Data Mining
Classification: Alternative Techniques

Bayesian Classifiers

Introduction to Data Mining

by

Tan, Steinbach, Karpatne, Kumar

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$$

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

- Bayes theorem:

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

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?
  - Let us formulate this question as a math formula.
  - Let the event of having a stiff neck be S. Let the event of having meningitis be M, what formula can describe this question?

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables

- Given a record with attributes (X_1, X_2,…, X_d)
  - Goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes P(Y| X_1, X_2,…, X_d)

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

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

- Can we estimate P(Evade = Yes | X) and P(Evade = No | X)?

In the following we will replace

“Evade = Yes” by Yes, and

“Evade = No” by No

Using Bayes Theorem for Classification

- Approach:
  - compute posterior probability $P(Y \mid X_1, X_2, \ldots, X_d)$ using the Bayes theorem

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

  - *Maximum a-posteriori*: Choose Y that maximizes
    $P(Y \mid X_1, X_2, \ldots, X_d)$ (e.g. Y=yes or Y=no)

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

- How to estimate $P(X_1, X_2, \ldots, X_d \mid Y)$?

Using Bayes Theorem for Classification

$$\boxed{P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}}$$

- *Maximum a-posteriori*: Choose Y that maximizes

$$P(Y \mid X_1, X_2, \ldots, X_d)$$

- E.g. P(Y=iris Versicolor| petal length=2, petal width=1.5, sepal width=1.4, sepal length=3) =?

- P(Y=iris Setosa| petal length=2, petal width=1.5, sepal width=1.4, sepal length=3) =?
- P(Y=iris Virginica| petal length=2, petal width=1.5, sepal width=1.4, sepal length=3) =?

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:

- $$P(\text{Yes} \mid X) = \frac{P(X \mid \text{Yes})P(\text{Yes})}{P(X)}$$

- $$P(\text{No} \mid X) = \frac{P(X \mid \text{No})P(\text{No})}{P(X)}$$

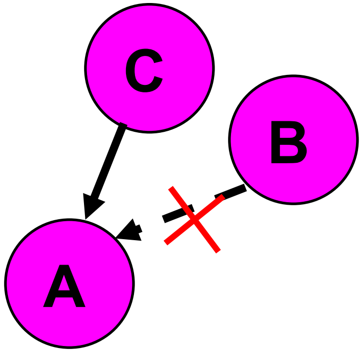
- How to estimate P(X | Yes) and P(X | No)?

How to compute $P(X_1, X_2, \ldots, X_d \mid Y)$?

- $$P(X_1X_2|Y) = \frac{P(X_1X_2Y)}{P(Y)} = \frac{P(X_1|X_2Y)P(X_2Y)}{P(Y)}$$
$$= P(X_1|X_2Y)\,P(X_2|Y)$$
$$= P(X_1|Y)\,P(X_2|Y)$$

$$P(X_1X_2X_3|Y) = P(X_1|X_2X_3Y)P(X_2|X_3Y)P(X_3|Y)$$ [how to prove?]

$$= P(X_1|Y)P(X_2|Y)P(X_3|Y)$$



A and **B** are conditionally independent given **C** if P(**A**|**BC**) = P(**A**|**C**)

Assumption: all attributes (X_1 to X_3) are independent

How to compute $P(X_1, X_2, \ldots, X_d \mid Y)$?

<u>Chain rule</u>:

$P(X1, X2, \ldots, Xd, Y)$

$= P(X1 | X2, \ldots, Xd, Y)\ P(X2, X3, \ldots, Xd, Y)$

$= P(X1 | X2, \ldots, Xd, Y)\ P(X2 | X3, \ldots, Xd, Y)\ P(X3, \ldots, Xd, Y)$

$= P(X1 | X2, \ldots, Xd, Y)\ P(X2 | X3, \ldots, Xd, Y)\ P(X3 | X4, \ldots, Xd, Y)\ P(X4, \ldots, Xd, Y)$

$= P(X1 | X2, \ldots, Xd, Y)\ P(X2 | X3, \ldots, Xd, Y)\ P(X3 | X4, \ldots, Xd, Y)\ P(X4 | X5, \ldots, Xd, Y)\ P(X5, \ldots Y)$

.

