

Naív Bayes-osztályozó

Csima Judit

BME, VIK,
Számítástudományi és Információelméleti Tanszék

2018. március 8. és 12.

- az attribútumoknak valószínűségi változókat feleltetünk meg
- az osztályattribútum diszkrét, a többi attribútum lehet folytonos vagy diszkrét valószínűségi változó
- az osztályattribútum értékét a megfelelő valváltozó többi valváltozóra vett feltételes eloszlása alapján becsüljük
- azaz $P(C | A_1, A_2, \dots, A_n)$ típusú feltételes valószínűségeket akarunk kiszámolni a training set alapján
- egy a_1, a_2, \dots, a_n érték n-eshez a predikció során azt a c_j címkét választjuk majd, amire $P(C = c_j | A_1 = a_1, A_2 = a_2, \dots, A_n = a_n)$ maximális

Szükséges fogalmak

- feltételes valószínűség: $P(X | Y) = \frac{P(X, Y)}{P(Y)}$
- Bayes-tétel: $P(X | Y) = \frac{P(Y | X)P(X)}{P(Y)}$
- X szerepét C játssza most, Y pedig a többi attribútumból álló összetett valváltozó lesz

Example of Bayes Theorem

□ Given:

- A doctor knows that meningitis causes stiff neck 50% of the time
- Prior probability of any patient having meningitis is 1/50,000
- Prior probability of any patient having stiff neck is 1/20

□ If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M | S) = \frac{P(S | M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayes tétel az osztályozásnál

- most $P(C | A_1, A_2, \dots, A_n)$ -ra lenne szükségünk
- ezt $\frac{P(A_1, A_2, \dots, A_n | C)P(C)}{P(A_1, A_2, \dots, A_n)}$ alakban tudjuk kiszámolni
- keressük azt a c_j címkét, amire a $\frac{P(A_1, A_2, \dots, A_n | C = c_j)P(C = c_j)}{P(A_1, A_2, \dots, A_n)}$ tört maximális
- mivel minden egyes $C = c_j$ esetben ugyanaz a nevező, ezért igazából az a kérdés, hogy számláló hol maximális
- ehhez kéne tudni a $P(A_1, A_2, \dots, A_n | C = c_j)$ és $P(C = c_j)$ értékeket

$P(A_1, A_2, \dots, A_n | C = c_j)$ és $P(C = c_j)$ kiszámolása

- $P(C = c_j) = \frac{n_j}{n} = c_j$ címkéjű sorok száma osztva az összes sor számával
- az A_1, A_2, \dots, A_n valváltozókról feltesszük, hogy feltételelesen függetlenek, ha C értéke adott
- azaz $P(A_1, A_2, \dots, A_n | C = c_j) = P(A_1 | C = c_j)P(A_2 | C = c_j) \dots P(A_n | C = c_j)$
- ezek után már csak $P(A_i = a_i | C = c_j)$ a kérdés minden i, j párra

$P(A_i = a_i | C = c_j)$ meghatározása

- ha A_i diszkrét valváltozó:

$P(A_i = a_i | C = c_j) = \frac{n_{ij}}{n_j}$ = a_i és c_j értéket felvevő sorok száma osztva az összes c_j címkéjű sor számával

- ha A_i folytonos valváltozó:

- feltételezzük, hogy normális eloszlású

- $P(A_i = a_i | C = c_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(a_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$

- kérdés σ_{ij} és μ_{ij} értéke

- ezeket közelítsük a training set alapján: mintából számolt átlag és szórás

- kérdés σ_{ij} és μ_{ij} értéke
- ezeket közelítsük a training set alapján: mintából számolt átlag és szórás
- μ_{ij} = az A_i oszlopbeli értékek átlaga azon sorokat nézve csak, ahol a c_j címke van
- σ_{ij} = a c_j címkéjű sorokban az A_i attribútumértékek szórása

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

□ Class: $P(C) = N_C/N$

– e.g., $P(\text{No}) = 7/10$,
 $P(\text{Yes}) = 3/10$

□ For discrete attributes:

$$P(A_i | C_k) = |A_{ik}| / N_C$$

– where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k

– Examples:

$$P(\text{Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes})=0$$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(A_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- One for each (A_i, c_j) pair

- For (Income, Class=No):

- If Class=No

- ◆ sample mean = 110

- ◆ sample variance = 2975

$$P(\text{Income} = 120 | \text{No}) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

- ha már minden feltételes valószínűséget kiszámoltunk
- egy új sor osztályzásakor az A_i attribútumok a_i értékei alapján
- minden c_j címkére
 $P(A_1 | C = c_j)P(A_2 | C = c_j) \dots P(A_n | C = c_j)P(C = c_j)$
kiszámolása
- az lesz a jósolt címke, amelyik c_j -re ez maximális

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$P(A|M)P(M) > P(A|N)P(N)$

=> Mammals

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No: sample mean=110

sample variance=2975

If class=Yes: sample mean=90

sample variance=25

$$\begin{aligned} P(X|\text{Class}=\text{No}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{No}) \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024 \end{aligned}$$

$$\begin{aligned} P(X|\text{Class}=\text{Yes}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{Yes}) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} = 0 \end{aligned}$$

Since $P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

Therefore $P(\text{No}|X) > P(\text{Yes}|X)$

$\Rightarrow \text{Class} = \text{No}$

Mi van, ha a feltételes valószínűség 0?

- előfordulhat, hogy valami i, j esetén $P(A_i | C = c_j)$ nulla, mert nincs ilyen teszt sor
- ekkor hiába tűnik a többi a_i alapján nagy esélyesnek egy c_j címke, biztosan nem választjuk
- megoldás, hogy máshogy becsüljük $P(A_i | C = c_j)$, mint eddig:
 - Laplace: $P(A_i | C = c_j) = \frac{n_{ij} + 1}{n_j + c_i}$, ahol c_i az A_i attribútum lehetséges értékeinek száma
 - ezzel a becsléssel sose kapok 0-t

- tanítási fázisan megbecslem a feltételes valószínűségeket
 - relatív gyakoriságok a training setben
 - Laplace-becslés verzióban ugyanez
 - folytonos változónál a normális eloszlás paraméterezése
- predikciókor az így kiszámolt feltételes valószínűségek segítségével megkeresem a legvalószínűbb címkét

R-ben mi van?

- pl. e1071 package
- `> m <- naiveBayes(Species ~ ., data = iris)`
- `> table(predict(m, iris), iris[,5])`

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	3	47