

Data Mining Classification: Alternative Techniques

Bayesian Classifiers

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

1

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:
$$P(Y | X) = \frac{P(X, Y)}{P(X)}$$

$$P(X | Y) = \frac{P(X, Y)}{P(Y)}$$
- Bayes theorem:

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

2

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables
- Given a record with attributes (X_1, X_2, \dots, X_d), the goal is to predict class Y

— Specifically, we want to find the value of Y that maximizes $P(Y|X_1, X_2, \dots, X_d)$

- Can we estimate $P(Y|X_1, X_2, \dots, X_d)$ directly from data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

2/08/2021

Introduction to Data Mining, 2nd Edition

3

3

Using Bayes Theorem for Classification

- Approach:

— compute posterior probability $P(Y | X_1, X_2, \dots, X_d)$ using the Bayes theorem

$$P(Y | X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_d | Y)P(Y)}{P(X_1, X_2, \dots, X_d)}$$

— *Maximum a-posteriori*: Choose Y that maximizes $P(Y | X_1, X_2, \dots, X_d)$

— Equivalent to choosing value of Y that maximizes $P(X_1, X_2, \dots, X_d | Y)P(Y)$

- How to estimate $P(X_1, X_2, \dots, X_d | Y)$?

2/08/2021

Introduction to Data Mining, 2nd Edition

4

4

Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- We need to estimate

$$P(\text{Evade} = \text{Yes} | X) \text{ and } P(\text{Evade} = \text{No} | X)$$

In the following we will replace

Evade = Yes by Yes, and

Evade = No by No

2/08/2021

Introduction to Data Mining, 2nd Edition

5

Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Using Bayes Theorem:

$$\square P(\text{Yes} | X) = \frac{P(X | \text{Yes})P(\text{Yes})}{P(X)}$$

$$\square P(\text{No} | X) = \frac{P(X | \text{No})P(\text{No})}{P(X)}$$

\square How to estimate $P(X | \text{Yes})$ and $P(X | \text{No})$?

2/08/2021

Introduction to Data Mining, 2nd Edition

6

6

Conditional Independence

- X and Y are conditionally independent given Z if $P(X|YZ) = P(X|Z)$
- Example: Arm length and reading skills
 - Young child has shorter arm length and limited reading skills, compared to adults
 - If age is fixed, no apparent relationship between arm length and reading skills
 - Arm length and reading skills are conditionally independent given age

2/08/2021

Introduction to Data Mining, 2nd Edition

7

Naïve Bayes Classifier

- Assume independence among attributes X_i when class is given:
 - $P(X_1, X_2, \dots, X_d | Y_j) = P(X_1 | Y_j) P(X_2 | Y_j) \dots P(X_d | Y_j)$
 - Now we can estimate $P(X_i | Y_j)$ for all X_i and Y_j combinations from the training data
 - New point is classified to Y_j if $P(Y_j) \prod P(X_i | Y_j)$ is maximal.

2/08/2021

Introduction to Data Mining, 2nd Edition

8

8

Naïve Bayes on Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120K)$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(X | \text{Yes}) =$$

$$\begin{aligned} P(\text{Refund} = \text{No} | \text{Yes}) \times \\ P(\text{Divorced} | \text{Yes}) \times \\ P(\text{Income} = 120K | \text{Yes}) \end{aligned}$$

$$P(X | \text{No}) =$$

$$\begin{aligned} P(\text{Refund} = \text{No} | \text{No}) \times \\ P(\text{Divorced} | \text{No}) \times \\ P(\text{Income} = 120K | \text{No}) \end{aligned}$$

2/08/2021

Introduction to Data Mining, 2nd Edition

9

9

Estimate Probabilities from Data

- $P(y) =$ fraction of instances of class y

- e.g., $P(\text{No}) = 7/10$,
 $P(\text{Yes}) = 3/10$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- For categorical attributes:

$$P(X_i = c | y) = n_c / n$$

- where $|X_i = c|$ is number of instances having attribute value $X_i = c$ and belonging to class y