.

$= P(X1 | X2, \ldots, Xd, Y)\ P(X2 | X3, \ldots, Xd, Y)\ P(X3 | X4, \ldots, Xd, Y) \ldots P(Xd|Y)\ P(Y)$

Thus

$P(X1, X2, \ldots, Xd | Y) = P(X2, X2, \ldots, Xd, Y) / P(Y)$

If X1, X2, …, Xd are all conditionally independent of each other

$P(X1, X2, \ldots, Xd | Y) = P(X1|Y)\ P(X2|Y) \ldots P(Xd|Y)$

Conditional independence

- E.g. one might observe that people with longer arms tend to have better reading skills. So, the two attributes: arm length and reading skill are dependent?
- If add the age as a new attribute, adults → longer arms, adults → better reading skill. The dependence between arm length and reading skill disappears.
- Thus Pr(reading skill|age, arm length)= Pr(reading skill |age).

Naïve Bayes Classifier

- Assume independence among attributes X_i when class is given:

$$P(X_1, X_2, \ldots, X_d \mid Y_j) = P(X_1 \mid X_2, \ldots, X_d, Y_j)\, P(X_2 \mid \ldots, X_d, Y_j) \ldots P(X_3 \mid \ldots, X_d, Y_j) = P(X_1 \mid Y_j)\, P(X_2 \mid Y_j) \ldots P(X_d \mid Y_j)$$

  – Now we can estimate $P(X_i \mid Y_j)$ for all X_i and Y_j combinations from the training data

  – **New point is classified to Y_j if $P(Y_j) \prod P(X_i \mid Y_j)$ is maximal.**

Naïve Bayes on 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

- P(X | Yes) =

  P(Refund = No | Yes) x

  P(Divorced | Yes) x

  P(Income = 120K | Yes)

- P(X | No) =

  P(Refund = No | No) x

  P(Divorced | No) x

  P(Income = 120K | No)

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

- Class: $P(Y) = N_c/N$
  - e.g., $P(No) = 7/10$,
    $P(Yes) = 3/10$

- For categorical attributes:

  $$P(X_i \mid Y_k) = |X_{ik}| / N_{c_k}$$

  - where $|X_{ik}|$ is number of instances having attribute value X_i and belonging to class Y_k
  - Examples:

P(Status=Married | No) = 4/7
P(Refund=Yes | Yes)=0

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 $P(X_i|Y)$

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:

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

  – One for each (X_i, Y_i) pair

- For (Income, Class=No):
  – If Class=No
    - sample mean = 110
    - sample variance = 2975

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

Example of Naïve Bayes Classifier

Given a Test Record:

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

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | 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 | No) = P(Refund=No | No) × P(Divorced | No) × P(Income=120K | No) = 4/7 × 1/7 × 0.0072 = 0.0006

- P(X | Yes) = P(Refund=No | Yes) × P(Divorced | Yes) × P(Income=120K | Yes) = 1 × 1/3 × 1.2 × 10^{-9} = 4 × 10^{-10}

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

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

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

- P(Yes) = 3/10
  P(No) = 7/10

- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0
P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

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

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:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of different values for attribute A (e.g. c=2 for Refund, c=3 for marital status)

p: probability of different an attribute value

m: parameter

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Use +1 pseudocount

Given X = (Refund = Yes, Divorced, 120K)

$P(X \mid No) = (2+1)/(6+2) \times (0+1)/(6+3) \times .. > 0$

$P(X \mid Yes) = 1/(3+2) \times (1+1)/(3+3) \times \ldots > 0$

Given X = (Refund = Yes, Divorced, 120K)

