Elective Subject (2) for Computer Science

DATA MINING

Dr. Salha Mohammed Alzahrani
College of Computers and Information Technology
Taif University
Saudi Arabia
s.zahrani@tu.edu.sa / admin@c2learn.com

Data Processing |

Recap of Lecture 2

- Standard Formulation
- Types of Variables
- Categorical and Continuous Attributes
- Data Preparation
  - Data Cleaning
  - Missing Values
  - Attributes Reduction
- The UCI Repository of Datasets
Classification | What is Classification?

Classification is the process of dividing up objects so that each is assigned to one of a number of *mutually exhaustive and exclusive* categories known as classes.

- Classification normally used with labelled datasets.
- The term *mutually exhaustive and exclusive* simply means that each object must be assigned to precisely one class, i.e. never to more than one and never to no class at all.
Naïve Bayes Classifiers is a method of classification that does not use rules, a decision tree or any other explicit representation of the classifier. It uses the probability theory to find the most likely classification of an unseen (unclassified) instance.

- **Naïve** = simple or straightforward
- **Bayes** = Reverend Thomas Bayes (1702–1761), an English Mathematician.

The probability of an event is the number of times we would expect an event to occur over a long series of trails.

- For Example, the probability that the 6.30 p.m. train from London to Oxford arrives on time.
- The probability is a number from 0 to 1 inclusive, with 0 indicating ‘impossible’ and 1 indicating ‘certain’.
Classification | The train example

- In the train example, we may have four mutually exclusive and exhaustive events (E):
  - E1 – train cancelled
  - E2 – train ten minutes or more late
  - E3 – train less than ten minutes late
  - E4 – train on time.

- The probability of an event \( P(E) \):
  - \( P(E1) = 0.05 \)
  - \( P(E2) = 0.1 \)
  - \( P(E3) = 0.15 \)
  - \( P(E4) = 0.7 \)

- \( P(E) \in [0,1] \)
- The sum of the four probabilities has to be 1.
  \[ P(E1) + P(E2) + P(E3) + P(E4) = 1 \]

Classification | The train example

- For classification tasks, the labelled dataset is called a training set.
- Each row of the training set is called an instance. An instance comprises the values of a number of attributes and the corresponding classification.
- The training set constitutes the results of a sample of trials that we can use to predict the classification of other (unclassified) instance.

Example:

<table>
<thead>
<tr>
<th>weekday</th>
<th>season</th>
<th>wind</th>
<th>rain</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>weekday</td>
<td>spring</td>
<td>none</td>
<td>none</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>winter</td>
<td>none</td>
<td>slight</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>winter</td>
<td>none</td>
<td>slight</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>winter</td>
<td>high</td>
<td>heavy</td>
<td>late</td>
</tr>
<tr>
<td>weekday</td>
<td>autumn</td>
<td>normal</td>
<td>none</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>summer</td>
<td>high</td>
<td>slight</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>summer</td>
<td>normal</td>
<td>none</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>winter</td>
<td>high</td>
<td>heavy</td>
<td>very late</td>
</tr>
<tr>
<td>weekday</td>
<td>summer</td>
<td>none</td>
<td>slight</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>spring</td>
<td>high</td>
<td>heavy</td>
<td>cancelled</td>
</tr>
<tr>
<td>weekday</td>
<td>summer</td>
<td>high</td>
<td>none</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>winter</td>
<td>normal</td>
<td>heavy</td>
<td>very late</td>
</tr>
<tr>
<td>weekday</td>
<td>autumn</td>
<td>high</td>
<td>slight</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>autumn</td>
<td>none</td>
<td>heavy</td>
<td>on time</td>
</tr>
<tr>
<td>holiday</td>
<td>spring</td>
<td>normal</td>
<td>slight</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>spring</td>
<td>normal</td>
<td>none</td>
<td>on time</td>
</tr>
<tr>
<td>weekday</td>
<td>spring</td>
<td>normal</td>
<td>slight</td>
<td>on time</td>
</tr>
</tbody>
</table>

Figure 2.1 The train Dataset
Classification | The prior probability

The prior probability is the probability of an event based on its observed frequency in a series of trials without any additional information.

\[ P(\text{class} = c) = \frac{\text{frequency of } c}{\text{total number of instances}} \]

Example:
- The prior probability of the train being on time:
  \[ P(\text{class} = \text{on time}) = \frac{14}{20} = 0.7 \]
- If we have no other information, this is the best we can do.
- If we have other (relevant) information, the position is different.

