School of Computing
UNIVERSITY OF GEORGIA

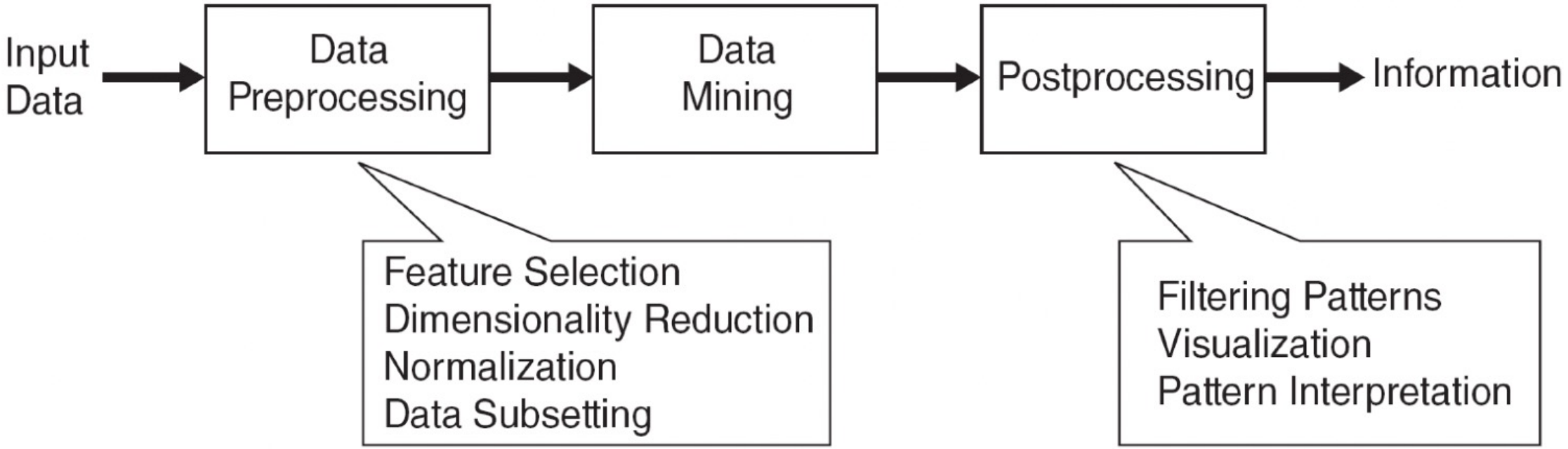
CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

October 10, 2023

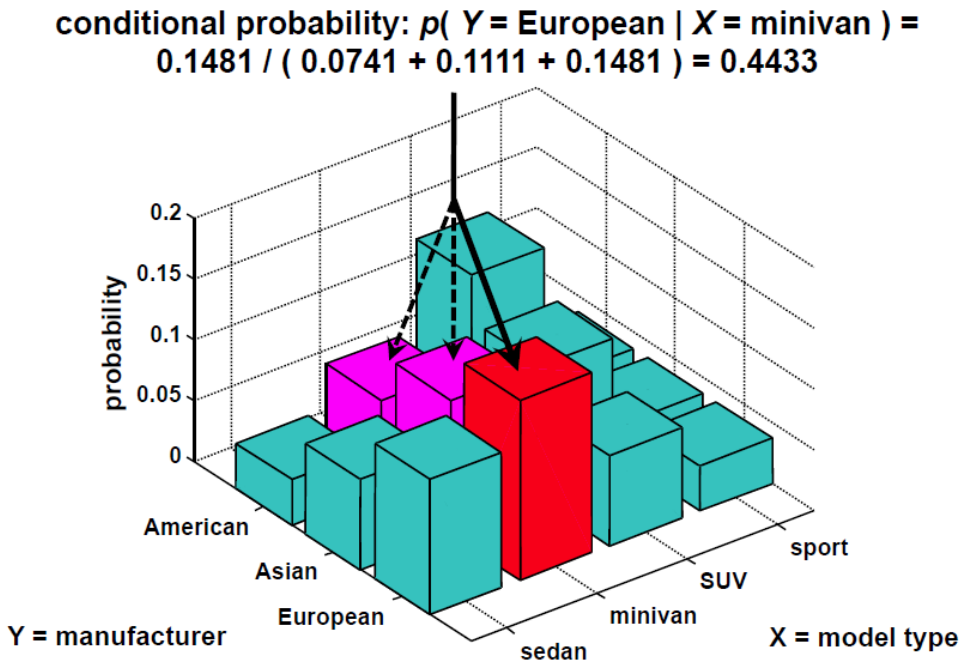
Recap: Data Mining Process



Naïve Bayes

Recap: Conditional Probability Distribution

- ***Conditional probability distribution*** is the probability distribution of one variable provided that another variable has taken a certain value
  - Denoted $P(X = x|\, Y = y)$
- Note that: $P(X = x|\, Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$



Recap: Independence

- Two random variables X and Y are ***independent*** if the occurrence of Y does not reveal any information about the occurrence of X
  - E.g., two successive rolls of a die are independent

