CSE4334/5334 Data Mining
Naïve Bayes Classifier

Chengkai Li
University of Texas at Arlington
Spring 2015

(Slides courtesy of Pang-Ning Tan, Michael Steinbach and Vipin Kumar)
Bayes Classifier

A probabilistic framework for solving classification problems

Conditional Probability:

\[
P(C \mid A) = \frac{P(A, C)}{P(A)}
\]

\[
P(A \mid C) = \frac{P(A, C)}{P(C)}
\]

Bayes theorem:

\[
P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}
\]
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?

\[
P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002
\]
Bayesian Classifiers

Consider each attribute and class label as random variables

Given a record with attributes \((A_1, A_2, \ldots, A_n)\)

- Goal is to predict class \(C\)
- Specifically, we want to find the value of \(C\) that maximizes \(P(C | A_1, A_2, \ldots, A_n)\)

Can we estimate \(P(C | A_1, A_2, \ldots, A_n)\) directly from data?
Bayesian Classifiers

Approach:

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

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

- Choose value of $C$ that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$

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

How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?
Naïve Bayes Classifier

Assume independence among attributes $A_i$ when class is given:

- $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j) P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$

- Can estimate $P(A_i \mid C_j)$ for all $A_i$ and $C_j$.

- New point is classified to $C_j$ if $P(C_j) \prod P(A_i \mid C_j)$ is maximal.
How to Estimate Probabilities from Data?

Class: \[ P(C) = \frac{N_c}{N} \]
- e.g., \[ P(\text{No}) = \frac{7}{10}, \]
  \[ P(\text{Yes}) = \frac{3}{10} \]

For discrete attributes:

\[ P(A_i | C_k) = \frac{|A_{ik}|}{N_c} \]
- where \(|A_{ik}|\) is number of instances having attribute \(A_i\) and belongs to class \(C_k\)

Examples:
- \[ P(\text{Status}=\text{Married} | \text{No}) = \frac{4}{7} \]
- \[ P(\text{Refund}=\text{Yes} | \text{Yes}) = 0 \]
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?

Normal distribution:

\[
P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma^2_{ij}}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma^2_{ij}}}
\]

- 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

\[
P(\text{Income} = 120 \mid \text{No}) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072
\]
Example of Naïve Bayes Classifier

Given a Test Record:

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

**naive Bayes Classifier:**

- \[ P(X|\text{Class=No}) = P(\text{Refund=No}|\text{Class=No}) \times P(\text{Married}|\text{Class=No}) \times P(\text{Income}=120\text{K}|\text{Class=No}) \]
  \[ = \frac{4}{7} \times \frac{4}{7} \times 0.0072 = 0.0024 \]

- \[ P(X|\text{Class=Yes}) = P(\text{Refund=No}|\text{Class=Yes}) \times P(\text{Married}|\text{Class=Yes}) \times P(\text{Income}=120\text{K}|\text{Class=Yes}) \]
  \[ = 1 \times 0 \times 1.2 \times 10^{-9} = 0 \]

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:

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Give Birth</th>
<th>Can Fly</th>
<th>Live in Water</th>
<th>Have Legs</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>python</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>salmon</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>whale</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>mammals</td>
</tr>
<tr>
<td>frog</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>komodo</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>bat</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>pigeon</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>cat</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>leopard shark</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>turtle</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>penguin</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>porcupine</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>eel</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>salamander</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>gila monster</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>platypus</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>owl</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>dolphin</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>mammals</td>
</tr>
<tr>
<td>eagle</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
</tbody>
</table>

A: attributes
M: mammals
N: non-mammals

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

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

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

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

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)