COMP 3605

Naïve Bayes Classifier

- The Naive Bayes classifier is a family of probabilistic classifiers based on applying Bayes' theorem with the assumption of independence between every pair of features.

- Despite its simplicity, Naive Bayes can be surprisingly accurate and is particularly useful for very large datasets.

- It's often used for text classification, spam filtering, sentiment analysis, and recommendation systems.

Bayes' Theorem:

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

Where:

- $P(A|B)$ is the posterior probability of class A given predictor B.
- $P(A)$ is the prior probability of class A.
- $P(B|A)$ is the likelihood which is the probability of predictor B given class A.
- $P(B)$ is the prior probability of predictor B.

EXAMPLE 1

.

TID	*age*	*income*	*student*	*credit_rating*	*buys_computer?*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Predict the class label for this new tuple:

X = (age = youth, income = medium, student = yes, credit_rating = fair)

i.e., X = (x1 = youth, x2 = medium, x3 = yes, x4 = fair)

	buys_computer	
age	yes	no
youth	2	3
middle_aged	4	0
senior	3	2

	buys_computer	
income	yes	no
low	3	1
medium	4	2
high	2	2

	buys_computer	
student	yes	no
yes	6	1
no	3	4

	buys_computer	
credit_ratting	yes	no
fair	6	2
excellent	3	3

1. **Calculate Prior Probabilities**:

   Calculate the probability of each class in the class_buy_computer column.

P(Yes) = 9/14

P(No) = 5/14

2. **Calculate Likelihoods:**
   For each feature (age, income, student, credit_rating), calculate the likelihood of that feature given each class.

   For Buy_Computer = Yes:

   P(Age = Youth | Yes) = 2/9
   P(Income = Medium | Yes) = 4/9
   P(Student = Yes | Yes) = 6/9
   P(Credit_Rating = Fair | Yes) = 6/9

   For Buy_Computer = No:

   P(Age = Youth | No) = 3/5
   P(Income = Medium | No) = 2/5
   P(Student = Yes | No) = 1/5
   P(Credit_Rating = Fair | No) = 2/5

3. **Calculate Posterior Probabilities:**

For the given example X, calculate the probability of each class given X by multiplying the prior probabilities by the likelihoods.

$$\dot{P}(\boldsymbol{X} \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

Compute P(X | yes) = P(age = youth | yes) *P(income = medium | yes)* P(student = yes | yes) *P(credit_rating = fair | yes)

= 2/9 x 4/9 x 6/9 x 6/9

Compute P(X | no) = P(age = youth | no) *P(income = medium | no)* P(student = yes | no) *P(credit_rating = fair | no)

= (3/5) * (2/5) * (1/5) * (2/5)

To find the class Ci that maximizes P(X | Ci)P(Ci), we compute:

P(Yes | X) = P(Yes) * P(Age = Youth | Yes) * P(Income = Medium | Yes) * P(Student = Yes | Yes) * P(Credit_Rating = Fair | Yes)

P(No | X) = P(No) * P(Age = Youth | No) * P(Income = Medium | No) * P(Student = Yes | No) * P(Credit_Rating = Fair | No)

Substitute the probabilities calculated above:

P(Yes | X) = (9/14) * (2/9) * (4/9) * (6/9) * (6/9) = 0.0282

P(No | X) = (5/14) * (3/5) * (2/5) * (1/5) * (2/5) = 0.0069

4.Prediction: The class with the highest posterior probability will be the prediction.

Since P(Yes | X) > P(No | X), the Naive Bayes Classifier predicts that the computer will be bought.(yes)

EXAMPLE 2:

The "Play Tennis" dataset is a well-known toy dataset often used to illustrate the concepts of classification algorithms. The dataset consists of 14 instances, each representing a day with corresponding weather conditions and whether or not tennis was played on that day.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Let’s say we have a new instance with the following features:

Outlook: Sunny

Temperature: Mild

Humidity: High

Wind: Strong

We want to predict whether tennis will be played under these conditions (Play Tennis: Yes/No).

Temperature	Yes	No
Hot	2	2
Mild	4	2
Cool	3	1

Humidity	Yes	No
High	3	4
Normal	6	1

Outlook	Yes	No
Sunny	2	3
Overcast	4	0
Rain	3	2

Wind	Yes	No
Strong	3	3
Weak	6	2

Predict a class label using the Naive Bayes Classifier for a new instance with the following features:

- Outlook: Sunny
- Temperature: Mild
- Humidity: High
- Wind: Strong

We want to predict whether tennis will be played under these conditions (Play Tennis: Yes/No).

1. **Calculate Prior Probabilities**

P(Yes) = 9/14 (9 days of playing tennis out of 14)

P(No) = 5/14 (5 days of not playing tennis out of 14)

2. Calculate Likelihoods

For the class label Yes:

P(Outlook=Sunny|Yes) = 2/9 (2 sunny days out of 9 days of playing tennis)

P(Temperature=Mild|Yes) = 4/9

P(Humidity=High|Yes) = 3/9

P(Wind=Strong|Yes) = 3/9

For the class label No:

P(Outlook=Sunny|No) = 3/5 (3 sunny days out of 5 days of not playing tennis)

P(Temperature=Mild|No) = 2/5

P(Humidity=High|No) = 4/5

P(Wind=Strong|No) = 3/5

2. **Calculate Posterior Probabilities**

$$\dot{P}(\boldsymbol{X} \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

P(X|Yes) = P(Outlook=Sunny|Yes) * P(Temperature=Mild|Yes) *
P(Humidity=High|Yes) * P(Wind=Strong|Yes)

P(X|Yes) = P (2/9) * (4/9) * (3/9) * (3/9)

P(X|No) = P(Outlook=Sunny|No) * P(Temperature=Mild|No) *
P(Humidity=High|No) * P(Wind=Strong|No)

P(X|No) = (3/5) * (2/5) * (4/5) * (3/5)

For the given example X = (Outlook=Sunny, Temperature=Mild, Humidity=High, Wind=Strong), calculate the posterior probabilities:

P(Yes|X) = P(Yes) * P(Outlook=Sunny|Yes) * P(Temperature=Mild|Yes) * P(Humidity=High|Yes) * P(Wind=Strong|Yes)

P(No|X) = P(No) * P(Outlook=Sunny|No) * P(Temperature=Mild|No) * P(Humidity=High|No) * P(Wind=Strong|No)

Substitute the calculated probabilities:

P(Yes|X) = (9/14) * (2/9) * (4/9) * (3/9) * (3/9) = 0.0053

P(No|X) = (5/14) * (3/5) * (2/5) * (4/5) * (3/5) = 0.0206

3. **Prediction**: The class with the highest posterior probability will be the prediction.

Since P(No | X) > P(Yes| X), the Naive Bayes Classifier predicts that tennis will not be played under the given conditions (Play Tennis: No).

EXERCISE 1:

Dataset:

Age	Income	Gender	Owns House	Buy Product
Young	High	M	N	No
Young	High	M	N	No
Middle	High	M	Y	Yes
Old	Medium	F	Y	Yes
Old	Low	F	Y	No
Old	Low	F	N	Yes
Middle	Low	F	Y	Yes
Young	Medium	M	N	No
Young	Low	F	Y	Yes
Old	Medium	F	Y	Yes
Young	Medium	F	Y	Yes
Middle	Medium	M	Y	Yes
Middle	High	F	Y	Yes
Old	Medium	M	N	No
Middle	Low	F	N	Yes

Predict if an individual with the following attributes will buy the product:

Age: Middle

Income: Low

Gender: F

Owns House: N