare la parola j in un documento di classe c .
- la somma di tutti i $T_{c,j}$ per una classe c è 1: $\sum_j T_{c,j} = 1$.

$$P(X=x | c=c) = \frac{\left(\sum_{j=1}^m x_{c,j}\right)!}{\prod_{j=1}^m x_{c,j}!} \prod_{j=1}^m \frac{x_{c,j}}{T_{c,j}} \approx \prod_{j=1}^m \frac{x_{c,j}}{T_{w,j}}$$

Dove $\left(\sum_{j=1}^m x_{c,j}\right)!$ è il coefficiente multinomiale che rappresenta il numero di modi in cui possiamo ottenere esattamente $x_{c,j}$ occorrenze di ogni parola; dato un totale di $\sum_{j=1}^m x_{c,j}$ parole, così diversi di ordinare le parole, mantenendo lo stesso numero di occorrenze di ciascuna.

$$\prod_{j=1}^m \frac{x_{c,j}}{T_{w,j}}$$
 probabilità di osservare esattamente $x_{c,j}$ occorrenze di ciascuna parola;

Immaginiamo di avere solo tre parole nel nostro vocabolario: "gatto", "cane", "animale". Supponiamo che per i documenti di classe "felini" abbiamo:

$\pi_{\text{felini}, \text{gatto}} = 0,5$ (probabilità della parola "gatto")

$\pi_{\text{felini}, \text{cane}} = 0,1$ (probabilità della parola "cane")

$\pi_{\text{felini}, \text{animale}} = 0,4$ (probabilità della parola "animale")

Per un documento con conteggi $x = (3, 1, 2)$ (3 "gatto", 1 "cane", 2 "animale"), la probabilità multinomiale (escludendo il coefficiente multinomiale) sarebbe:

$$P \propto (0,5)^3 \times (0,1)^1 \times (0,4)^2 = 0,125 \times 0,1 \times 0,16 = 0,002$$

VOLLIANO ADORSO STIMARE I PARAMETRI $\pi_{c,j}$ CHE MINIMIZZANO LA LIKELIHOOD DEI DATI OSSERVATI. LA LIKELIHOOD PUÒ ESSERE SCRITA COME:

$$\begin{aligned} l(\pi) &= \sum_c l_c(\pi_c) + \sum_j \text{penalizzazione} \\ l_c(\pi_c) &= \sum_{i \in c} \sum_{j \in c} x_{i,c,j} \log \pi_{c,j} \end{aligned}$$

[ESTIMANTE COME IL VISO PREFERITO È PUSSONE, OMOFORE IL TUTTO PER CLASSE (ADDANNO TUTTO IL COEFFICIENTE MULTINOMIALE PROVA A COSTARE E NON CI INTERESSA)]

VIREMO FATTI UNA DOCUMENTI

SOMMA ESTERNA SU TUTTI I VARIANTI I CHE APPARTENGONO ALLA CLASSE C

SOMMA INTERNA SU TUTTE LE M PRESENZE DELLA VOCALOGNIO (VOCALOGNIO DA UN DOCUMENTO E PAROLE MA $X_{i,c,j}$ È IL CONTEGGIO DELLA PAROLA J NEL VARIANTE I)

$$l_c(\pi_c) = \sum_{i \in c} \sum_{j \in c} x_{i,c,j} \log \pi_{c,j} = \sum_{j \in c} N_{c,j} \log \pi_{c,j}$$

Dove $N_{c,j}$ è il conteggio totale della parola j in tutti i documenti di classe c

$$N_{c,j} = \sum_{i \in c} x_{i,c,j}$$

LA SOLUZIONE UNISCONDO SEMPRE I MULTIPLOS DI VALORIE, È EFFETTIVAMENTE ULTRAVALE A PIENA VOLONTÀ VEDENDO CHE $\sum_j N_{c,j} \log \pi_{c,j}$ È UNO OTTELONICO

