Bayesian Classifiers

Data Mining

Prof. Sujee Lee

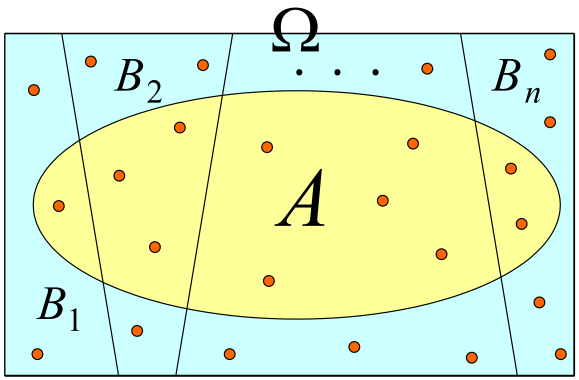
Department of Systems Management Engineering

Sungkyunkwan University

Bayes' Theorem

- **Joint probability**

$$P(X, Y)$$



- **Conditional probability**

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

- **Bayes' theorem**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \qquad Pr(B_i|A) = \frac{Pr(B_i)Pr(A|B_i)}{\sum_{j=1}^{n} Pr(B_j)Pr(A|B_j)}$$

- Example

There are 1,000 coins in a jar. 999 normal coins have both head and tail, but 1 coin has two heads. You randomly picked up a coin from the jar and tossed it 10 times. All outcomes are heads. What is the probability you picked up the coin having two heads?

Bayesian Classifiers

- **Bayesian Classifiers**
  - Consider each independent (predictor) variable X_i and the class variable Y as random variables.
  - Goal : Given a record with variables $(X_1, X_2, \cdots, X_p)$, predict its class Y.
  - Specifically, we want to find the value of Y that maximizes

$$P(Y|X_1, X_2, \cdots, X_p)$$

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

Bayesian Classifiers (cont.)

- **Approach**:
  - Compute the posterior probability $P(Y|X_1, X_2, \cdots, X_p)$ for all values of Y using the Bayes' theorem

$$P(Y|X_1, X_2, \cdots, X_p) = \frac{P(X_1, X_2, \cdots, X_p|Y)P(Y)}{P(X_1, X_2, \cdots, X_p)}$$

    - **Prior: the initial belief about the probability of a certain setting/an event**
    - **Likelihood: how probable the observed data set is for the given setting**
    - **Posterior: the updated probability of the setting/event after considering the observed data**

  - Choose value of Y that maximizes

$$P(Y|X_1, X_2, \cdots, X_p)$$

  - It is equivalent to choosing value of Y that maximizes

$$P(X_1, X_2, \cdots, X_p|Y)P(Y)$$

  - How to estimate $P(X_1, X_2, \cdots, X_p|Y)P(Y)$?

Naïve Bayes Classifier

- **The Chain Rule**

$$P(X_1, X_2, \ldots, X_p|Y)P(Y)$$

$$= \frac{P(X_1, X_2, \ldots, X_p, Y)}{P(X_2, X_3, \ldots, X_p, Y)} \cdot \frac{P(X_2, X_3, \ldots, X_p, Y)}{P(Y)} \cdot P(Y)$$

$$= P(X_1|X_2, X_3, \ldots, X_p, Y) \cdot P(X_2, X_3, \ldots, X_p|Y) \cdot P(Y)$$

$$= P(X_1|X_2, X_3, \ldots, X_p, Y) \cdot P(X_2|X_3, X_4, \ldots, X_p, Y) \cdot P(X_3, X_4, \ldots, X_p|Y) \cdot P(Y)$$

$$\vdots$$

$$= P(X_1|X_2, X_3, \ldots, X_p, Y) \cdot P(X_2|X_3, X_4, \ldots, X_p, Y) \cdots P(X_p|Y) \cdot P(Y)$$

Not easy yet…

Naïve Bayes Classifier (cont.)

- **Conditional Independence**
  - Let X, Y, and Z denote three sets of random variables. The variables in X are said to be conditionally independent of Y, given Z, if the following condition holds.