Classification | The conditional (or posterior) probability

The conditional probability (also called posterior probability) is the probability of an event based on its observed frequency in a series of trials and given that we have additional information.

\[ P(\text{class} = c \mid a = v) = \frac{\text{frequency of } c \text{ given attribute } a}{\text{total number of instances in } a} \]

Example:
\[ P(\text{class} = \text{on time} \mid \text{season} = \text{winter}) \] is the probability that the class is on time given that the season is winter.
Classification

The conditional (or posterior) probability

The train example:

- \( P(\text{class = on time} \mid \text{season = winter}) = \frac{2}{6} = 0.33 \)
- \( P(\text{class = late} \mid \text{season = winter}) = \frac{1}{6} = 0.17 \)
- \( P(\text{class = very late} \mid \text{season = winter}) = \frac{3}{6} = 0.5 \)
- \( P(\text{class = cancelled} \mid \text{season = winter}) = \frac{0}{6} = 0 \)

Classification

Naïve Bayes Algorithm

The Naïve Bayes algorithm combines the prior probability and conditional probabilities in a single formula, to calculate the probability of alternative classifications.

- The method uses prior probability as before.
- But posterior probabilities the other way round from before.

Example:
We use the conditional probability that the season is \textit{winter} given that the class is \textit{on time}, i.e. \( P(\text{season = winter} \mid \text{class = on time}) \).
Classification | Naïve Bayes Algorithm

Naïve Bayes Classification

Given a set of \( k \) mutually exclusive and exhaustive classifications \( c_1, c_2, \ldots, c_k \), which have prior probabilities \( P(c_1), P(c_2), \ldots, P(c_k) \), respectively, and \( n \) attributes \( a_1, a_2, \ldots, a_n \) which for a given instance have values \( v_1, v_2, \ldots, v_n \) respectively, the posterior probability of class \( c_i \) occurring for the specified instance can be shown to be proportional to

\[
P(c_i) \times P(a_1 = v_1 \mid c_i) \times P(a_2 = v_2 \mid c_i) \times \ldots \times P(a_n = v_n \mid c_i)
\]

Making the assumption that the attributes are independent, the value of this expression can be calculated using the product

\[
P(c_i) \times P(a_1 = v_1 \mid c_i) \times P(a_2 = v_2 \mid c_i) \times \ldots \times P(a_n = v_n \mid c_i)
\]

We calculate this product for each value of \( i \) from 1 to \( k \) and choose the classification that has the largest value.

Figure 2.3 The Naïve Bayes Classification Algorithm

---

For the train dataset, we can tabulate all the conditional and prior probabilities, as shown in this table.

<table>
<thead>
<tr>
<th></th>
<th>class = on time</th>
<th>class = late</th>
<th>class = very late</th>
<th>class = cancelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>day</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weekday</td>
<td>9/14 = 0.64</td>
<td></td>
<td>3/8 = 0.37</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>saturday</td>
<td>2/14 = 0.14</td>
<td>1/2 = 0.5</td>
<td>0/3 = 0</td>
<td>1/1 = 1</td>
</tr>
<tr>
<td>sun</td>
<td>1/14 = 0.07</td>
<td>0/2 = 0</td>
<td>0/3 = 0</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>holiday</td>
<td>2/14 = 0.14</td>
<td>0/2 = 0</td>
<td>0/3 = 0</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spring</td>
<td>4/14 = 0.29</td>
<td>0/2 = 0</td>
<td>0/3 = 0</td>
<td>1/1 = 0</td>
</tr>
<tr>
<td>summer</td>
<td>6/14 = 0.43</td>
<td>0/2 = 0</td>
<td>0/3 = 0</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>autumn</td>
<td>2/14 = 0.14</td>
<td>0/2 = 0</td>
<td>1/3 = 0.33</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>winter</td>
<td>2/14 = 0.14</td>
<td>2/2 = 1</td>
<td>2/3 = 0.67</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>wind</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>5/14 = 0.36</td>
<td>0/2 = 0</td>
<td>0/3 = 0</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>high</td>
<td>4/14 = 0.29</td>
<td>1/2 = 0.5</td>
<td>1/3 = 0.33</td>
<td>1/1 = 1</td>
</tr>
<tr>
<td>normal</td>
<td>5/14 = 0.36</td>
<td>1/2 = 0.5</td>
<td>2/3 = 0.67</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>rain</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>5/14 = 0.36</td>
<td>1/2 = 0.5</td>
<td>1/3 = 0.33</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>slight</td>
<td>8/14 = 0.57</td>
<td>0/2 = 0</td>
<td>0/3 = 0</td>
<td>0/1 = 0</td>
</tr>
<tr>
<td>heavy</td>
<td>1/14 = 0.07</td>
<td>1/2 = 0.6</td>
<td>2/3 = 0.67</td>
<td>1/1 = 1</td>
</tr>
<tr>
<td>Prior Probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14/20 = 0.70</td>
<td>2/20 = 0.10</td>
<td>3/20 = 0.15</td>
<td>1/20 = 0.05</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.2 Conditional and Prior Probabilities: train Dataset
Classification | Naïve Bayes Algorithm: classification of unseen instance

