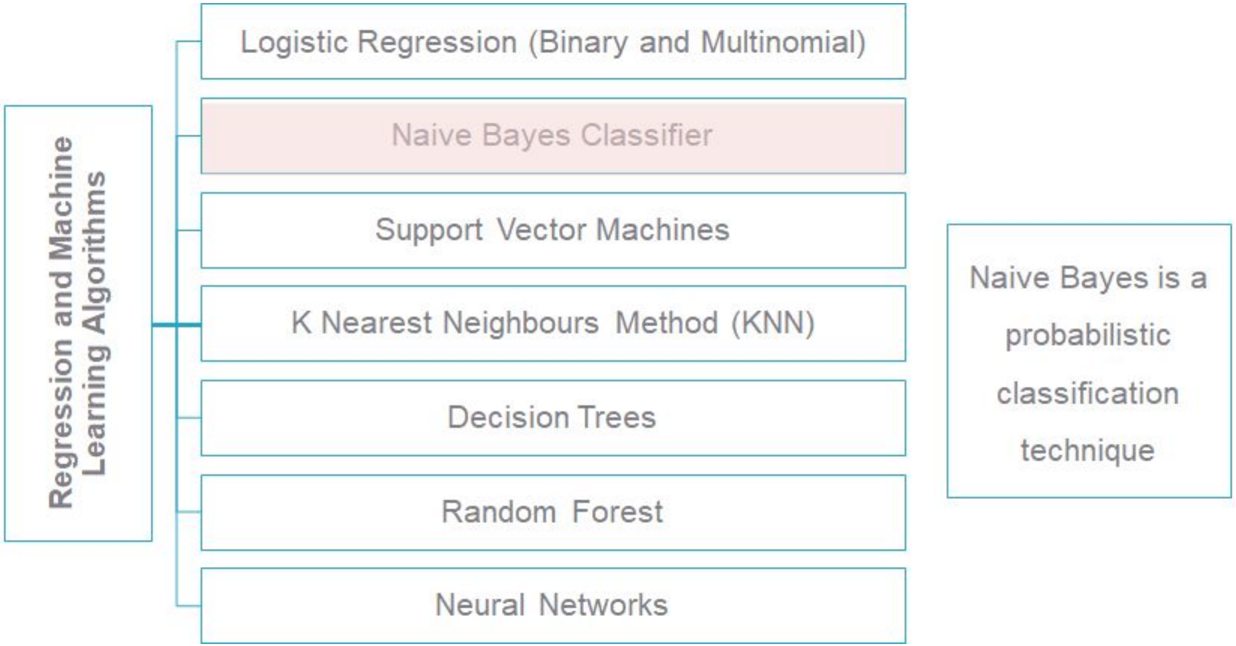
Naive Bayes Classifier I

Classifier Based on Bayes' Theorem

Contents

Classification Methods

Apart from logistic regression, several types of machine learning algorithms are effective in classification and prediction.



About Naive Bayes Classifier

- Simple **probabilistic classifier based on Bayes Theorem.**
- It can be used as an **alternative method to logistic regression (Binary or Multinomial).**
- It **assumes conditional independence among the predictors.**
- It is particularly **suited when the dimensionality of the inputs is high.**

Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

Conditional Probability

The conditional probability of an event B is the probability that event B will occur given the knowledge that an event A has already occurred. This probability is written as $P(B|A)$.

- If A and B are independent events then

$$P(B|A) = P(B)$$

- An unbiased die, with numbers 1-6 is tossed

A: Getting a number greater than 1

B: Getting an even number

$$P(A) = 5/6$$

$$P(B) = 3/6$$

$$P(B|A) = 3/5$$

Here the sample space has 5 points given A has occurred.

Bayes Theorem

Conditional Probability

Posterior Probability

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

Prior Probability

Prior Probability

where

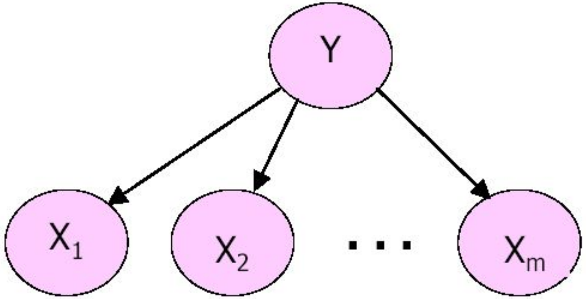
P(A) : Prior probability or marginal probability of A

