

# **Data Mining**

## **Classification: Naïve Bayes**

---

Lecture Notes for Chapter 5

Introduction to Data Mining  
by  
Tan, Steinbach, Kumar

# Bayes Classifier

---

- | A probabilistic framework for solving classification problems

- | Conditional Probability:

$$P(C | A) = \frac{P(A, C)}{P(A)}$$

$$P(A | C) = \frac{P(A, C)}{P(C)}$$

- | Bayes theorem:

$$P(C | A) = \frac{P(A | C)P(C)}{P(A)}$$

# 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?

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

# Bayesian Classifiers

---

- Consider each attribute and class label as random variables
- Given a record with attributes  $(A_1, A_2, \dots, A_n)$ 
  - Goal is to predict class  $C$
  - Specifically, we want to find the value of  $C$  that maximizes  $P(C | A_1, A_2, \dots, A_n)$
- Can we estimate  $P(C | A_1, A_2, \dots, A_n)$  directly from data?

# Bayesian Classifiers

## □ Approach:

- compute the posterior probability  $P(C \mid A_1, A_2, \dots, A_n)$  for all values of  $C$  using the Bayes theorem

$$P(C \mid A_1 A_2 \dots A_n) = \frac{P(A_1 A_2 \dots A_n \mid C) P(C)}{P(A_1 A_2 \dots A_n)}$$

- Choose value of  $C$  that maximizes  $P(C \mid A_1, A_2, \dots, A_n)$
- Equivalent to choosing value of  $C$  that maximizes  $P(A_1, A_2, \dots, A_n \mid C) P(C)$

## □ How to estimate $P(A_1, A_2, \dots, A_n \mid C)$ ?

# Naïve Bayes Classifier

---

- Assume independence among attributes  $A_i$  when class is given:
  - $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C) P(A_2 | C) \dots P(A_n | C)$
  - Can estimate  $P(A_i | C)$  for all  $A_i$  and  $C$ .
  - New point is classified to  $C_j$  if  $P(C_j) \prod P(A_i | C_j)$  is maximal.

# How to Estimate Probabilities from Data?

<i>Tid</i>	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:  $P(C) = N_c/N$

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

| For discrete attributes:

$$P(A_i | C_k) = |A_{ik}| / N_{C_k}$$

- where  $|A_{ik}|$  is number of instances having attribute  $A_i$  and belongs to class  $C_k$

– Examples:

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

$$P(\text{Refund}=\text{Yes}|\text{Yes})=0$$

# How to Estimate Probabilities from Data?

---

- | For continuous attributes:
  - **Discretize** the range into bins
    - ◆ one ordinal attribute per bin
    - ◆ violates independence assumption <sup>k</sup>
  - **Two-way split:**  $(A < v)$  or  $(A > v)$ 
    - ◆ choose only one of the two splits as new attribute
  - **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, can use it to estimate the conditional probability  $P(A_i|c)$



# How to Estimate Probabilities from Data?

<i>Tid</i>	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:

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

— One for each  $(A_i, c_i)$  pair

| For (Income, Class=No):

— If Class=No

◆ sample mean = 110

◆ sample variance = 2975

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

# Example of Naïve Bayes Classifier

**Given a Test Record:**

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

**Given a Test Record 2: X= (Ref=Yes, Single, Income= 80K)**

naive Bayes Classifier:

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

For taxable income:

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

- $P(X|\text{Class}=\text{No}) = P(\text{Refund}=\text{No}|\text{Class}=\text{No})$   
                                  $\times P(\text{Married}|\text{Class}=\text{No})$   
                                  $\times P(\text{Income}=120\text{K}|\text{Class}=\text{No})$   
                                  $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $P(X|\text{Class}=\text{Yes}) = P(\text{Refund}=\text{No}|\text{Class}=\text{Yes})$   
                                  $\times P(\text{Married}|\text{Class}=\text{Yes})$   
                                  $\times P(\text{Income}=120\text{K}|\text{Class}=\text{Yes})$   
                                  $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

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

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

# Naïve Bayes Classifier

---

- | If one of the conditional probability is zero, then the entire expression becomes zero
- | Probability estimation:

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

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

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

c: number of classes

p: prior probability

m: parameter

# 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**

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

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

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

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

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

# 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)