Esercitazione 8 27/11/2023

Naive bayes

Assume target function $f : X \to V$, where each instance x described by attributes $\langle a_1, a_2 \dots a_n \rangle$. Most probable value of $f(x)$ is:

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \dots a_n)$$

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)}$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \dots a_n | v_j) P(v_j)$$

Need a lot of data to estimate these many numbers!

How can we deal with this?

Answer: Make independence assumptions

$X_1, \dots, X_m$ are **conditionally independent** given Y.

Naive Bayes assumption:

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

Naive Bayes assumption

$X_1, \dots, X_m$ are **conditionally independent** given Y.

CONDITIONALLY INDIPENDENT?

Focus: Conditional independence

Definition: X is *conditionally independent* of Y given Z if the probability distribution governing X is independent of the value of Y given the value of Z; that is, if

$(\forall x_i, y_j, z_k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_l)$

more compactly, we write

$$P(X|Y,Z) = P(X|Z)$$

- Or equivalently

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

Focus: Conditional independence

$$P(X|Y,Z) = P(X|Z)$$

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

- **e.g.,** $P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$

  Note: does NOT mean Thunder is independent of Rain

Esempio

Example: Naive Bayes

- Complete the Table 1 (for x_6) using Map hypothesis (i.e., insert the corresponding label for T).
- In this problem, likelihood and prior have to be estimated using observations $\{x_1, x_2, \ldots, x_5\}$.
- Moreover assume $A, B, C, T \in \{0, 1\}$.

Table 1: Training

Esempi	A	B	C	T
x_1	1	1	1	0
x_2	0	1	1	0
x_3	1	0	1	1
x_4	0	0	0	1
x_5	0	1	0	1
x_6	1	1	0	

Es 4: «Complete the table»

- In this problem, likelihood and prior have to be estimated using observations $\{x_1, x_2, \ldots, x_5\}$.

Table 1: Training

Esempi	A	B	C	T
x_1	1	1	1	0
x_2	0	1	1	0
x_3	1	0	1	1
x_4	0	0	0	1
x_5	0	1	0	1
x_6	1	1	0	

For the first hypothesis, i.e., $h_0 = (T = 0)$ we have

$$P(T = 0|A = 1, B = 1, C = 0) = \frac{P(A = 1, B = 1, C = 0|T = 0)P(T = 0)}{P(x_6)}$$
$$\propto P(A = 1, B = 1, C = 0|T = 0)P(T = 0)$$
$$\propto P(A = 1|T = 0)P(B = 1|T = 0)P(C = 0|T = 0)P(T = 0)$$
$$\propto 1/2 * 1 * 0 * 2/5$$

Then we have $P(h_0|x_6) \propto 1/2 * 1 * 0 * 2/5 = 0$

For the second hypothesis $h_1 = [T = 1]$ we have

$$P(T = 1|A = 1, B = 1, C = 0) = \frac{P(A = 1, B = 1, C = 0|T = 1)P(T = 1)}{P(x_6)}$$
$$\propto P(A = 1, B = 1, C = 0|T = 1)P(T = 1)$$
$$\propto P(A = 1|T = 1)P(B = 1|T = 1)P(C = 0|T = 1)P(T = 1)$$
$$\propto 1/3 * 1/3 * 2/3 * 3/5$$

Then we have $P(h_1|x_6) \propto 1/3 * 1/3 * 2/3 * 3/5 [= 2/45]$

Remark: What is P(T= 1|A= 1, B= 1, C= 0) ?

Notice that, we have obtained

$$P(h_0|x_6) =$$
$$P(T=0|A=1,B=1,C=0) = \frac{P(A=1,B=1,C=0|T=0)P(T=0)}{P(x_6)}$$
$$\propto P(A=1,B=1,C=0|T=0)P(T=0)$$

Here we don't consider (denominator) $P(x_6)$ as it does not affect the final result, i.e., h_{Map} hypothesis. In fact, both $P(h_1|x_6)$ and $P(h_2|x_6)$ are normalized by $P(x_6)$

Similarly, we have for h_1

$$P(h_1|x_6) =$$
$$P(T=1|A=1,B=1,C=0) = \frac{P(A=1,B=1,C=0|T=1)P(T=1)}{P(x_6)}$$
$$\propto P(A=1,B=1,C=0|T=1)P(T=1)$$

What is P(T= 1|A= 1, B= 1, C= 0) ?

