Naïve Bayes Classifier

Example:

This data source has description for every client.

Each client is described by four attributes of "ownership", "Marital Status (MS)", "Income" And "Bankruptcy (BR)".

The attribute BR is the class attribute And it has two class values of "Not Bankrupted" (N) and "Bankrupted" (Y).

Rec#	Ownership	MS	Income	BR
1	O	S	125K	N
2	R	M	100K	N
3	R	S	70K	N
4	O	M	120K	N
5	R	D	95K	Y
6	R	M	60K	N
7	O	D	220K	N
8	R	S	85K	Y
9	R	M	75K	N
10	R	S	90K	Y

X is the record of a new client whose class label is unknown

What is the class value for X?

If H is the hypothesis that X belongs to "Not Bankrupted" class, then

P(BR="N" | X) is the probability for X to be labeled as "Not Bankrupted"

If H is the hypothesis that X belongs to "Bankrupted" class, then

P(BR="Y" | X) is the probability for X to be labeled as "Bankrupted"

MAX[P(BR="N" | X), P(BR="Y" | X)]

determines the class for X.

P(BR="N" | X)= $\frac{P(X|BR="N")\,P(BR = "N")}{P(X)}$

Given X: (R, M. 120K)

Predict BR for X.

MAX [P(BR=“NO” | X), P(BR=“YES” | X)] determines the class for X.

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

X: (R, M. 120K)

$$P(BR=\text{"NO"} \mid X)= \frac{P(BR=\text{"NO"}) \prod_{i=1}^{d} P(A_i=v \mid BR=\text{"NO"})}{P(X)}$$

$$P(BR=\text{"YES"} \mid X)= \frac{P(BR=\text{"YES"}) \prod_{i=1}^{d} P(A_i=v \mid BR=\text{"YES"})}{P(X)}$$

Where:

d is the number of attributes in D

A_i is the i-th attribute of X

v is the value for A_i.

$\prod_{i=1}^{d} P(A_i = v \mid BR = \text{"NO"}) =$
P(Ownership="R" | BR= "NO") *
P(MS="M" | BR= "NO") *
P(Income="120K" | BR= "NO") =

4/7

Out of the 7 records with BR="NO",
4 of them have the Ownership= "R"

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

X: (R, M. 120K)

P(BR="NO" | X)= $\frac{P(BR="NO")\prod_{i=1}^{d} P(A_i=v\,|BR="NO")}{P(X)}$

P(BR="YES" | X)= $\frac{P(BR="YES")\prod_{i=1}^{d} P(A_i=v\,|BR="YES")}{P(X)}$

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

Where:

d is the number of attributes in D

A_i is the i-th attribute of X

v is the value for A_i.

$$\prod_{i=1}^{d} P(A_i = v\,|BR = "NO")$$

P(Ownership="R" | BR= "NO") *
P(MS="M" | BR= "NO") *
P(Income="120K" | BR= "NO") =
4/7 * 4/7 * ? =

The last Probability deals with the

Attribute Income that has continuous values. Let us calculate that probability and then comeback to this slide.

$$P(A_i=\text{"v"} \mid BR=\text{"NO"}) = \frac{1}{\sqrt{2\pi}\,\sigma} * e^{-\frac{(v-\bar{a})^2}{2\sigma^2}}$$

Where, $\bar{a}$ and standard deviation, σ, are calculated for INCOME values of all the records in D with BR = "NO"

$\bar{a} =$

(125+100+70+120+60+220+75+90)/7= 110

$\sigma^2 = \frac{\sum(x_i-\bar{a})^2}{n(n-1)}$ =[(125-110)^2+(100-110)^2+ (70-110)^2+(120-110)^2+(60-110)^2+(220-110)^2+ (75-110)^2+(90-110)^2] / 7*6 = 2975

σ = 54.54

P(Income="120K" | BR= "NO") =

$$\frac{1}{\sqrt{2\pi}\,54.54} * e^{-\frac{(120-110)^2}{2*2975}} = 0.0072$$

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

X: (R, M. 120K)

$$P(BR=\text{"NO"} \mid X) = \frac{P(BR=\text{"NO"}) \prod_{i=1}^{d} P(A_i=v \mid BR=\text{"NO"})}{P(X)}$$

$$P(BR=\text{"YES"} \mid X) = \frac{P(BR=\text{"YES"}) \prod_{i=1}^{d} P(A_i=v \mid BR=\text{"YES"})}{P(X)}$$

Where:

d is the number of attributes in D

A_i is the i-th attribute of X

v is the value for A_i.

$$\prod_{i=1}^{d} P(A_i = v \mid BR = \text{"NO"})$$

P(Ownership="R" | BR= "NO") *
P(MS="M" | BR= "NO") *
P(Income="120K" | BR= "NO") =
4/7 * 4/7 * ? =

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

X: (R, M. 120K)

P(BR="NO" | X)= $\frac{P(BR=\text{"NO"})\prod_{i=1}^{d} P(A_i=v\,|BR=\text{"NO"})}{P(X)}$

P(BR="YES" | X)= $\frac{P(BR=\text{"YES"})\prod_{i=1}^{d} P(A_i=v\,|BR=\text{"YES"})}{P(X)}$

Where:

d is the number of attributes in D

A_i is the i-th attribute of X

v is the value for A_i.

$$\prod_{i=1}^{d} P(A_i = v\,|BR = \text{"NO"})$$

P(Ownership="R" | BR= "NO") *
P(MS="M" | BR= "NO") *
P(Income="120K" | BR= "NO") =
4/7 * 4/7 * 0.0072 = 0.0024

$P(BR = \text{"NO"})$ = 7/10

P(BR="NO" | X)= (0.7*0.0024)/P(x)

We ignore P(X) because both probability is divided by the same value

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

X: (R, M. 120K)

$$P(BR=\text{"NO"} \mid X) = \frac{P(BR=\text{"NO"}) \prod_{i=1}^{d} P(A_i=v \mid BR=\text{"NO"})}{P(X)}$$

$$P(BR=\text{"YES"} \mid X) = \frac{P(BR=\text{"YES"}) \prod_{i=1}^{d} P(A_i=v \mid BR=\text{"YES"})}{P(X)}$$

Where:

d is the number of attributes in D

A_i is the i-th attribute of X

v is the value for A_i.

$$\prod_{i=1}^{d} P(A_i = v \mid BR = \text{"NO"})$$

P(Ownership="R" | BR= "NO") *

P(MS="M" | BR= "NO") *

P(Income="120K" | BR= "NO") =

4/7 * 4/7 * 0.0072 = 0.0024

$P(BR = \text{"NO"})$ = 7/10

P(BR="NO" | X)= (0.7*0.0024)/P(x)

We ignore P(X) because both probability is divided by the same value

P(BR="NO" | X)= (0.7*0.0024)= 0.00168 P(BR="YES" | X)= 0

Dataset D

Rec#	Ownership	MS	Income	BR
1	O	S	125K	NO
2	R	M	100K	NO
3	R	S	70K	NO
4	O	M	120K	NO
5	R	D	95K	YES
6	R	M	60K	NO
7	O	D	220K	NO
8	R	S	85K	YES
9	R	M	75K	NO
10	R	S	90K	YES

Winner is BR="NO"