- Examples:

$$\begin{aligned} P(\text{Status}=\text{Married} | \text{No}) &= 4/7 \\ P(\text{Refund}=\text{Yes} | \text{Yes}) &= 0 \end{aligned}$$

2/08/2021

Introduction to Data Mining, 2nd Edition

10

10

Estimate Probabilities from Data

- For continuous attributes:
 - **Discretization:** Partition the range into bins:
 - ◆ Replace continuous value with bin value
 - Attribute changed from continuous to ordinal
 - **Probability density estimation:**
 - ◆ Assume attribute follows a normal distribution
 - ◆ Use data to estimate parameters of distribution (e.g., mean and standard deviation)
 - ◆ Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

2/08/2021

Introduction to Data Mining, 2nd Edition

11

11

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(X_i | Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- One for each (X_i, Y_j) pair

- For (Income, Class=No):

- If Class=No
 - ◆ sample mean = 110
 - ◆ sample variance = 2975

$$P(\text{Income} = 120 | \text{No}) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

2/08/2021

Introduction to Data Mining, 2nd Edition

12

12

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120K)$$

Naïve Bayes Classifier:

$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$
 $P(\text{Refund} = \text{No} | \text{No}) = 4/7$
 $P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$
 $P(\text{Refund} = \text{No} | \text{Yes}) = 1$
 $P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$
 $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$
 $P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$
 $P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$
 $P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$
 $P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$

For Taxable Income:

If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- $P(X | \text{No}) = P(\text{Refund} = \text{No} | \text{No}) \times P(\text{Divorced} | \text{No}) \times P(\text{Income} = 120K | \text{No}) = 4/7 \times 1/7 \times 0.0072 = 0.0006$
- $P(X | \text{Yes}) = P(\text{Refund} = \text{No} | \text{Yes}) \times P(\text{Divorced} | \text{Yes}) \times P(\text{Income} = 120K | \text{Yes}) = 1 \times 1/3 \times 1.2 \times 10^{-9} = 4 \times 10^{-10}$

Since $P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

Therefore $P(\text{No}|X) > P(\text{Yes}|X)$
=> Class = No

2/08/2021

Introduction to Data Mining, 2nd Edition

13

13

Naïve Bayes Classifier can make decisions with partial information about attributes in the test record

Even in absence of information about any attributes, we can use Apriori Probabilities of Class Variable:

Naïve Bayes Classifier:

$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$
 $P(\text{Refund} = \text{No} | \text{No}) = 4/7$
 $P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$
 $P(\text{Refund} = \text{No} | \text{Yes}) = 1$
 $P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$
 $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$
 $P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$
 $P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$
 $P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$
 $P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$

$$P(\text{Yes}) = 3/10$$

$$P(\text{No}) = 7/10$$

If we only know that marital status is Divorced, then:

$$P(\text{Yes} | \text{Divorced}) = 1/3 \times 3/10 / P(\text{Divorced})$$

$$P(\text{No} | \text{Divorced}) = 1/7 \times 7/10 / P(\text{Divorced})$$

If we also know that Refund = No, then

$$P(\text{Yes} | \text{Refund} = \text{No}, \text{Divorced}) = 1 \times 1/3 \times 3/10 / P(\text{Divorced, Refund} = \text{No})$$

$$P(\text{No} | \text{Refund} = \text{No}, \text{Divorced}) = 4/7 \times 1/7 \times 7/10 / P(\text{Divorced, Refund} = \text{No})$$

If we also know that Taxable Income = 120, then

$$P(\text{Yes} | \text{Refund} = \text{No}, \text{Divorced, Income} = 120) = 1.2 \times 10^{-9} \times 1 \times 1/3 \times 3/10 / P(\text{Divorced, Refund} = \text{No}, \text{Income} = 120)$$

$$P(\text{No} | \text{Refund} = \text{No}, \text{Divorced, Income} = 120) = 0.0072 \times 4/7 \times 1/7 \times 7/10 / P(\text{Divorced, Refund} = \text{No}, \text{Income} = 120)$$

2/08/2021

Introduction to Data Mining, 2nd Edition

14

14

Issues with Naïve Bayes Classifier

Given a Test Record:

