Data Mining Midterm Exam

10.04.2014

Instructions for Students

- Write your name, student number, and signature on the exam sheet.
- The duration of the whole mid-term exam is 1 hour and 30 minutes.
- This is a closed book exam. The only resources allowed to use are blank paper, pens, and your head.

Good luck!

Reserved for the Teacher

<table>
<thead>
<tr>
<th>Max. points</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>
Multiple Choice Questions (4 points)

1. Assume that one of the attributes describing students is Exam which can take one of two values “pass” or “fail”. If we want to put more emphasis on students who passed the exam when analyzing the data, what should be the type of the Exam attribute:

- symmetric binary
- asymmetric binary
- nominal

2. Normalization is used when attribute values are measured on different scales to convert them to a common scale. The goal is to avoid that an attribute dominates some other attributes when computing similarity between data objects. For example, the age attribute goes from 0 to 100 while incomes are in the order of thousands of Euros. What is the problem that you might encounter doing normalization?

- sensitivity to outliers
- biased similarity measures
- unbounded values

3. Imagine you are working in a bank and you are asked to manage loan applications. Your task is to select the criteria that the financial committee should take into account to make the decision about approving the loan or not. By analyzing past loan applications, you describe each applicant by a set of features and assign to him/her a class label “successful” or “failed”. Which of the following techniques help you to achieve your task?

- Bayesian classifiers
- Nearest Neighbor classifiers
- Decision trees

4. Imagine you are responsible for fixing the price of a new mid-class Italian product in the Chinese market. The first thing you need to do is to analyze the average amount that Chinese costumers spend for similar products. However, the data about such costumers is huge, so you need to take your decision about the price based on some samples. The price you decide would be more suitable when:

- the sample contains costumers with representative incomes
- the sample does not contain costumers of high quality products and costumers of low quality products
- the sample contains costumers with representative spending behaviors
Classification Algorithms (5 points)

1. Briefly, what are the main steps to build a Bayesian Network?

2. In decision trees, attribute selection techniques decide the goodness of a nominal attribute based on the purity of its relevant partitions. Each partition contains only tuples that have the same value of that attribute. Explain how attribute selection techniques deal with numeric attributes?

3. Some nonlinear regression models can be converted to linear models by applying transformations to the predictor variables. Show how the nonlinear regression equation $y = \alpha X^\beta$ can be converted to a linear regression equation solvable by the method of least squares.

4. Given the data in the table below, and given a data tuple having the values “systems”, “26...30”, and “4650K” for the attributes department, age, and salary, respectively, what would a Naive Bayesian classification of the status for the tuple be?

<table>
<thead>
<tr>
<th>department</th>
<th>status</th>
<th>age</th>
<th>salary</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>sales</td>
<td>senior</td>
<td>31...35</td>
<td>46K...50K</td>
<td>30</td>
</tr>
<tr>
<td>sales</td>
<td>junior</td>
<td>26...30</td>
<td>26K...30K</td>
<td>40</td>
</tr>
<tr>
<td>sales</td>
<td>junior</td>
<td>31...35</td>
<td>31K...35K</td>
<td>40</td>
</tr>
<tr>
<td>systems</td>
<td>junior</td>
<td>21...25</td>
<td>46K...50K</td>
<td>20</td>
</tr>
<tr>
<td>systems</td>
<td>senior</td>
<td>31...35</td>
<td>66K...70K</td>
<td>5</td>
</tr>
<tr>
<td>systems</td>
<td>junior</td>
<td>26...30</td>
<td>46K...50K</td>
<td>3</td>
</tr>
<tr>
<td>systems</td>
<td>senior</td>
<td>41...45</td>
<td>66K...70K</td>
<td>3</td>
</tr>
<tr>
<td>marketing</td>
<td>senior</td>
<td>36...40</td>
<td>46K...50K</td>
<td>10</td>
</tr>
<tr>
<td>marketing</td>
<td>junior</td>
<td>31...35</td>
<td>41K...45K</td>
<td>4</td>
</tr>
<tr>
<td>secretary</td>
<td>senior</td>
<td>46...50</td>
<td>36K...40K</td>
<td>4</td>
</tr>
<tr>
<td>secretary</td>
<td>junior</td>
<td>26...30</td>
<td>26K...30K</td>
<td>6</td>
</tr>
</tbody>
</table>
Problem Solving (6 points)

1. Assume we have two classes: positive opinion and negative opinion.

   (a) Describe how would you classify a sentence into these two classes.

   (b) Based on your approach, how would you classify the following sentences:
       Sentence1: “the company is doing a good job by hiring people with excellent competencies”
       Sentence2: “employees are competent but they work in bad conditions which often leads them to depression”

   (c) Do encounter any problem with the following sentences?
       Sentence3: “Marco is very serious in his work”
       Sentence4: “Maria’s health condition is serious”
2. Assume the Weather changes from \textit{sunny} to \textit{rainy} in a random way.

(a) Draw a Markov Model that represents the situation and give the prior distribution on the states as well as the transition matrix.

(b) You observe that 70\% of people are in a good mood when it is sunny and 40\% of people are in a bad mood when it is raining. Assuming that the current mood state does not depend on the previous mood state, how would you update your model to capture mood change.

(c) Draw the corresponding Bayesian network for the first three days.