P(A | B) : Conditional probability of A given B

P(B | A) : Conditional probability of B given A

P(B) : Prior or marginal probability of B

Naive Bayes Framework



Y : Categorical Dependent Variable

X_i : Categorical/Continuous Independent Variable

Objective: To estimate Y given the values of X_i's or

To estimate $P(Y|X_1, X_2, \dots, X_m)$ using the Naïve Bayes Classifier

Assumption: All X_i's are conditionally independent of each other

Naive Bayes Framework - Example

Consider a simple example where Y is binary (response to a certain question) with 2 independent categorical variables X_1 and X_2

We classify	**Y = 1 "Buyer"** **Y = 0 "Non-Buyer"**
Let $\mathbf{X_1}$ denote **age of the individual**	X_1 = 0 for age group 25-30 years X_1 = 1 for age group 31-40 years
Let $\mathbf{X_2}$ denote **gender**	X_2 = 0 if Gender=female X_2 = 1 if Gender=male

Classification Rule

For the given values of X_1 and X_2 we want to know if the individual will be a potential buyer or not. Using Naive Bayes classifier we estimate:

$$P(Y = 0|X_1 = a_1, X_2 = a_2)$$

&

$$P(Y = 1|X_1 = a_1, X_2 = a_2)$$

where a_1 and a_2 are values of X_1 and X_2 for a particular respondent

We classify $Y = 0$ if $P(Y = 0|X_1 = a_1, X_2 = a_2) > 0.5$ OR

$Y = 1$ if $P(Y = 1|X_1 = a_1, X_2 = a_2) > 0.5$

In the general case i.e. when Y has more than 2 categories we compare $P(Y = y_k \mid X)$ for all values of y_k and classify $Y = y_k$ for which $P(Y = y_k \mid X)$ is the maximum

Expected Output

Once the classification rule is applied the output can be shown as follows:

Case#	X1	X2	$P(Y=1/X_1,X_2)$	$P(Y=0/X_1,X_2)$	Y classified as
1	1	0	0.44	0.56	0
2	1	1	0.7	0.3	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
240	0	0	0.2	0.8	0

Advantages of Naive Bayes Method

- Classification rule is simple to understand.
- The method requires a small amount of training data to estimate the parameters necessary for classification.
- The evaluation of the classifier is quick and easy.
- The method can be a good alternative to logistic regression.

Limitations of Naive Bayes Method

- Assumption of conditional independence of the independent variables is highly impractical.
- In case of continuous independent variables the density function must be known or assumed to be normal.
- In case of categorical independent variables the probabilities cannot be calculated if the count in any conditional category is zero. For instance: If there are no respondents in the age group 25-30 yrs. then $P(X_1=0 \mid Y=1) = 0$

?

How to deal with such cases?
If a category has zero entries we replace 0 by 0.5/n (n = sample size) so that the probability expression does not reduce to zero.

Case Study – Modeling Loan Defaults

Background

- A bank possesses demographic and transactional data of its loan customers. If the bank has a model to predict defaulters it can help in loan disbursal decision making.

Objective

- To predict whether the customer applying for the loan will be a defaulter or not.

Available Information

- Sample size is 700
- Independent Variables: Age group, Years at current address, Years at current employer, Debt to Income Ratio, Credit Card Debts, Other Debts. The information on predictors was collected at the time of loan application process.
- Dependent Variable: Defaulter (=1 if defaulter ,0 otherwise). The status is observed after loan is disbursed.

Bank Loan Data

Independent Variables

Dependent Variable

SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTE
1	3	17	12	9.3	11.36	5.01	1
2	1	10	6	17.3	1.36	4	0

Column	Description	Type	Measurement	Possible Values
SN	Serial Number	numeric	-	-
AGE	Age Groups	Categorical	1(<28 years),2(28-40 years),3(>40 years)	3
EMPLOY	Number of years customer working at current employer	Continuous	-	Positive value
ADDRESS	Number of years customer staying at current address	Continuous	-	Positive value
DEBTINC	Debt to Income Ratio	Continuous	-	Positive value
CREDDEBT	Credit Card Debt	Continuous	-	Positive value
OTHDEBT	Other Debt	Continuous	-	Positive value
DEFAULTER	Whether customer defaulted on loan	Binary	1(Defaulter), 0(Non-Defaulter)	2

Logistic Regression in R

```
# Importing data and checking data structure
```

```
bankloan<-read.csv("BANK LOAN.csv",header=T)

str(bankloan)
```

```
# Output
```

```
> str(bankloan)
'data.frame':   700 obs. of  8 variables:
 $ SN      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ AGE     : int  3 1 2 3 1 3 2 3 1 2 ...
 $ EMPLOY  : int  17 10 15 15 2 5 20 12 3 0 ...
 $ ADDRESS : int  12 6 14 14 0 5 9 11 4 13 ...
 $ DEBTINC : num  9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
 $ CREDDEBT : num  11.36 1.36 0.86 2.66 1.79 ...
 $ OTHDEBT  : num  5.01 4 2.17 0.82 3.06 ...
 $ DEFAULTER: int  1 0 0 0 1 0 0 0 1 0 ...
```

```
bankloan$AGE<-factor(bankloan$AGE)

riskmodel<-glm(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT+OTHDEBT,
               family=binomial,data=bankloan)
```

glm() fits a generalised linear model. **family=binomial** ensures that a binary regression is used.

