

CSE4392/5334
DATA MINING

Lecture 7:
Classification (3)

CSE4392/5334 Data Mining, Fall 2009
Department of Computer Science and Engineering, University of Texas at Arlington
Chengkai Li (Slides courtesy of Vipin Kumar)

Bayes Classifier

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:
 - Team A wins $P(W=A) = 0.65$
 - Team B wins $P(W=B) = 0.35$
 - If team A won, the probability that team B hosted the game $P(H=B | W=A) = 0.30$
 - If team B won, the probability that team B hosted the game $P(H=B | W=B) = 0.75$
- If team B is the next host, which team has a better chance to win?

$$P(W = A | H) = \frac{P(H | W)P(W)}{P(H)}$$

$$P(W = A | H = B) = \frac{P(H = B | W = A)P(W = A)}{P(H = B)} = \frac{0.30 \times 0.65}{P(H = B)}$$

$$P(W = B | H = B) = \frac{P(H = B | W = B)P(W = B)}{P(H = B)} = \frac{0.75 \times 0.35}{P(H = B)}$$
- And how big is the chance?

$$P(H = B) = P(H = B, W = A) + P(H = B, W = B) = P(H = B | W = A)P(W = A) + P(H = B | W = B)P(W = B)$$

$$= 0.30 \times 0.65 + 0.75 \times 0.35$$

$$P(W = B | H = B) = \frac{0.75 \times 0.35}{0.30 \times 0.65 + 0.75 \times 0.35}$$

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 $P(C | A_1, A_2, \dots, A_n)$ for all values of C using the Bayes theorem
$$P(C | A_1, A_2, \dots, A_n) = \frac{P(A_1, A_2, \dots, A_n | C)P(C)}{P(A_1, A_2, \dots, A_n)}$$
- Choose value of C that maximizes $P(C | A_1, A_2, \dots, A_n)$
- Equivalent to choosing value of C that maximizes $P(A_1, A_2, \dots, A_n | C)P(C)$
- How to estimate $P(A_1, A_2, \dots, A_n | C)$?

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 $P(A_i | C_j)$ for all A_i and C_j .
 - New point is classified to C_i if $P(C_i) \prod P(A_j | C_i)$ is maximal.

How to 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: $P(C) = N_c / N$
 - e.g., $P(\text{No}) = 7/10$, $P(\text{Yes}) = 3/10$
- For discrete attributes:
 - $P(A_i | C_k) = |A_{ik}| / N_{kc}$
 - 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
 - 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?

| 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:
 - $P(A_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(A_i - \mu_j)^2}{2\sigma_j^2}}$
 - One for each (A_i, c_j) pair
- For $(\text{Income}, \text{Class}=\text{No})$:
 - If $\text{Class}=\text{No}$
 - 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:

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

naïve Bayes Classifier:

For refund=Yes:

- $P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$
- $P(\text{Refund}=\text{No}|\text{No}) = 4/7$
- $P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$
- $P(\text{Refund}=\text{No}|\text{Yes}) = 1$

For marital status=Single:

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

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

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:
 - Original: $P(A_i | C) = \frac{N_{ic}}{N_c}$
 - Laplace: $P(A_i | C) = \frac{N_{ic} + 1}{N_c + c}$
 - 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 |
| seal | 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 |
| lizard | 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)