Let's compute now $P(x_6)$

$$P(x_6) = \sum_i P(x_6, h_i) = \sum_{i\in\{0,1\}} P(A=1,B=1,C=0|T=i)P(T=i)$$
$$= P(A=1|T=0)P(B=1|T=0)P(C=0|T=0)P(T=0)+$$
$$+ P(A=1|T=1)P(B=1|T=1)P(C=0|T=1)P(T=1)$$

We have

$$P(x_6) = (1/2*1*0*2/5) + (1/3*1/3*2/3*3/5) = (1/3*1/3*2/3*3/5)[= 2/45]$$

What is P(T= 1 | A= 1, B= 1, C= 0) ?

Then for $h_1 = [T = 1]$, [Similarly for $h_0 = [T = 0]$]

$$P(h_1|x_6) =$$

$$P(T = 1|A = 1, B = 1, C = 0) = \frac{P(A = 1, B = 1, C = 0|T = 1)P(T = 1)}{P(x_6)} =$$

$$= \frac{P(A = 1|T = 1)P(B = 1|T = 1)P(C = 0|T = 1)P(T = 1)}{P(x_6)}$$

$$= \frac{(1/3 * 1/3 * 2/3 * 3/5)}{(1/3 * 1/3 * 2/3 * 3/5)} = 1$$

Notice that, more easily here we have $P(h_1|x_6) = 1$, so clearly $P(h_0|x_6) = 0$

Issues with Naive Bayes

- Generally features are not conditionally independent given the label

$$P(\mathbf{x}|y) \neq \prod P(x_j|y)$$

Nonetheless, NB is the single most used classifier particularly when data is limited, works well

Issues with Naive Bayes

- Not enough training data to get good estimates of the probabilities from counts
- What if we never see a particular feature with a particular label? Eg: Suppose we never observe Temperature = cold with PlayTennis= Yes
  - $P(X_1 = a \mid Y = b) = 0$

that will make the probabilities zero

Answer: Smoothing
- Add fake counts (very small numbers so that the counts are not zero)

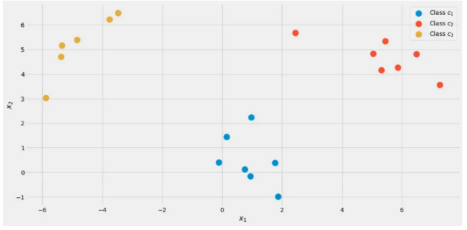
Gaussian Naive Bayes

- Naive Bayes estimates probabilities based on the class frequencies of each feature in the training data.
- How does it calculate the frequency of continuous variables?
  1. Assume that all the features are following a gaussian i.e, normal distribution, or
  2. Recode the feature (variable) values into quartiles, such that e.g.,
     - values less than the 25th percentile are assigned a 1,
     - 25th to 50th a 2,
     - 50th to 75th a 3 and
     - greater than the 75th percentile a 4.

     → Calculations are merely done on these categorical bins.

Example

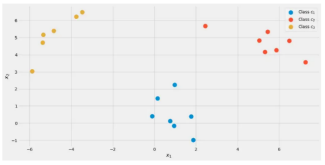
- Let us use a toy dataset with **two real features** x_1, x_2, and **three classes** c_1, c_2, c_3 in the following.



The data, visualized. Image by the Author.

A simple prior for Gaussian Bayes

- **class probability** $p(c)$, the probability that some class c is observed in the labeled dataset.
  - just compute the relative frequencies of the classes and use them as the probabilities. We can use our dataset to see what this means exactly.



The data, visualized. Image by the Author.

7 out of 20 points labeled class c_1 (blue): $p(c_1)$=7/20.
7 points for class c_2 (red): $p(c_2)$=7/20.
The last class c_3 (yellow) has only 6 points: $p(c_3)$=6/20.

Remark: Prior

- You can, however, also use another *prior* distribution, if you like.
- For example, if you know that this dataset is not representative of the true population because class c_3 should appear in 50% of the cases, then you set
- $p(c_1)$=0.25,
- $p(c_2)$=0.25 and
- $p(c_3)$=0.5.

Likelihood

- we have real feature values and decide for a **Gaussian** distribution, hence the name Gaussian naive Bayes.

$$p(x_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_{i,j}}{\sigma_{i,j}}\right)^2} \text{ for } i = 1, 2 \text{ and } j = 1, 2, 3$$

$\mu_{i,j}$ is the mean and
$\sigma_{i,j}$ is the standard deviation that we have to estimate from the data.

This means that we get one mean for each feature i coupled with a class c_j, in our case 2*3=6 means.
The same goes for the standard deviations.