Hands on Lab on Naive Bayes Algorithm

AGENDA OF THE VIDEO

- What is Naive Bayes Algorithm?
- Bayes Theorem
- Discussion on Prior and Posterior Probability
- What do we mean by feature Independence?
- Algorithm Description
- Distributions of Likelihood (Features): Gaussian, Bernoulli, Multinomial
- Predicting whether the person is male or female (Gaussian Likelihood)
- Coding the Naive Bayes Algorithm for Gaussian Likelihood..
- Sklearn Class for Naive Bayes Algorithm
- Task For you………. (with some other likelihood)
- One more Task with Real World Dataset

What is Naive Bayes Algorithm?

Naive Bayes Algorithm is an algorithm based on one of the theorem in probability theory called "Bayes theorem". It is mostly used for classification problems with special use in text related data.

This algorithm is called "naive" because of its over simplified assumption of independence between features.

It uses Bayes Theorem to predict the posterior probability of a feature belonging to a given class given its prior probability.

Bayes Theorem

Bayes Theorem gives us a mathematical formula to calculate the posterior probability.

Posterior probability is a probability after some information has been received. So, the probability without any information will be known as prior probability.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Prior and Posterior Probabilities of an event

If we have an event and we compute its probability without any information, then that probability will be known as prior probability.

Now, if we compute the probability of the same event given some additional information, then the same probability will be known as posterior probability.

The probability of rain is (without any information) will most often be ½. Now, if I say, I have the data for humidity, wind and cloud, then this probability will be updated based on some evidence.

Feature Independence

Consider a problem of predicting whether a customer of a bank will default or not.

Now, the feature might be: No of family member, Income, Asses Value etc.

Now, when we say that naive bayes assumes independence between features, what we are essentially trying to convey is these above feature will have no inter dependence.

So, no of family members will be independent of Income. But it is rarely the case in real world which is why we call naive bayes algorithm as "naive".

Algorithm Description

We have a set of features say (X1, X2, …, Xn) and a variable for class Y which has k class say (y1, y2...yk),

Given a new sample vector say X_cap. So, we compute its class by calculating

P(yi given X_cap) for all i = 1 to k.

Then predict that class for which the above probability is maximum. We compute the above probability by using Bayes Theorem that we discussed earlier.

Algorithm Description (Continued)

P(yi given X_cap) = (P(X_cap given yi) * P(yi))/P(X_cap)

We can get rid of the denominator and the formula reduces to:

P(yi given X_cap) = P(X_cap given yi) * P(yi)
                          = Likelihood * Prior

How to come up with the distribution of Likelihood?

It depends on feature to feature.

For continuous data, the Gaussian Likelihood will be used.

For binary data, the bernoulli Likelihood will be used.

For the multiple class data, the multinomial Likelihood will be used.

Gaussian Likelihood

We split the data into separate dataset for each individual class (yi). If there are k distinct class values, we will have k subsets of data.

For each data, we need to compute each features mean and variance. Then, to calculate the likelihood probability, we use the following formula:

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

Predict the person: male or female?

person	height	weight	footSize
male	6	180	12
male	5.92	190	11
male	5.58	170	12
male	5.92	165	10
female	5	100	6
female	5.5	150	8
female	5.42	130	7
female	5.75	150	9

Coding the Naive Bayes algorithm in Scratch...

Gaussian Naive Bayes Class in Sklean

Task For you….

Build a naive bayes classification model using this dataset about whether a particular person should go out and play or not based on these two variables on whether the day was sunny or clouds were there.

I have created this dataset myself. So, please do not start to judge the authenticity as it might contain some unbelievable combination of observations.

Play or not	Sunny Day	Clouds
Play	Yes	No
Play	Yes	Yes
Play	Yes	No
Play	Yes	No
No Play	No	Yes
No Play	No	Yes
No Play	Yes	Yes
No Play	No	No

Task For you….

For this task, I want you to visit Kaggle.com which is one of the most powerful website for participating in data science competitions related to machine learning model building.

I want you to go to search bar and search for "Quora Insincere Question Classification Data".

I want you to download the dataset and fit a Multinomial Naive Bayes algorithm to predict whether the question is insincere or not.

Thank You….