**National University of Computer & Emerging Sciences**

**Karachi Campus**



**Project Report**

**Data Structures**

**Section: J**

**[Linear regression and Naive Bayes algorithm C++]**

**Group Members:**

**19K-0249 Haris Aqeel**

**19K-0271 Ayesha Saif**

**19K-1326 Muhammad Jahanzeb**

Introduction

The project is able to predict the specie of an iris flower up to a minimum accuracy of 90% using either of the two methods provided to the user. As the program is run, it provides the user with three options, first two being the method of prediction and the last one being the choice to exit the program. The two methods used in the project are (1) Linear Regression & (2) Naïve Bayes theorem. Each of these method uses both a different strategy and a different set of data to come to the mathematically most probable outcome possible.

What is Linear Regression?

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

WORKING:

STEP 1: Assume a mathematical relationship between the target and the predictor(s). “The relationship can be a straight line (linear regression) or a polynomial curve (polynomial regression) or a non-linear relationship (non-linear regression)”

STEP 2: Create a scatter plot of the target variable and predictor variable (simplest and most popular way).

Image for post

STEP 3: Find the most-likely values of the coefficients in the mathematical formula

FUNCTIONALITY:

Regression analysis comprises of the entire process of identifying the target and predictors,finding the relationship, estimating the coefficients, finding the predicted values of target, and finally evaluating the accuracy of the fitted relationship.

What is Naïve Bayes?

Naive Bayes classifiers are linear classifiers based on Bayes’ theorem. The model generated is probabilistic. It estimates conditional probability which is the probability that something will happen, given that something else has already occurred. For example, the given mail is likely spam given the appearance of the words such as ‘prize’.

It is called naive due to the assumption that the features in the dataset are mutually independent. In the real-world, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well.

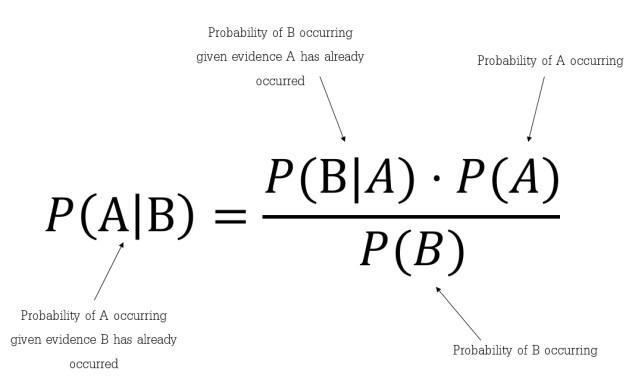
The idea is to factor all available evidence in form of predictors into the naive Bayes rule to obtain a more accurate probability for class prediction. Being relatively robust, easy to implement, fast, and accurate, naive Bayes classifiers are used in many different fields.

WORKING:

**Bayes’ Theorem** is a simple mathematical formula used for calculating conditional probabilities.

**Conditional probability** is a measure of the probability of an event occurring given that another event has (by assumption, presumption, assertion, or evidence) occurred.

The formula is: —



Which tells us: how often A happens given that B happens, written **P(A|B)**also called posterior probability, When we know: how often B happens given that A happens, written **P(B|A)** and how likely A is on its own, written **P(A)** and how likely B is on its own, written **P(B).**

In simpler terms, Bayes’ Theorem is a way of finding a probability when we know certain other probabilities.

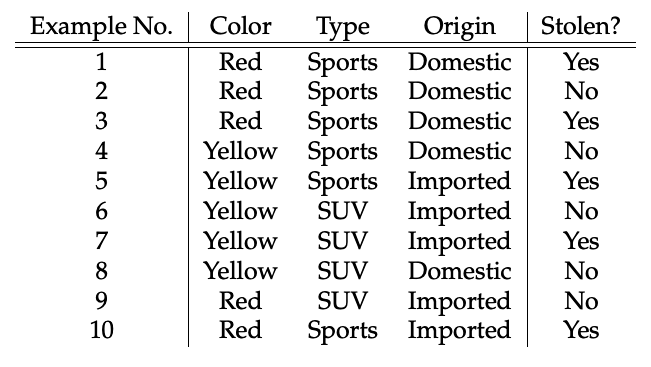
The fundamental Naïve Bayes assumption is that each feature makes an:

* independent
* equal contribution to the outcome.