$X = (\text{Married})$

Naïve Bayes Classifier:

$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$
 $P(\text{Refund} = \text{No} | \text{No}) = 4/7$
 $P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$
 $P(\text{Refund} = \text{No} | \text{Yes}) = 1$
 $P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$
 $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$
 $P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$
 $P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$
 $P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$
 $P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$

$$P(\text{Yes}) = 3/10$$

$$P(\text{No}) = 7/10$$

$$P(\text{Yes} | \text{Married}) = 0 \times 3/10 / P(\text{Married})$$

$$P(\text{No} | \text{Married}) = 4/7 \times 7/10 / P(\text{Married})$$

For Taxable Income:

If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

2/08/2021

Introduction to Data Mining, 2nd Edition

15

15

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7				
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

$P(\text{Refund} = \text{Yes} | \text{No}) = 2/6$
 $P(\text{Refund} = \text{No} | \text{No}) = 4/6$
 $\rightarrow P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$
 $P(\text{Refund} = \text{No} | \text{Yes}) = 1$
 $\rightarrow P(\text{Marital Status} = \text{Single} | \text{No}) = 2/6$
 $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 0$
 $P(\text{Marital Status} = \text{Married} | \text{No}) = 4/6$
 $P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$
 $P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$
 $P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0/3$

For Taxable Income:

If class = No: sample mean = 91
sample variance = 685
If class = Yes: sample mean = 90
sample variance = 25

Given $X = (\text{Refund} = \text{Yes}, \text{Divorced}, 120\text{K})$

$$P(X | \text{No}) = 2/6 \times 0 \times 0.0083 = 0$$

$$P(X | \text{Yes}) = 0 \times 1/3 \times 1.2 \times 10^{-9} = 0$$

Naïve Bayes will not be able to
classify X as Yes or No!

2/08/2021

Introduction to Data Mining, 2nd Edition

16

16

Issues with Naïve Bayes Classifier

- If one of the conditional probabilities is zero, then the entire expression becomes zero
- Need to use other estimates of conditional probabilities than simple fractions
- Probability estimation:

original: $P(X_i = c|y) = \frac{n_c}{n}$

n : number of training instances belonging to class y

n_c : number of instances with $X_i = c$ and $Y = y$

v : total number of attribute values that X_i can take

p : initial estimate of $(P(X_i = c|y))$ known apriori

m : hyper-parameter for our confidence in p

Laplace Estimate: $P(X_i = c|y) = \frac{n_c + 1}{n + v}$

$m -$ estimate: $P(X_i = c|y) = \frac{n_c + mp}{n + m}$

2/08/2021

Introduction to Data Mining, 2nd Edition

17

17

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

$$P(A|M)P(M) > P(A|N)P(N)$$

=> Mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

2/08/2021

Introduction to Data Mining, 2nd Edition

18

18

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Redundant and correlated attributes will violate class conditional assumption
 - Use other techniques such as Bayesian Belief Networks (BBN)

2/08/2021

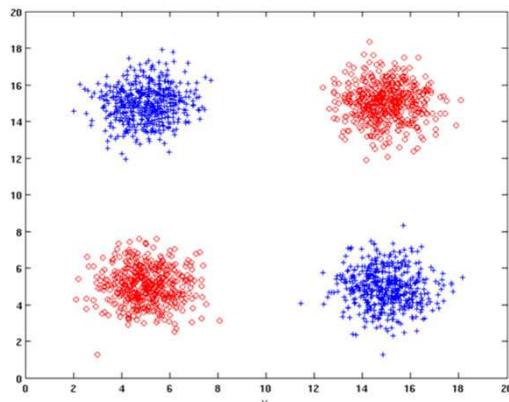
Introduction to Data Mining, 2nd Edition

19

19

Naïve Bayes

- How does Naïve Bayes perform on the following dataset?