CONSIDERANDO IL VINCULO $\sum_j \pi_{c,j} = 1$, DOPO TUTTO È UN COEFFICIENTE IN CASO GENERICO (≥ 0)

$$\therefore \pi_{c,j} = \frac{N_{c,j}}{N_c}$$

DOVE:

$N_{c,j}$ È IL CONTEGGIO TOTALE DELLA PAROLA J NEL DOCUMENTO DI CLASSE C

$$N_c = \sum_j N_{c,j} È IL CONTEGGIO TOTALE DI TUTTE LE PAROLE NEI DOCUMENTI DI CLASSE C$$

IL PROBABILITÀ OTTIMALE PER LA PAROLA J NELLA CLASSE C È LA MEDIANA RELATIVA CON CUI LA PAROLA J APPARE NEI DOCUMENTI DI CLASSE C.

SE CONSIDERIAMO UN PROBLEMA DI CLASSIFICAZIONE BINARIA, IL LOG-LIKELIHOOD RATIO (LR)

$$LR(x) = \log \frac{P(X=x|L_{\text{true}})}{P(X=x|L_{\text{false}})} = \sum_{j=1}^n x_{0j} \log \pi_{0j} - \sum_{j=1}^n x_{1j} \log \pi_{1j} = x^T b$$

Dove:

x è la vettore dei controlli, che ospitano nel campo di test

le più bassi (L_{true})

o se (L_{false})

b è un vettore di coefficienti = $(\log \pi_{01} - \log \pi_{11}, \dots, \log \pi_{0n} - \log \pi_{1n})$

Il rapporto di verosimiglianza ha una forma di una funzione univara, divenendo una funzione lineare del vettore dei controlli x . Ciò significa che la superficie di decisione fra le 2 classi sarà un iperplano (n-dimensione si ha considerando la prior probabilità delle 2 classi che dipende dal dataset di apprendimento)

(SULLE STESE VENE FATTO UN ESEMPIO, M QUI LI CONFERMANO SULLA TEORIA)

Un esempio pratico

Consideriamo due classi di documenti: "tecnologia" e "sport". Supponiamo di avere un vocabolario di 5 parole: "computer", "calcio", "dati", "squadra", "internet". Per i documenti di classe "tecnologia", contiamo:

"computer": 50 occorrenze
"calcio": 5 occorrenze
"dati": 40 occorrenze
"squadra": 10 occorrenze
"internet": 45 occorrenze
Totale parole: 150

Per i documenti di classe "sport", contiamo:

"computer": 8 occorrenze
"calcio": 60 occorrenze
"dati": 15 occorrenze
"squadra": 50 occorrenze
"internet": 7 occorrenze
Totale parole: 160

Le stime di massima verosimiglianza sarebbero:

Per "tecnologia":

tecnologia.computer = 50/150 = 0.333
tecnologia.calcio = 5/150 = 0.033
tecnologia.dati = 40/150 = 0.267
tecnologia.squadra = 10/150 = 0.067
tecnologia.internet = 45/150 = 0.300

Per "sport":

sport.computer = 8/160 = 0.057
sport.calcio = 60/160 = 0.429
sport.dati = 15/160 = 0.107
sport.squadra = 50/160 = 0.357
sport.internet = 7/160 = 0.050

Queste probabilità riflettono bene l'intuizione che parole come "computer" e "internet" sono più probabili in documenti di tecnologia, mentre parole come "calcio" e "squadra" sono più probabili in documenti di sport.