$$P(X|Y,Z) = P(X|Z)$$

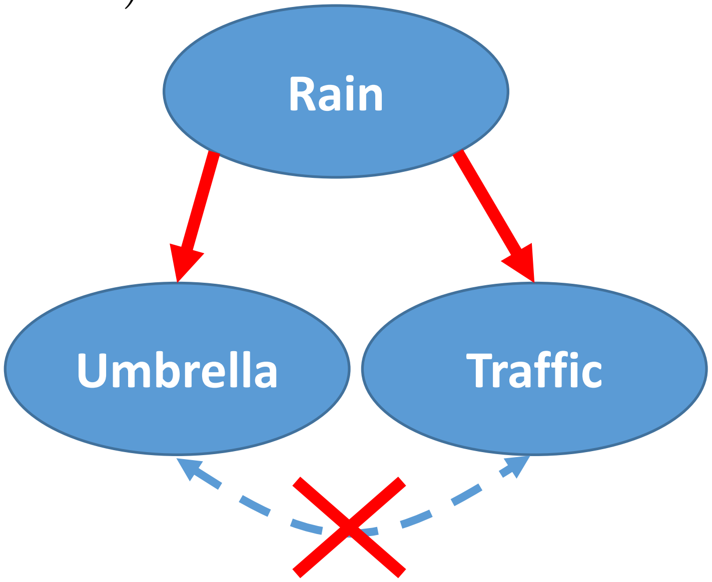
  - Example

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella})$$
$$= P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

Conditional Independence

between Umbrella and Traffic

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella})$$
$$= P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$



Naïve Bayes Classifier (cont.)

- Your life becomes much easier with Conditional Independence.
  - Assume independence among variables ($X_i's$) when class is given.

$$P(X_1, X_2, \ldots, X_p|Y)P(Y)$$
$$= P(X_1|X_2, X_3, \ldots, X_p, Y) \cdot P(X_2|X_3, X_4, \ldots, X_p, Y) \cdots P(X_p|Y) \cdot P(Y)$$
$$= P(X_1|Y) \cdot P(X_2|Y) \cdots P(X_p|Y) \cdot P(Y)$$

- Can estimate $P(X_i|Y = j)$ for all X_i and $Y = j$.
- A new data is classified to j if $P(Y = j) \cdot \prod_{i=1}^{p} P(X_i|Y = j)$ is maximal.

How to Estimate Probabilities from Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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: $P(Y=j) = N(y_j)/N$
  - e.g.,

$$P(Y = \text{No}) = \frac{7}{10}, \; P(Y = \text{Yes}) = \frac{3}{10}$$

- For discrete attributes:

$$P(X_i|Y=j) = |X_{ij}|/N(y_j)$$

  - where $|X_{ij}|$ is the number of instances having variable X_i and belongs to class j
  - Examples:

$$P(\text{Marital Status} = \text{Married}|\text{Defaulted Borrower} = \text{No}) = \frac{4}{7}$$

$$P(\text{Home Owner} = \text{Yes}|\text{Defaulted Borrower} = \text{Yes}) = 0$$

How to Estimate Probabilities from Data (cont.)

- **For continuous variables:**
  - **Discretize** the range into bins
    - one ordinal attribute per bin
    - violates independence assumption
  - **Two-way split:** $(X \leq v)$ or $(X > v)$
    - choose only one of the two splits as new attribute
  - **Probability density estimation:**
    - Assume variable follows a normal distribution
    - Use data to estimate parameters of distribution

      (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability

$$P(X_i|Y)$$

How to Estimate Probabilities from Data (cont.)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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:

$$P(X_i|Y=j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(X_i-\mu_{ij})^2}{2\sigma_{ij}^2} \right)$$

  – One for each $(X_i, Y=j)$ pair

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

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

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{No}, \text{Married}, \text{Income} = 120)$$

naive 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) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

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

$$\prod P(X|DB=No) = P(HW=No|DB=No) \times P(MS=Married|\ DB=No) \times P(AI=120K|\ DB=No)$$
$$= 4/7 \times 4/7 \times 0.0072 = 0.0024$$

$$\prod P(X|DB=Yes) = P(HW=No|\ DB=Yes) \times P(MS=Married|\ DB=Yes) \times P(AI=120K|\ DB=Yes)$$
$$= 1 \times 0 \times 1.2 \times 10^{-9} = 0$$

Since $P(No)\prod P(X|No) > P(Yes)\prod P(X|Yes)$

=> Predict as (Defaulted Borrower = No)

Naïve Bayes Classifier

- **Probability Smoothing**
  - If one of the conditional probability is zero, then the entire expression becomes zero.
  - Probability estimation:
    - Original:

$$P(X_i|Y=j)=\frac{|X_{ij}|}{N(y_j)}$$

    - Laplace:

$$P(X_i|Y=j)=\frac{|X_{ij}|+1}{N(y_j)+J}$$

    - m-estimate:

$$P(X_i|Y=j)=\frac{|X_{ij}|+mp}{N(y_j)+m}$$

J: number of classes

p: prior probability

m: sample size

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

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

A: attributes

M: mammals

N: non-mammals

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

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

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

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

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

=> Mammals

Summary: Naïve Bayes Classifer

- 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 Networks (BN)