Classification: Naive Bayes Classifiers

Prof. (Dr.) Honey Sharma

Reference Book: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, A
Introduction to Statistical Learning with Applications in R

Conditional Probability

If events are dependent, then their probability is expressed by conditional probability. The probability that *A* occurs given that *B* is denoted by P(A|B).

Suppose, A and B are two events associated with a random experiment. The probability of A under the condition that B has already occurred and P(B)≠0 is given by

$P(A \mid B)$=(Number of events in B which are favourable to A)/(Number of events in B)

$=P(A \cap B)/P(B)$

$$P(A \cap B) = P(A).P(B \mid A), \quad \text{if } P(A) \neq 0$$

$$P(A \cap B) = P(B).P(A \mid B), \quad \text{if } P(B) \neq 0$$

For three events A, B and C

$$P(A \cap B \cap C) = P(A).P(B).P(C \mid A \cap B)$$

Note:

$P(A \mid B) = 0$ if events are mutually exclusive

$P(A \mid B) = P(A)$ if A and B are independent

Generalization of Conditional Probability:

P(A | B)=P(A∩B)/P(B)=P(B∩A)/P(B)

= P(B|A)·P(A)/P(B) ∵ P(A∩B)=P(B|A) ·P(A) = P(A|B) ·P(B)

By the law of total probability : P(B) = P[(B∩A) ∪ (B∩A')],

$$P(A \mid B) = \frac{P(B|A) \bullet P(A)}{P[(B \cap A) \cup (B \cap A')} = \frac{P(B \mid A) \bullet P(A)}{P(B \mid A) \bullet P(A) + P(B \mid A') \bullet P(A')}$$

Total Probability

Let be $E_1, E_2, \ldots\ldots E_n$ n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with them, then

$$P(A) = P(E_1).P(A \mid E_1) + P(E_2).P(A \mid E_2) + \ldots + P(E_n).P(A|E_n)$$

Bayes' Theorem of Probability

Let $E_1, E_2, \ldots\ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 or E_2 or……E_n , then

$$P(E_i|A) = \frac{P(E_i).P(A|E_i)}{\sum_{i=1}^{n} P(E_i).P(A|E_i)}$$

Prior and Posterior Probabilities

- P(A) and P(B) are called prior probabilities
- P(A|B), P(B|A) are called posterior probabilities

Example: This table shows that the event Y has two outcomes namely A and B, which is dependent on another event X with various outcomes like x_1, x_2 and x_3.

- Case1: Suppose, we don't have any information of the event A. Then, from the given sample space, we can calculate P(Y = A) = 5/10 = 0.5
- Case2: Now, suppose, we want to calculate P(X = x_2|Y =A) = 2/5 = 0.4 .

The later is the conditional or posterior probability, where as the former is the prior probability.

X	Y
x_1	A
x_2	A
x_3	B
x_3	A
x_2	B
x_1	A
x_1	B
x_3	B
x_2	B
x_2	A

Naïve Bayesian Classifier

From Bayes' theorem on conditional probability, we have

$$P(Y \mid X) = \frac{P(X|Y) \bullet P(Y)}{P(X)}$$

$$P(Y \mid X) = \frac{P(X|Y) \bullet P(Y)}{\sum_{i=1}^{k} P(X|Y=y_i) \bullet P(Y=y_i)}$$

Note:

- P(X) is called the evidence (also the total probability) and it is a constant.
- The probability P(Y|X) (also called class conditional probability) is therefore proportional to P(X|Y)·P(Y).

Thus, P(Y|X) can be taken as a measure of Y given that X.

$$P(Y|X) \approx P(X \mid Y) \bullet P(Y)$$

- Suppose, for a given instance of X (say x = (x_1……..x_n)).
- There are any two class conditional probabilities namely P(Y=y_i|X=x) and P(Y=y_j | X=x).
- If P(Y=y_i | X=x) > P(Y=y_j | X=x), then we say that y_i is more stronger than y_j for the instance X = x.
- The strongest y_i is the classification for the instance X = x.

Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late

Days	Season	Fog	Rain	Class
Weekday	Summer	High	None	On Tim
Weekday	Winter	Normal	Heavy	Very La
Saturday	Autumn	High	Slight	On Tim
Weekday	Autumn	None	Heavy	On Tim
Holiday	Spring	Normal	Slight	On Tim
Weekday	Spring	Normal	None	On Tim
Weekday	Spring	Normal	Heavy	On Tim

https://docs.google.com/spreadsheets/d/1UfDW1pNvrr-oWK3D3GnTUFWo496AnhOqnuhVZjvZdak/edit#gid=1189150732

Instance:

Days:WeekDay, **Season**:Winter **Fog**:High, **Rain**: Heavy **Class???**

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.14 = 0.0025

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.0 = 00.0000

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

Case3 is the strongest; Hence correct classification is Very Late

Pros & Cons

- This algorithm works very fast and can easily predict the class of a test dataset.
- You can use it to solve multi-class prediction problems as it's quite useful with them.
- Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.

However, it has a number of potential problems

- If your test data set has a categorical variable of a category that wasn't present in the training data set, the Naive Bayes model will assign it zero probability and won't be able to make any predictions in this regard. This phenomenon is called 'Zero Frequency,' and you'll have to use a smoothing technique to solve this problem.
- It assumes that all the features are independent. While it might sound great in theory, in real life, you'll hardly find a set of independent features.

Zero Frequency problem

It occurs when any condition having zero probability in the whole multiplication of the likelihood makes the whole probability zero. In such a case, Laplace Estimator is used:

$$P(x_i \mid y_i) = \frac{n_c + \alpha}{n + \alpha m}$$

where, **n_c** = number of instances where xi = x and yi = y,
n = number of instances where yi = y,
m =the number of different (unique) attribute values.
Alpha =represents the smoothing parameter, Using higher alpha values will push the likelihood towards a value of 0.5, i.e., the probability equal to 0.5. Since we are not getting much information from that, it is not preferable. Therefore, it is preferred to use alpha=1.

https://docs.google.com/spreadsheets/d/1UfDW1pNvrr-oWK3D3GnTUFWo496AnhOqnuhVZjvZdak/edit#gid=1189150732

Instance:

Days:WeekDay, **Season**:Winter **Fog**:High, **Rain**: Heavy **Class???**

Case1: Class = On Time : 0.003478608

Case2: Class = Late : 0.00132

Case3: Class = Very Late : 0.006066225

Case4: Class = Cancelled : 0.0005

Case3 is the strongest; Hence correct classification is Very Late

Continuous Attributes

- Let x1, x2, …, xn be the values of a numerical attribute in the training data set.

https://docs.google.com/spreadsheets/d/1UfDW1pNvrr-oWK3D3GnTUFWo496AnhOqnuhVZjvZdak/edit#gid=1189150732

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$