Let us take an example to get some better intuition. Consider the car theft problem with attributes Color, Type, Origin, and the target, Stolen can be either Yes or No.

### Example

The dataset is represented as below.



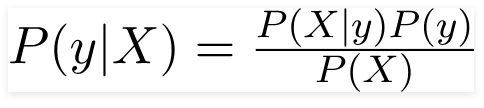
Concerning our dataset, the concept of assumptions made by the algorithm can be understood as:

* We assume that no pair of features are dependent. For example, the color being ‘Red’ has nothing to do with the Type or the Origin of the car. Hence, the features are assumed to be **Independent**.
* Secondly, each feature is given the same influence(or importance). For example, knowing the only Color and Type alone can’t predict the outcome perfectly. So none of the attributes are irrelevant and assumed to be contributing **Equally** to the outcome.

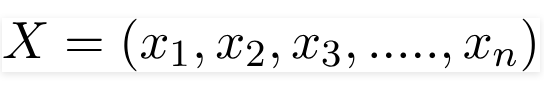
**Note:** The assumptions made by Naïve Bayes are generally not correct in real-world situations. The independence assumption is never correct but often works well in practice. **Hence the name ‘Na**ï**ve’.**

Here in our dataset, **we need to classify whether the car is stolen, given the features of the car**. The columns represent these features and the rows represent individual entries. If we take the first row of the dataset, we can observe that the car is stolen if the Color is Red, the Type is Sports and Origin is Domestic. So we want to classify a Red Domestic SUV is getting stolen or not. Note that there is no example of a Red Domestic SUV in our data set.

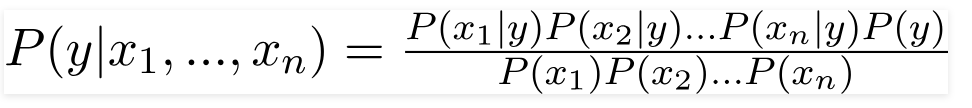
According to this example, Bayes theorem can be rewritten as:



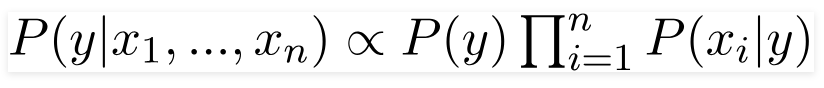
The variable **y** is the class variable (stolen?), which represents if the car is stolen or not given the conditions. Variable **X**represents the parameters/features.  
**X** is given as,



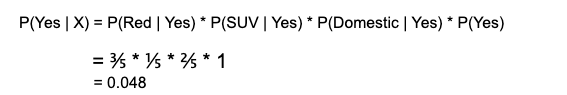
Here x1, x2 …. xn represent the features, i.e they can be mapped to Color, Type, and Origin. By substituting for **X**and expanding using the chain rule we get,



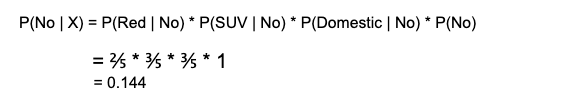
Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remains static. Therefore, the denominator can be removed and proportionality can be injected.



As per the equations discussed above, we can calculate the posterior probability P(Yes | X) as :



and, P(No | X):



Since 0.144 > 0.048, which means given the features RED SUV and Domestic, our example gets classified as ’NO’ the car is not stolen.

FUNCTIONALITY:

Naïve Bayes algorithms are often used in sentiment analysis, spam filtering, recommendation systems, etc. They are quick and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent.

CONCLUSION:

In conclusion, the program is able to predict an outcome via two different methods using the data set provided to certain level of acceptable accuracy. This kind of prediction is very useful when a large amount of data is provided where each data member is independent of the other and an estimate has to be made of the most likely outcome such as in weather forecasting, marketing effectiveness, pricing and promotions on sales of a product, etc. and as in our case, predicating the specie of a flower.