Using the values in each of the columns in the table, we can obtain the posterior probabilities for each possible classification for the unseen instance:

| weekday | winter | high | heavy | ????
|---------|--------|------|-------|-------|
- **class = on time**: 0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013
- **class = late**: 0.10 \times 0.50 \times 1.00 \times 0.50 \times 0.50 = 0.0125
- **class = very late**: 0.15 \times 1.00 \times 0.67 \times 0.33 \times 0.67 = 0.0222
- **class = cancelled**: 0.05 \times 0.00 \times 0.00 \times 1.00 \times 1.00 = 0.0000

**The largest value is for class = very late**

---

Classification | Naïve Bayes Algorithm: summary of steps

**Naïve Bayes Algorithm**

**Step 1**: define the classes
**Step 2**: Calculate the prior probability of each class.
**Step 3**: Calculate the posterior probability of each attribute given that each class (in a table).
**Step 4**: Using the values in each of the columns in the table, calculate the posterior (conditional) probabilities for each possible classification for a given unseen (unclassified) instance.
**Step 4**: Choose the largest value as a classification of the given unseen instance.
Classification | Self-assessment Exercise

- Using the Naïve Bayes classification algorithm with the *train* dataset, calculate the most likely classification for the following unseen instances.

<table>
<thead>
<tr>
<th>weekday</th>
<th>summer</th>
<th>high</th>
<th>heavy</th>
<th>????</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunday</td>
<td>summer</td>
<td>normal</td>
<td>slight</td>
<td>????</td>
</tr>
</tbody>
</table>

Data Mining Intro. | Data mining and knowledge discovery

Summary & Checklist

- What is Classification?
- Naïve Bayes Classifiers
- Probability of an event
- The train example
- The prior probability
- The conditional (or posterior) probability
- Naïve Bayes Algorithm
- Naïve Bayes Algorithm: The train example
- Naïve Bayes Algorithm: classification of unseen instance
- Naïve Bayes Algorithm: summary of steps
- Self-assessment Exercise
Each student should prepare her own course portfolio!

Portfolios should include the following parts:

1) Course Syllabus
2) Lecture notes (slides)
3) Assignments
4) Quizzes
5) Mid-term exam and answer sheet.
6) Research articles and other supporting materials.
7) Lab lecture notes, exercises, and MATLAB codes.
8) Glossary

Portfolios will be checked regularly by the instructor.

Students who prepare good course portfolios may be given a BONUS +2/+5 on their examinations, if needed.

Reminder

Next Lecture!

Next Lecture...

Classification using Nearest Neighbour Algorithm (Ch. 2)

- Be ready!
- Do your self-assessment exercise.
- Prepare your course portfolio.
- Download & print the lecture notes before your class.
Data Mining Project
20 marks for writing a review paper

- What is a review paper?
- Discussion on project for writing a concise review paper.
- Group formulation and leader selection.
- Format and tips for paper writing.
- More: www.c2learn.com/data_mining/

Thank You!

Website: www.c2learn.com
✉️ s.zahrani@tu.edu.sa
admin@c2learn.com