Model Summary

```
summary(riskmodel)
```

```
# Output
```

summary() generates model summary.

```
Call:
glm(formula = DEFAULTER ~ AGE + EMPLOY + ADDRESS + DEBTINC +
    CREDDEBT + OTHDEBT, family = binomial, data = bankloan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3495  -0.6601  -0.2974   0.2509   2.8583

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78821    0.26407  -2.985  0.00284 **
AGE2         0.25202    0.26651   0.946  0.34433
AGE3         0.62707    0.36056   1.739  0.08201 .
EMPLOY      -0.26172    0.03188  -8.211  < 2e-16 ***
ADDRESS     -0.09964    0.02234  -4.459 8.22e-06 ***
DEBTINC      0.08506    0.02212   3.845  0.00012 ***
CREDDEBT     0.56336    0.08877   6.347 2.20e-10 ***
OTHDEBT      0.02315    0.05709   0.405  0.68517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 553.41  on 692  degrees of freedom
AIC: 569.41

Number of Fisher Scoring iterations: 6
```

Interpretation: EMPLOY, ADDRESS, DEBTINC and CREDDEBT are statistically significant.

Excluding Insignificant Variables

```
riskmodel<-glm(DEFAULTER~EMPLOY+ADDRESS+DEBTINC+CREDDEBT,
               family=binomial,data=bankloan)

summary(riskmodel)
```

```
# Output
```

```
Call:
glm(formula = DEFAULTER ~ EMPLOY + ADDRESS + DEBTINC + CREDDEBT,
    family = binomial, data = bankloan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4483  -0.6396  -0.3108   0.2583   2.8496

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79107    0.25154  -3.145  0.00166 **
EMPLOY      -0.24258    0.02806  -8.646  < 2e-16 ***
ADDRESS     -0.08122    0.01960  -4.144 3.41e-05 ***
DEBTINC      0.08827    0.01854   4.760 1.93e-06 ***
CREDDEBT     0.57290    0.08725   6.566 5.17e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 556.74  on 695  degrees of freedom
AIC: 566.74

Number of Fisher Scoring iterations: 6
```

Interpretation :
All four variables remain significant.

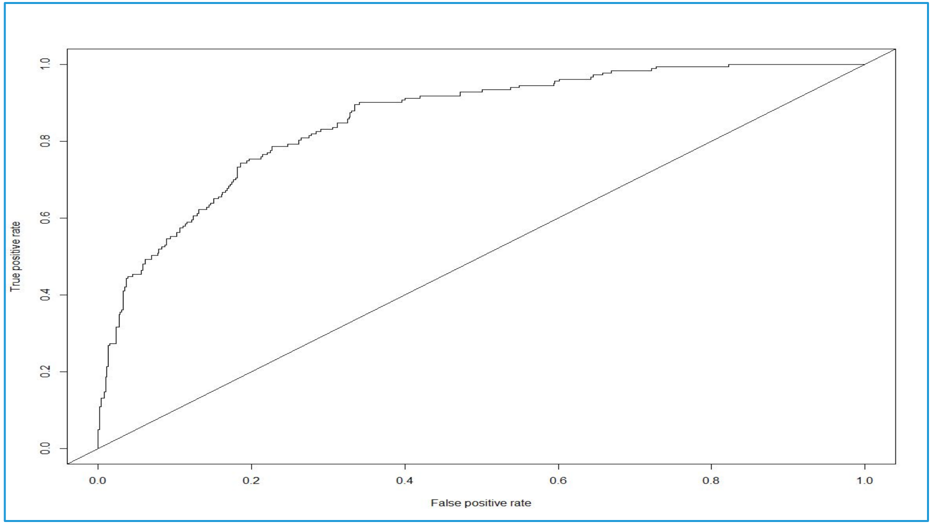
ROC Curve and Area Under ROC Curve

```
# ROC Curve
install.packages("ROCR")
library(ROCR)

bankloan$predprob<-fitted(riskmodel)

pred<-prediction(bankloan$predprob,bankloan$DEFAULTER)

perf<-performance(pred,"tpr","fpr")

plot(perf)

abline(0,1)
```

- **prediction()** function prepares data required for ROC curve.
- **performance()** function creates performance objects, "tpr“ (True positive rate), "fpr” (False positive rate).
- **plot()** function plots the objects created using performance
- **abline()** adds a straight line to the plot.