$P(X \mid No) = 2/6 \times 0 \times 0.0083 = 0$

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

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

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

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

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

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

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

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

=> Mammals

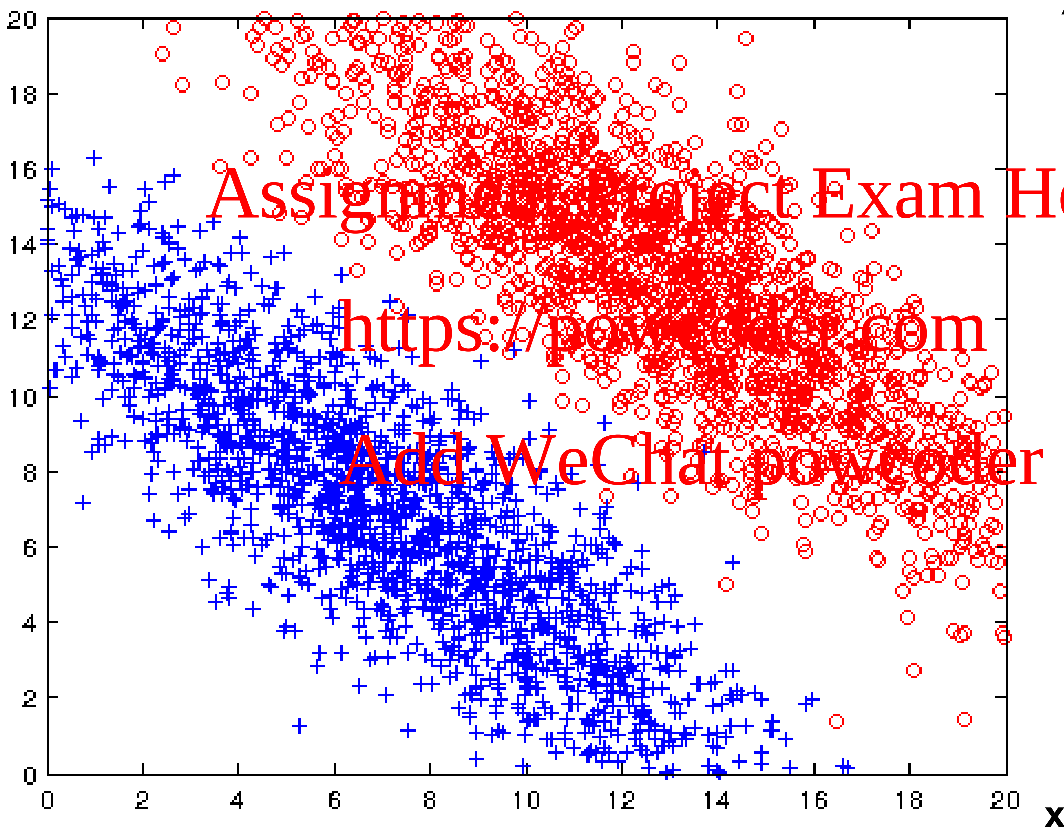
In-class exercise (submit your equations)

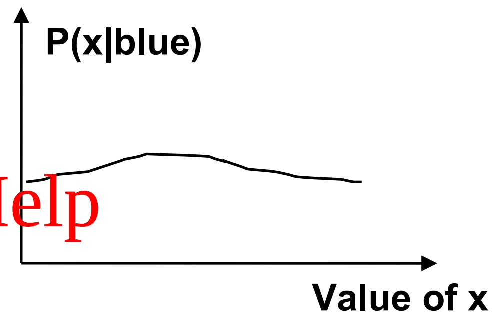
Suppose there are two full bowls of cookies. Bowl #1 has 10 chocolate chip and 30 plain cookies, while bowl #2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?

Write your equations and compute the probability. It should be 0.6.

Naïve Bayes

- How does Naïve Bayes perform on the following dataset?



