Data Mining
Classification: Alternative Techniques

Lecture Notes for Chapter 5

Introduction to Data Mining
by
Tan, Steinbach, Kumar
Instance-based classification

Bayesian classification

Lecture of 3 March 2016
## Instance-Based Classifiers

### Set of Stored Cases

<table>
<thead>
<tr>
<th>Atr1</th>
<th>.......</th>
<th>AtrN</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B</td>
</tr>
</tbody>
</table>

- Store the training records
- Use training records to predict the class label of unseen cases

### Unseen Case

<table>
<thead>
<tr>
<th>Atr1</th>
<th>.......</th>
<th>AtrN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Instance Based Classifiers

Examples:

- **Rote-learner**
  
  Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly.

- **Nearest neighbor**
  
  Uses k “closest” points (nearest neighbors) for performing classification.
Nearest Neighbor Classifiers

Basic idea:
- If it walks like a duck, quacks like a duck, then it’s probably a duck

Training Records

Compute Distance

Choose k of the “nearest” records

Test Record
Nearest-Neighbor Classifiers

- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of $k$, the number of nearest neighbors to retrieve

- To classify an unknown record:
  - Compute distance to other training records
  - Identify $k$ nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)
Definition of Nearest Neighbor

K-nearest neighbors of a record \( x \) are data points that have the \( k \) smallest distance to \( x \)

(a) 1-nearest neighbor  
(b) 2-nearest neighbor  
(c) 3-nearest neighbor
1 nearest-neighbor

Voronoi Diagram
Nearest Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

\[
d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}
\]

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, \( w = 1/d^2 \)
Choosing the value of k:

- If k is too small, sensitive to noise points
- If k is too large, neighborhood may include points from other classes
Nearest Neighbor Classification...

- Scaling issues
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M
Nearest Neighbor Classification...

- Problem with Euclidean measure:
  - High dimensional data
  - curse of dimensionality
  - Can produce counter-intuitive results

\[
\begin{align*}
\text{vs} \quad & 1 1 1 1 1 1 1 1 1 1 1 0 \\
& 0 1 1 1 1 1 1 1 1 1 1 1 \\
& \downarrow \quad d = 1.4142 \\
& 1 0 0 0 0 0 0 0 0 0 0 0 0 \\
& 0 0 0 0 0 0 0 0 0 0 0 0 1 \\
& \quad d = 1.4142
\end{align*}
\]

- Solution: Normalize the vectors to unit length
k-NN classifiers are lazy learners

- It does not build models explicitly
- Unlike eager learners such as decision tree induction and rule-based systems
- Classifying unknown records are relatively expensive
Example: PEBLS

- PEBLS: Parallel Examplar-Based Learning System (Cost & Salzberg)
  - Works with both continuous and nominal features
    - For nominal features, distance between two nominal values is computed using modified value difference metric (MVDM)
  - Each record is assigned a weight factor
  - Number of nearest neighbor, k = 1
**Example: PEBLS**

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Distance between nominal attribute values:**

\[
d(Single,\text{Married}) = |2/4 - 0/4| + |2/4 - 4/4| = 1
\]

\[
d(Single,\text{Divorced}) = |2/4 - 1/2| + |2/4 - 1/2| = 0
\]

\[
d(\text{Married},\text{Divorced}) = |0/4 - 1/2| + |4/4 - 1/2| = 1
\]

\[
d(\text{Refund} = \text{Yes}, \text{Refund} = \text{No}) = |0/3 - 3/7| + |3/3 - 4/7| = 6/7
\]
Example: PEBLS

Distance between record X and record Y:

\[
\Delta(X, Y) = w_X w_Y \sum_{i=1}^{d} d(X_i, Y_i)^2
\]

where:

\[
w_X = \frac{\text{Number of times X is used for prediction}}{\text{Number of times X predicts correctly}}
\]

- \( w_X \approx 1 \) if X makes accurate prediction most of the time
- \( w_X > 1 \) if X is not reliable for making predictions
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?

\[
P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times \frac{1}{50000}}{\frac{1}{20}} = 0.0002
\]
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:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j) P(A_2 | C_j) \ldots P(A_n | C_j)$
  - Can estimate $P(A_i | C_j)$ for all $A_i$ and $C_j$.
  - 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?

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

- **For discrete attributes:**
  \[
P(A_i \mid C_k) = \frac{|A_{ik}|}{N_{ck}}
  \]
  - where \(|A_{ik}|\) is number of instances having attribute \(A_i\) and belongs to class \(C_k\)
  - Examples:
    \[
P(\text{Status}=\text{Married} \mid \text{No}) = \frac{4}{7} \quad \text{P(Refund=Yes} \mid \text{Yes})=0
    \]
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:
  \[ P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma^2_{ij}}} e^{-\frac{(A_i - \mu_j)^2}{2\sigma^2_{ij}}} \]
  - One for each \((A_i, c_i)\) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

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

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

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

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

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

\[
P(A|M)P(M) > P(A|N)P(N)
\]

\[
=> \text{Mammals}
\]
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)