Classificazione di un nuovo documento

Per classificare un nuovo documento, calcoliamo la probabilità che appartenga a ciascuna classe utilizzando la formula multinomiale e la regola di Bayes. Supponiamo di avere un nuovo documento con conteggio

"computer": 3 occorrenze
"calcio": 0 occorrenze
"dati": 2 occorrenze
"squadra": 1 occorrenza
"internet": 4 occorrenze

Calcoliamo (escludendo il coefficiente multinomiale e assumendo probabilità a priori uguali per le classi):

$$P(X = L_{\text{true}}) \propto (0.333)^3 \times (0.429)^0 \times (0.107)^2 \times (0.357)^1 \times (0.050)^4 \approx 7.96 \times 10^{-6}$$

$$P(X = L_{\text{false}}) \propto (0.057)^3 \times (0.429)^0 \times (0.107)^2 \times (0.357)^1 \times (0.050)^4 \approx 4.84 \times 10^{-6}$$

Il documento ha una probabilità molto maggiore di appartenere alla classe "tecnologia", il che è coerente con il contenuto che presenta molte parole legate alla tecnologia

Abbiamo visto fino ad ora 2 modelli diversi, per modellare un dataset di eventi catenari, in realtà questi metodi sono matematicamente equivalenti.

① Modello basato su variabili casuarine indipendenti

In questo approccio modelliamo il dataset come un insieme di n variabili casuarine (catenarie, indipendenti)

$$X_i \in \{1, \dots, m\}$$

X_i è un token, elemento singolo del dataset.

La distribuzione è descritta da un vettore di probabilità π

La probabilità che un token abbia il valore i è $P(X_i = i) = \pi_i$.

Per esempio X_i potrebbe rappresentare una svolta verso la i -ma posizione specifica del documento, T_i , la probabilità di osservare la parola j in qualsiasi posizione.

② Modello basato su contelli di eventi

In questo caso modelliamo il dataset attraverso un vettore casuale di contelli degli eventi $Y = \{Y_1, \dots, Y_m\}$ dove:

Ogni componente Y_j è una variabile casuale che rappresenta il numero di occorrenze dell'evento j nel dataset.

La distribuzione è ancora descritta da un vettore di probabilità π , che rappresenta le probabilità dei singoli eventi.

Ovviamente in entrambi i casi il tutto sarà poi condizionato dalla classe.

La relazione tra i due modelli è data da

$$Y_j = \sum_{i=1}^n I[X_i = j]$$

Dove $I[X_i = j]$ è una funzione indicatrice che vale 1 se $X_i = j$ e 0 altrimenti. In altre parole, Y_j conta quante volte l'evento j si verifica nel dataset (è effettivamente così).

Definizione

① Stima di Massima Verosimiglianza

In entrambi i casi, la soluzione di Maximum Likelihood per π corrisponde alle frequenze relative di occorrenze. Se confrontiamo le log-likelihood, sono identiche:

Per il modello basato su variabili casuarine indipendenti

$$L_X(\pi) = \prod_{i=1}^n \log \pi_{X_i} = \prod_{j=1}^m N_j \log \pi_j \quad (\text{perché le parole sono ormai tutte presenti})$$

N_j è il numero di volte che il valore j appare nel dataset.

Per il modello basato sui contelli, la log-likelihood è

$$L_Y(\pi) = \log \frac{(\prod_{j=1}^m Y_j)!}{\prod_{j=1}^m N_j!} + \sum_{j=1}^m Y_j \log \pi_j = \sum_{j=1}^m N_j \log \pi_j$$

Dove π è la costante (che non dipende da π).

Notiamo che le 2 funzioni sono identiche (no cambia π è inutile), di conseguenza i

VALORI DI TI CHE MASSIMIZZANO UNA, MASSIMIZZANO ANCHE L'ALTRA.

LA PROBABILITÀ BAYESIANA, SIA NELLA STIMA, LA CLASSIFICAZIONE PONTE ALI STESSI RISULTATI SE APPUNTO SULI STESSI PAMPANTI, QUINDI LA POSTERIOR PROBABILITY CONSIDERANDO STIMA LIKELIHOOD + STIMA PAMPANTISI NELLA STIMA.

