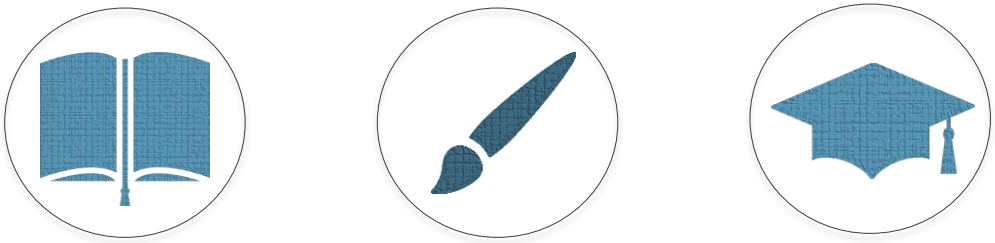

Data Science Concepts

Lesson06-Naïve Bayes Classifier

Objective

After completing this lesson you will be able to:

- Understand Bayes Theorem
- Explain Naïve Bayes Classifier
- Describe the simplifying assumption of Naïve Bayes Classifier
- Explain the steps in building Naïve Bayes model

Naïve Bayes

One of the ways to assign an individual/object/data to a particular **class** is to use Naïve Bayes

Based on Bayes Theorem which provides away to calculate the probability of a **class** given prior knowledge of the problem.

Bayes Theorem

$$P(c|d) = \frac{P(d|c) * P(c)}{P(d)}$$

- $P(c|d)$ is the probability of class given the data d. This is called the posterior probability.
- $P(d|c)$ is the probability of data d given that the class to which it belongs was true.
- $P(c)$ is the probability of class c being true (regardless of the data). This is called the prior probability of class.
- $P(d)$is the probability of the data (regardless of the class).

$$Posterior\ Probability = \frac{conditional\ probability * prior\ probability}{evidence}$$

Naïve Bayes Classifier

Compute the posterior probability for a number of different class and select the class with the highest probability.

Objective function is:

$$Maximum\ a\ posterior\ \ class\ [MAP(c)] = \max[P(c|d)]$$

Simplifying assumption of Naïve Bayes

1. Each input value is assumed to be conditionally independent given the outcome variable

$$P(d_1, d_2, d_3|c) = P(d_1|c) * P(d_2|c) * P(d_3|c)$$

How likely it is to observe a particular pattern (d_1, d_2, d_3) given that it belongs to class c?

2. The samples are I.I.D (Independent and Identically Distributed)

It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each class are simplified by making the assumption that inputs are mutually independent.

Representation of Naïve Bayes Model

Class Probabilities: The probabilities of each class in the training dataset.

Conditional Probabilities: The conditional probabilities of each input value given each class value.

Steps in Naïve Bayes Model:

1. Compute the class probability from the data
2. Compute the conditional probability

Example

Renege data with 8995 observation across four variables. Below is a subset of data

DOJ Extended	Gender	Candidate Source	Status
Yes	Female	Agency	Joined
No	Male	Employee Referral	Joined
No	Male	Agency	Joined
No	Male	Employee Referral	Joined
Yes	Male	Employee Referral	Joined
Yes	Male	Employee Referral	Joined
Yes	Male	Employee Referral	Joined
Yes	Female	Direct	Joined
No	Female	Employee Referral	Joined
No	Male	Employee Referral	Joined
No	Male	Employee Referral	Not Joined
No	Male	Employee Referral	Joined
Yes	Male	Agency	Not Joined
No	Male	Direct	Not Joined
No	Male	Employee Referral	Not Joined
No	Male	Direct	Joined
Yes	Male	Agency	Not Joined

Probability Calculation

Calculate the class probability and conditional probability using frequency count

Class Probability	
Joined	Not Joined
0.813007	0.186993

Conditional Probabilities	Probability
P(DOJ Extension = Yes|Status = Joined)	0.469164502
P(DOJ Extension = No|Status = Joined)	0.530835498
P(DOJ Extension = Yes|Status = Not Joined)	0.461355529
P(DOJ Extension = No|Status = Not Joined)	0.538644471
P(Gender = Male|Status = Joined)	0.825242718
P(Gender = Female|Status = Joined)	0.174757282
P(Gender = Male|Status = Not Joined)	0.837693222
P(Gender = Female|Status = Not Joined)	0.162306778
P(Candidate Source = Agency|Status = Joined)	0.268015862
P(Candidate Source = Direct|Status = Joined)	0.538356352
P(Candidate Source = Employee Referal|Status = Joined)	0.193627786
P(Candidate Source = Agency|Status = Not Joined)	0.371581451
P(Candidate Source = Direct|Status = Not Joined)	0.513674197
P(Candidate Source = Employee Referal|Status = Not Joined)	0.114744352

Prediction Calculation

Look up the unique combination across three input variables

DOJ Extended	Gender	Candidate Source	Joined	Not Joined	Prediction
Yes	Female	Agency	0.017865506	0.02262147	Not Joined
No	Male	Employee Referral	0.068961031	0.009681516	Joined
No	Male	Agency	0.095454534	0.031352058	Joined
Yes	Male	Employee Referral	0.060949329	0.008292336	Joined
Yes	Female	Direct	0.035885969	0.007192584	Joined
No	Female	Employee Referral	0.014603512	0.001875837	Joined
Yes	Male	Agency	0.084364891	0.026853419	Joined
No	Male	Direct	0.19173699	0.043341085	Joined
Yes	Male	Direct	0.169461518	0.037122166	Joined
Yes	Female	Employee Referral	0.012906917	0.001606677	Joined
No	Female	Direct	0.040603127	0.008397528	Joined
No	Female	Agency	0.020213901	0.0060746	Joined

$$P(Joined|DOJ\ Extended, Gender, Candidate\ Source) = P(DOJ\ Extended|Joined) * P(Gender|Joined) * P(Candidate\ Source|Joined) * P(Joined)$$

Summing up

Tag the prediction for each observation in the dataset

DOJ Extended	Gender	Candidate Source	Status	Prediction
Yes	Female	Agency	Joined	Not Joined
No	Male	Employee Referral	Joined	Joined
No	Male	Agency	Joined	Joined
No	Male	Employee Referral	Joined	Joined
Yes	Male	Employee Referral	Joined	Joined
Yes	Male	Employee Referral	Joined	Joined
Yes	Male	Employee Referral	Joined	Joined
Yes	Female	Direct	Joined	Joined
No	Female	Employee Referral	Joined	Joined
No	Male	Employee Referral	Joined	Joined
No	Male	Employee Referral	Not Joined	Joined
No	Male	Employee Referral	Joined	Joined
Yes	Male	Agency	Not Joined	Joined
No	Male	Direct	Not Joined	Joined
No	Male	Employee Referral	Not Joined	Joined
No	Male	Direct	Joined	Joined
Yes	Male	Agency	Not Joined	Joined

End of Lesson06-Naïve Bayes Classifier