ROC Curve and Area Under ROC Curve

```
# Output
```




```
auc<-performance(pred,"auc")

auc@y.values
[[1]]
[1] 0.8556193
```

Estimates area under the ROC curve. Here it is 0.8556

Naive Bayes Method in R

```
# Install and load package "e1071".
# Model Fitting
```

```
install.packages("e1071")
library(e1071)

riskmodel2<-naiveBayes(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT+OTHDEBT,
                       data=bankloan)

riskmodel2
```

- **naiveBayes()** fits a Naive Bayes algorithm.
- It computes the conditional posterior probabilities of customer being defaulter/Non defaulter given values of independent variables using the Bayes rule.

Naive Bayes Model Output

```
# Output
```

```
> riskmodel2

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.7385714 0.2614286 

Conditional probabilities:
   AGE
Y           1         2         3
  0 0.3017408 0.4313346 0.2669246
  1 0.4699454 0.3333333 0.1967213

   EMPLOY
Y       [,1]     [,2]
  0 9.508704 6.663741
  1 5.224044 5.542946

   ADDRESS
Y       [,1]     [,2]
  0 8.945841 7.000621
  1 6.393443 5.925208

   DEBTINC
Y        [,1]     [,2]
  0  8.679304 5.615197
  1 14.727869 7.902798

   CREDDEBT
Y       [,1]     [,2]
  0 1.245397 1.422238
  1 2.423770 3.232645

   OTHDEBT
Y       [,1]     [,2]
  0 2.773230 2.813970
  1 3.863388 4.263394
```

Interpretation :

- Output shows a list of tables, one for each predictor variable. If the variable is categorical it shows the conditional probabilities for each class. For a numeric variable, for each target class, mean and standard deviation are shown.
- Eg. For EMPLOY, mean for "Defaulter" status = 0 is 9.51 and sd is 6.66.

Predicted Probabilities

```
# Predicted Probabilities
prednb<-predict(riskmodel2,bankloan,type='raw')

head(prednb)
```

- **predict()** returns predicted probabilities based on the model results and historical data.
- **type="raw"** returns raw probabilities. If not specified, predicted class is returned for each case

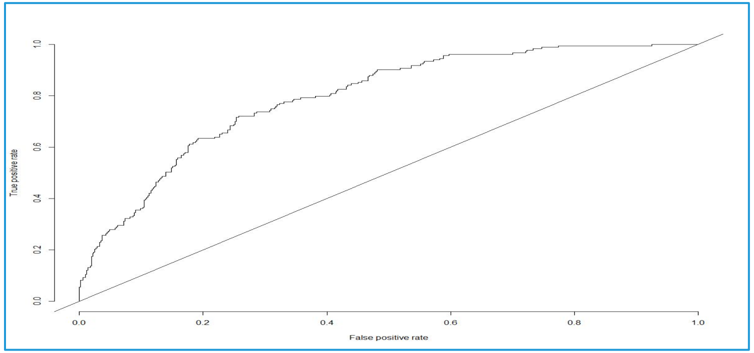
```
# Output
```

```
> head(prednb)
                0           1
[1,] 4.269360e-08 0.999999957
[2,] 7.182632e-01 0.281736806
[3,] 9.930388e-01 0.006961177
[4,] 9.885979e-01 0.011402100
[5,] 4.889496e-01 0.511050425
[6,] 9.177660e-01 0.082234006
```

Interpretation :
Column 2 gives probability of default (=1)

ROC Curve and Area Under ROC Curve

```
# ROC Curve and Area Under ROC Curve
pred<-prediction(prednb[,2],bankloan$DEFAULTER)
perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)
```




```
# Area Under ROC Curve
auc<-performance(pred,"auc")
auc@y.values
[[1]]
[1] 0.794971
```

*The column having probability of the event under study must be selected while creating the prediction object. In this case, we are predicting the likelihood of default and default is represented by 1, hence column index [,2] is taken.

Quick Recap

Conditional Probability and Bayes' Theorem

- The conditional probability of an event B is the probability that event B will occur given the knowledge that an event A has already occurred.
- P(B|A) = P(A|B) P(B) / P(A)

Naive Bayes Classifier

- To estimate Y given the values of X_i's or $P(Y|X_1, X_2, \ldots, X_m)$ using the Naïve Bayes Classifier.
- **Assumption:** All X_i's are conditionally independent of each other.
-

Naive Bayes in R

- `naiveBayes()` in package **`e1071`**