Conditional independence of attributes is violated

2/08/2021

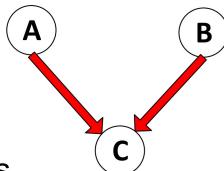
Introduction to Data Mining, 2nd Edition

20

20

Bayesian Belief Networks

- Provides graphical representation of probabilistic relationships among a set of random variables
- Consists of:
 - A directed acyclic graph (dag)
 - ◆ Node corresponds to a variable
 - ◆ Arc corresponds to dependence relationship between a pair of variables
 - A probability table associating each node to its immediate parent



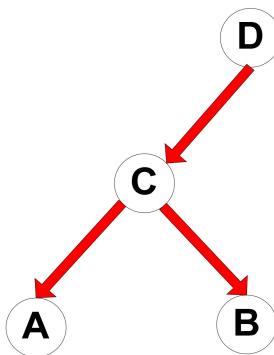
2/08/2021

Introduction to Data Mining, 2nd Edition

21

21

Conditional Independence



D is parent of C

A is child of C

B is descendant of D

D is ancestor of A

- A node in a Bayesian network is conditionally independent of all of its nondescendants, if its parents are known

2/08/2021

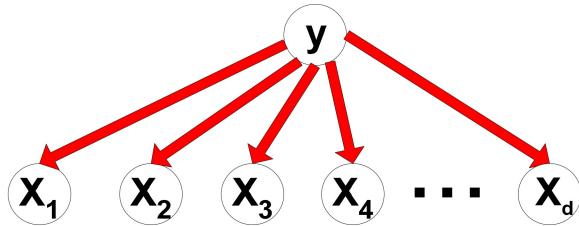
Introduction to Data Mining, 2nd Edition

22

22

Conditional Independence

- Naïve Bayes assumption:



2/08/2021

Introduction to Data Mining, 2nd Edition

23

23

Probability Tables

- If X does not have any parents, table contains prior probability $P(X)$
- If X has only one parent (Y), table contains conditional probability $P(X|Y)$
- If X has multiple parents (Y_1, Y_2, \dots, Y_k), table contains conditional probability $P(X|Y_1, Y_2, \dots, Y_k)$



2/08/2021

Introduction to Data Mining, 2nd Edition

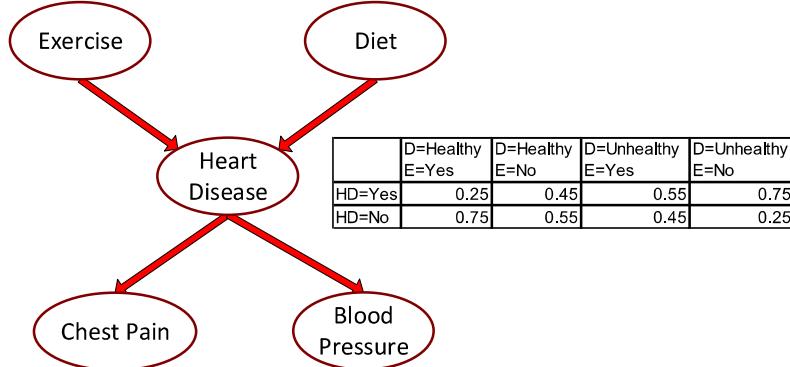
24

24

Example of Bayesian Belief Network

Exercise=Yes	0.7
Exercise>No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



2/08/2021

Introduction to Data Mining, 2nd Edition

25

Example of Inferencing using BBN

- Given: $X = (E=No, D=Yes, CP=Yes, BP=High)$
 - Compute $P(HD|E,D,CP,BP)$?
- $P(HD=Yes| E=No,D=Yes) = 0.55$
 $P(CP=Yes| HD=Yes) = 0.8$
 $P(BP=High| HD=Yes) = 0.85$
 - $P(HD=Yes| E=No,D=Yes,CP=Yes,BP=High)$
 $\propto 0.55 \times 0.8 \times 0.85 = 0.374$
- $P(HD=No| E=No,D=Yes) = 0.45$
 $P(CP=Yes| HD=No) = 0.01$
 $P(BP=High| HD=No) = 0.2$
 - $P(HD=No| E=No,D=Yes,CP=Yes,BP=High)$
 $\propto 0.45 \times 0.01 \times 0.2 = 0.0009$

Classify X as Yes

2/08/2021

Introduction to Data Mining, 2nd Edition

26

26