Consideriamo, un campione di test LE VEROSIMILANZE CONDIZIONATE ALLE CLASSI SONO PROPORZIONALI TRA I DUE COEFFICIENTI MULTINOMIALE - COSTANTE - NEI MAPPANTI SI SEMPRE

$$f_{X|C}(x_t|c) = f(y_t) \cdot f_{Y|C}(y_t|c)$$

DIPENDONO SOLO DAI MAPPANTI
DEI TEST

$$\text{La LLR} \Rightarrow \log \frac{f_{X|C}(x_t|h_0)}{f_{X|C}(x_t|h_1)} = \log \frac{f_{Y|C}(y_t|h_0)}{f_{Y|C}(y_t|h_1)}$$

$$\text{la probabilità a posteriori a sua volta } P(C_0=c|X=x) = \frac{f_{X|C}(x_t|c)P(c)}{\sum_c f_{X|C}(x_t|c)P(c)} = \frac{f_{Y|C}(y_t|c)P(c)}{\sum_c f_{Y|C}(y_t|c)P(c)} = P(C_0=c|Y=y)$$

QUINDI INDEPENDENTEMENTE DAL AUDITO SCELTO, LA CLASSIFICAZIONE DI NUOVI CAMPIONI SONO IDENTICHE.

Un Esempio Concreto per Consolidare la Comprensione

Immaginiamo di avere un piccolo corpus di tre documenti con un vocabolario di solo quattro parole (a, b, c, d):

Documento 1: "a b a d"
Documento 2: "b c d a"
Documento 3: "a a b d"

Rappresentazione con il modello X

Nel modello XX

K rappresentiamo ogni posizione come una variabile casuale categorica:

X1-X2-X3-X4-X5-X6-X7-X8-X9-X10-X11-X12

X1-X2-...-X12 rappresentano le 12 parole nei 12 documenti

Per esempio: X1=a

X1=a, X2=b, X3=b

X3=b, ... X10=d X11= (12) = d

X12=d

La log-verosimiglianza cambierà

\log(multinomialog(multilog(a)+multilog(b)+multilog(c)+multilog(d))) - \ell(X(pi)) = \log(pi_a) + \log(pi_b) + \log(pi_c) + \log(pi_d) + \ell(pi)

\ell(multilog(a)+multilog(b)+multilog(c)+multilog(d))

Raggruppando i termini:

\ell(multilog(a)+multilog(b)+multilog(c)+multilog(d)) = 5\log(pi_a) + 3\log(pi_b) + 3\log(pi_c) + 3\log(pi_d)

\ell(multilog(a)+multilog(b)+multilog(c)+multilog(d))

Rappresentazione con il modello Y

Nel modello YY

Y rappresentiamo direttamente i contagi:

Y=(5, 3, 1, 3)

Y=(5, 3, 1, 3) indicando 5 occorrenze di "a", 3 di "b", 1 di "c" e 3 di "d"

La log-verosimiglianza sarebbe:

\ell(multilog(5, 3, 1, 3) + multilog(multilog(pi_a)+multilog(pi_b)+multilog(pi_c)+multilog(pi_d))) - \ell(X(pi)) = \log(binom(12)(5, 3, 1, 3)) + 5\log(pi_a) + 3\log(pi_b) + 3\log(pi_c) + 3\log(pi_d)

\ell(multilog(5, 3, 1, 3) + multilog(multilog(pi_a)+multilog(pi_b)+multilog(pi_c)+multilog(pi_d)))

Il primo termine è lo logaritmo del coefficiente multinomiale che è costante rispetto a pi(pi)

Il resto della formula è identico a \ell(X(pi)) - \ell(X(pi))

X(pi)

Stima di Massima Verosimiglianza

In entrambi i casi la stima di massima verosimiglianza sarebbe:

$$pi_a = 12/40 = 0.47 \quad pi_a = 5/12 = 0.417$$

$$pi_b = 12/40 = 0.47$$

$$pi_c = 3/12 = 0.25 \quad pi_c = 3/12 = 0.25$$

$$pi_d = 12/40 = 0.483 \quad pi_d = 3/12 = 0.083$$

$$pi_d = 12/40 = 0.483$$

$$pi_d = 3/12 = 0.25 \quad pi_d = 3/12 = 0.25$$

$$pi_d = 3/12 = 0.25$$