- Therefore, we can write: $P(X|\ Y) = P(X)$
  - The following notation is used: $X \perp Y$
  - Also note that for independent random variables: $P(X, Y) = P(X)P(Y)$

- Two random variables X and Y are ***conditionally independent*** given another random variable Z if and only if $P(X, Y|Z) = P(X|Z)P(Y|Z)$
  - This is denoted as $X \perp Y|Z$

Recap: Bayes' Theorem

- ***Bayes' theorem*** – allows to calculate conditional probabilities for one variable when conditional probabilities for another variable are known

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

- Also known as Bayes' rule
- Multiplication rule for the joint distribution is used: $P(X, Y) = P(Y|\,X)P(X)$

- The terms are referred to as:
  - $P(X)$, the prior probability, the initial degree of belief for X
  - $P(X|\,Y)$, the posterior probability, the degree of belief after incorporating the knowledge of Y
  - $P(Y|\,X)$, the likelihood of Y given X
  - P(Y), the evidence
  - Bayes' theorem: $\textbf{posterior probability} = \frac{\textbf{likelihood} \times \textbf{prior probability}}{\textbf{evidence}}$

Recap: Generalized Product (Chain) Rule

- $P(a, b, c, d, \dots) = P(a)P(b|a)\ P(c|a, b)P(d|a, b, c)\ \dots$

Recap: Generalized Product (Chain) Rule

- $P(a, b, c, d, \dots) = P(a)P(b|a)\ P(c|a, b)P(d|a, b, c)\ \dots$

- Test: $P(c, d, \dots . |a, b)$ =?

Recap: Generalized Product (Chain) Rule

- $P(a, b, c, d, \dots) = P(a)P(b|a)\ P(c|a, b)P(d|a, b, c)\ \dots$

- Test: $P(c, d, \dots . |a, b)$ =?
- $P(c, d, \dots . |a, b) = P(c|a, b)\ P(d|a, b, c)\ \dots \dots$

Recap: Bayes Rule

- Bayes Rule:

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

- Test:
  - Dangerous fires are rare (1%),
  - Smoke is fairly common (10%) due to barbecues,
  - 90% of dangerous fires make smoke.
  - What's probability of dangerous Fire when there is Smoke?

Recap: Bayes Rule

- Bayes Rule:

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

- Test:
  - Dangerous fires are rare (1%),
  - Smoke is fairly common (10%) due to barbecues,
  - 90% of dangerous fires make smoke.
  - What's probability of dangerous Fire when there is Smoke?
  - $P(F|S) = \frac{P(S|F)P(F)}{P(S)} = \frac{0.9*0.01}{0.1} = 0.09$

Using Bayes Theorem for Classification

- Consider each attribute and class label as random variables

- Given a record with attributes (X1, X2,…, Xd), the goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes P(Y| X1, X2,…, Xd)

- Can we estimate P(Y| X1, X2,…, Xd) directly from data?

Using Bayes Theorem for Classification

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

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

  - *Maximum a-posteriori*: Choose Y that maximizes
    $P(Y \mid X_1, X_2, \ldots, X_d)$
  - Equivalent to choosing value of Y that maximizes
    $P(X_1, X_2, \ldots, X_d \mid Y)\ P(Y)$

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

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(Evade = Yes | X) and P(Evade = No | X)

  In the following we will replace

  Evade = Yes by Yes, and

  Evade = No by No

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:

- $P(\text{Yes} \mid X) = \dfrac{P(X \mid \text{Yes})P(\text{Yes})}{P(X)}$
- $P(\text{No} \mid X) = \dfrac{P(X \mid \text{No})P(\text{No})}{P(X)}$
- How to estimate P(X | Yes) and P(X | No)?

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

Naïve Bayes Classifier

- Assume independence among attributes X_i when class is given:
  - $P(X_1, X_2, \ldots, X_d \mid Y_j) = P(X_1 \mid Y_j)\, P(X_2 \mid Y_j) \ldots P(X_d \mid Y_j)$
  - 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

- P(y) = fraction of instances of class y
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10

- 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:

    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

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(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

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

P(Yes | Divorced) = 1/3 x 3/10 / P(Divorced)
P(No | Divorced) = 1/7 x 7/10 / P(Divorced)

If we also know that Refund = No, then

P(Yes | Refund = No, Divorced) = 1 x 1/3 x 3/10 / P(Divorced, Refund = No)

P(No | Refund = No, Divorced) = 4/7 x 1/7 x 7/10 / P(Divorced, Refund = No)

If we also know that Taxable Income = 120, then

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

P(No | Refund = No, Divorced Income = 120) = 0.0072 x 4/7 x 1/7 x 7/10 / P(Divorced, Refund = No, Income = 120)

Issues with Naïve Bayes Classifier

Given a Test Record: X = (Married)

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
7	Yes	Divorced	220K	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)

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

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

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:

n: number of training instances belonging to class y

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

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

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

v: total number of attribute values that X_i can take

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

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

m: hyper-parameter for our confidence in p

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)