Term Project Report

CS40003 - Data Analytics

Group ID - 22, Project ID - 7

"Probabilistic classification using Statistical Naïve Bayes' classifier"

TEAM MEMBERS-

1. Saurabh Singh, 17MT30018 Signature:-_____
2. Aniket Niranjan Mishra, 17MT3FP20 Signature:-_____

Problem description-

a) Obtain the appropriate contingency table from a training data set comprising the prior and posterior probabilities. You should split data into two sets in an appropriate manner.

b) Test the classifiers using k-fold cross validation technique. Run with different value of k and then choose the optimum result.

c) Furnish the accuracy using an appropriate confusion matrix and report the performance evaluation with different matrix (e.g., Precision, Recall, F1 score, etc.).

Dataset description-

CMC (Contraceptive Method Choice) data for the year 2016-17

Independent attributes:-

1. Wife's age (numerical)
2. Wife's education (categorical)
    a. 1=low, 2, 3, 4=high
3. Husband's education (categorical)
    a. 1=low, 2, 3, 4=high
4. Number of children ever born (numerical)
5. Wife's religion (binary)
    a. 0=Non-Islam, 1=Islam
6. Wife's now working? (binary)
    a. 0=Yes, 1=No
7. Husband's occupation (categorical)
    a. 1, 2, 3, 4
8. Standard-of-living index (categorical)
    a. 1=low, 2, 3, 4=high
9. Media exposure (binary)
    a. 0=Good, 1=Not good

Dependent attributes:-

Contraceptive method used (class attribute) 1=No-use, 2=Long-term, 3=Short-term

The various describing functions applied on the variables are as follows-

	age	edu_wife	edu_husband	children	religion	working	job_husband	living_std	media	contraceptive
count	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000
mean	32.538357	2.958588	3.429735	3.261371	0.850645	0.749491	2.137814	3.133741	0.073999	1.919891
std	8.227245	1.014994	0.816349	2.358549	0.356559	0.433453	0.864857	0.976161	0.261858	0.876376
min	16.000000	1.000000	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000
25%	26.000000	2.000000	3.000000	1.000000	1.000000	0.000000	1.000000	3.000000	0.000000	1.000000
50%	32.000000	3.000000	4.000000	3.000000	1.000000	1.000000	2.000000	3.000000	0.000000	2.000000
75%	39.000000	4.000000	4.000000	4.000000	1.000000	1.000000	3.000000	4.000000	0.000000	3.000000
max	49.000000	4.000000	4.000000	16.000000	1.000000	1.000000	4.000000	4.000000	1.000000	3.000000

Naive Bayesian Classifier theory

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high.

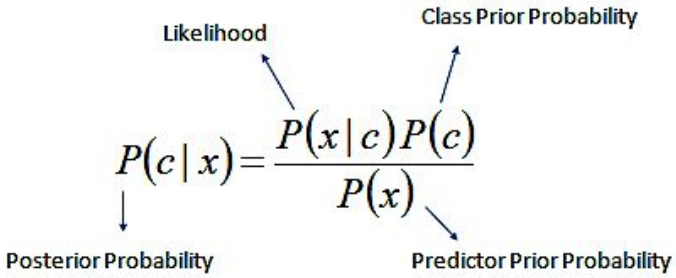
This classifier assumes-

1. The classes are mutually exclusive and exhaustive.
2. The attributes are independent given the class.

Hence the word **naive** is used.

If we have a certain event c and features x1, x2, x3, etc.

We first calculate P(x1 | c) , P(x2 | c) … [probability of x1 given event c happened] and then select the feature x with maximum probability value. The formulae used are-



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

P(A) and P(B) are called prior probabilities, whereas P(A|B), P(B|A) are called posterior probabilities (After all given evidence or background information has been taken into account).

Posterior probability = prior probability + new evidence (called likelihood)

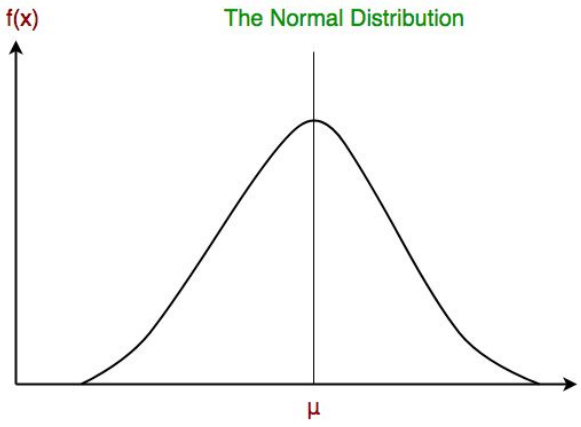
Pros and Cons of Naive Bayes' classifier-

1. All attributes should be categorical.
2. The estimation is poor on availability of less data.
3. The assumption that the input features are independent does not hold good in most cases. In the case of our data, the Wife's Working and Wife's Education are logically correlated, but we assume that they are not so.
4. The classifier does not work well on continuous (numerical) attributes hence we decided to remove the two continuous variables i.e. Wife's age and No. of children.

Gaussian Naive Bayes classifier

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a **Gaussian distribution (normal distribution)**.

The plot as well as the conditional probability according to the Gaussian assumption is-



$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

Naive Bayes implementation in Python from scratch

The steps to be followed for implementing naive bayes using python are:-

1. Differentiate into classes
    a. We will need to calculate the probability of data by the class they belong to.This means that we will first need to separate our training data by class.
2. Prepare a summary of the dataset
    a. We need two statistics from a given set of data, the Mean and Standard Deviation.
    b. The mean is the average value and the sample standard deviation is calculated as the mean difference from the mean value.
3. Prepare summary of data by class
    a. We require statistics from our training dataset organized by class.
4. Gaussian Probability Density Function
    a. Calculating the probability or likelihood of observing a given real-value is difficult.One way we can do this is to assume that values are drawn from a distribution, such as a bell curve or Gaussian distribution.
    b. A Gaussian distribution can be summarized using only two numbers: the mean and the standard deviation. We can estimate the probability of a given value. This is called a Gaussian Probability Distribution Function.
5. Calculating Class Probabilities
    a. Probabilities are calculated separately for each class. This means that we first calculate the probability that a new piece of data belongs to the first class, then calculate probabilities that it belongs to the second class, and so on for all the classes.
6. Predicting class based on highest probability
7. Validation using K-fold cross validation