Naive Bayes classifier and Maximum Likelihood Estimation

Q.1. Using the Naive Bayes Classifiers classify potential customers who are more likely to purchase a loan by the *Bank_Personal_Loan_Modelling* dataset. The dataset description is given below.

Dataset:

The *Bank_Personal_Loan_Modelling* dataset includes 5000 observations with fourteen variables divided into four different measurement categories. The binary category has five variables, including the target variable personal loan, also securities account, CD account, online banking and credit card. The interval category contains five variables: age, experience, income, CC avg and mortgage. The ordinal category includes the variables family and education. The last category is nominal with ID and Zip code. The variable ID does not add any interesting information e.g. individual association between a person (indicated by ID) and loan does not provide any general conclusion for future potential loan customers. Therefore, it will be neglected in the examination.

Note: Don’t use python libraries, implement functions by yourself.

Step 1: Calculate the prior probability for given class labels

Step 2: Find Likelihood probability with each attribute for each class

Step 3: Put these values in Bayes Formula and calculate posterior probability.

Step 4: See which class has a higher probability, given the input belongs to the higher probability class.

Que. Using the Bayesian classifier, make a function which takes input as Age, Income and Education . Output: Predict whether loan will be given or not.

Que. Do the performance analysis of the above model using precision, recall and F1-score.

Q.2. Implement the same by using a Gaussian Naive Bayes classifier.

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

Gaussian Naive Bayes classifier

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values.

Que. Using the **Gaussian Naive Bayes classifier**, make a function which takes input as Age, Income and Education . Output: Predict whether loan will be given or not.

Que. Also the plot graph and verify the bell curve.

Que. Do the performance analysis of the above model using precision, recall and F1-score.

Q.3. Assume that *m* random vectors, each of size *p*: $X^{(1)}, X^{(2)},., X^{(m)}$ where each random vector can be interpreted as an observation (data point) across *p* variables.

Consider a sample multivariate Normal Distribution model

$$f_{\mathbf{X}^{(i)}}(\mathbf{x}^{(i)}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu)^{\mathbf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)\right)$$

The likelihood function is defined as the product of the individual densities.

$$l(\mu, \mathbf{\Sigma}|\mathbf{x}^{(i)}) = \prod_{i=1}^{m} f_{\mathbf{X}^{(i)}}(\mathbf{x}^{(i)}|\mu, \mathbf{\Sigma})$$

The log-likelihood function is defined as

$$l(\mu, \Sigma;) = -\frac{mp}{2}\log(2\pi) - \frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{x}^{(i)} - \mu)^{\mathbf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu)$$

The maximum likelihood estimate is

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}^{(\mathbf{i})} = \bar{\mathbf{x}}$$

$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^{(\mathbf{i})} - \hat{\mu})(\mathbf{x}^{(\mathbf{i})} - \hat{\mu})^{T}$$

A binary classification problem between male and female individuals using height, weight and BMI is given in the dataset (500_Person_Gender_Height_Weight_Index).

1. Calculate the distribution of males using maximum likelihood.
2. Calculate the distribution of females using maximum likelihood.
3. When you get a new unlabelled data point, calculate the probability of that new data point belonging to both distributions, and assign it to the class (male or female) for which the distribution yields the highest probability.