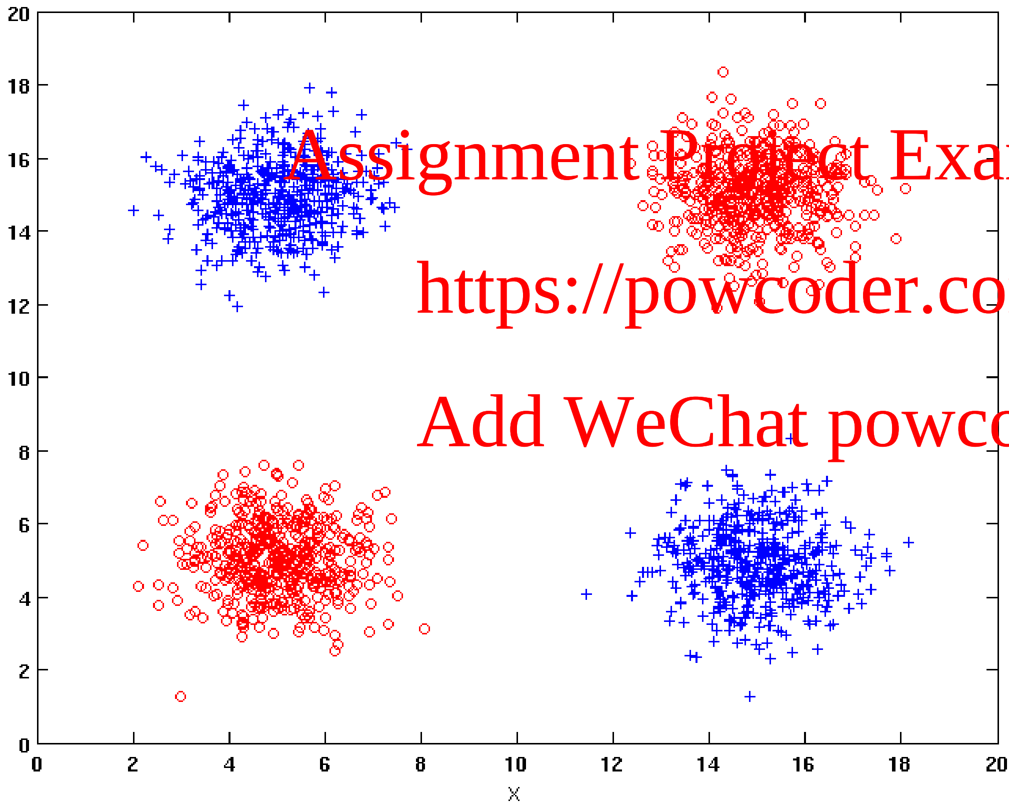


Conditionally independent: x and y inside each class: think about their histogram (distributions)

Naïve Bayes can construct oblique decision boundaries

Naïve Bayes

- How does Naïve Bayes perform on the following dataset?



1. How about decision tree?

2. Input: x=16, y=16

Conditional independence of attributes is violated

Naïve Bayes classifier summary: a different view

Bayesian inference derives the posterior probability as a consequence of two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

where

- H stands for any *hypothesis* whose probability may be affected by data (called *evidence* below). Often there are competing hypotheses, and the task is to determine which is the most probable.
- $P(H)$, the *prior probability*, is the estimate of the probability of the hypothesis H *before* the data E, the current evidence, is observed.
- E, the *evidence*, corresponds to new data that were not used in computing the prior probability.
- $P(H \mid E)$, the *posterior probability*, is the probability of H *given* E, i.e., *after* E is observed. This is what we want to know: the probability of a hypothesis *given* the observed evidence.
- $P(E \mid H)$ is the probability of observing E *given* H, and is called the *likelihood*. As a function of E with H fixed, it indicates the compatibility of the evidence with the given hypothesis. The likelihood function is a function of the evidence, E, while the posterior probability is a function of the hypothesis, H.
- $P(E)$ is sometimes termed the marginal likelihood or "model evidence". This factor is the same for all possible hypotheses being considered (as is evident from the fact that the hypothesis H does not appear anywhere in the symbol, unlike for all the other factors), so this factor does not enter into determining the relative probabilities of different hypotheses.

Naïve Bayes (Summary)

- Robust to isolated noise points

- Handle missing values by ignoring the instance during probability estimate calculations

- Robust to irrelevant attributes

- **Independence assumption may not hold for some attributes**
  - **Use other techniques such as Bayesian Belief Networks (BBN)**