Subject Withdrawal Suggestion Using Naïve Bayes
Chonthicha Luaarunsawad, Panarat Duke, Peerapol Moemeng
Faculty of Science and Technology, Assumption University
Bangkok, Thailand
{u4412831, u4425022, peerapolmmn}@au.edu

Abstract

Every semester after the midterm exams, almost all advisors have been asked by some of their students if they should drop this subject. This research attempts to model the answer to this question. The paper discusses how a machine-learning algorithm called, the ‘Bayesian Network’, is used in this context, including what are the problems we are facing in this attempt.

Keywords: Bayesian learning, formalism, machine-learning algorithm, rule-based algorithm, decision tree-type algorithm, classifier method.

Introduction

The Bayesian Network (BN) formalism was invented to allow efficient representation of mathematic reasoning with uncertain knowledge. This approach allows for learning from experience, and it combines the best of classical AI and neural networks. It promoted the idea of normative expert systems; ones that act rationally according to the laws of decision theory and do not try to imitate the thought processes of human experts.

This paper discusses two methods that allow us to model the advice that an advisor might be able to give to a concerned student about proceeding further in a given course. In addition, two other algorithms are mentioned to compare with Bayesian Network, namely Rule-based and Decision Tree type algorithms.

The constraint of this project is time limitation and available information limitation, leading to the difficulty of designing networks (i.e. the attributes will be limited).

What is Bayesian Network?

A Bayesian Network is a data structure, which is used to answer any answer (query) by summing up all the relevant joint entries. The answer is the optimal decision that can be made by reasoning out these probabilities together with observed data.

The basis for all Bayesian learning algorithm is the Bayes Rule.

\[
P(h|D) = \frac{P(D|h)P(h)}{P(D)}
\]

Where:

- \(P(h)\) = prior probability of hypothesis \(h\)
- \(P(D)\) = prior probability of training data \(D\)
- \(P(h|D)\) = posterior probability of \(h\) given \(D\)
- \(P(D|h)\) = probability of \(D\) given \(h\)

BN derived from Bayes rule is represented as:

\[
P(x_1, \ldots, x_n) = \prod P(x_i|\text{parents}(x_i))
\]

Node ordering in BN is a correct representation of the domain only if each node is conditionally independent of its predecessors.

The notion of conditional probabilities is:

\[
P(Y) = \sum P(Y|z) P(z)
\]

The notion of independence between propositions \(a\) and \(b\) can be written as:

\[
P(a|b) = P(a) \text{ or } P(b|a) = P(b) \text{ or } P(a\lor b)
\]

= \(P(a)P(b)\)
Available Information: The student’s scores in subject (SC-2212) in three semesters consist of:
- Quiz
- Midterm
- Assignment
- Absent (for a ‘no-show’)

Partial designs of a withdrawal suggestion program is as follows:
1. Task T: Estimated probability of the student who failed the subject.
2. Performance measure P: Percent of correct prediction.
3. Training experience E: New training example in different cases.

The three items above correspond to the specification of the learning task.

BN Method

Two methods are adopted to apply in this project.

Naïve Bayes Classifier, one highly practical Bayesian learning method is shown in the first part of this paper while the Bayesian Network is shown in the second part.

Naïve Bayes Classifier Method

This is a highly practical Bayesian learning method, and it is used together with our application because of its simplicity and efficiency of method.

The learner is asked to predict the target value, or classification, for this new instance.

Naïve Bayes Classifier:

\[
v_{NB} = \arg\max_{i} P(v_j) P(p(a_i|v_j))
\]

Where \(v_{NB}\) denotes the target value output by Naïve Basian Classifier.

The hypothesis, in this method, is formed without searching, but counting the frequency of various data combinations within the training examples.

Our task is to predict the target value (yes or no) of the target concept that failed in this new instance.

From the available data of the subject SC-2212, we defined all the variables as listed below:

\[
h = \{\text{failed (yes, no)}\}
\]

\[
D = \{\text{quiz (less, more)},
\text{Midterm (less, more)},
\text{Assignment (less, more)},
\text{Absent (0,1,2,3,4)}\}
\]

\[
d_1 = \{\text{less,less,more,3}\}
\]

Table 1 is created by adjusting raw data to fit with this method.

Table 1. Data transformation from raw data to comply with the expected representation

<table>
<thead>
<tr>
<th>Student</th>
<th>Quiz</th>
<th>Assignment</th>
<th>Midterm</th>
<th>Absent</th>
<th>Failed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>less</td>
<td>less</td>
<td>less</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>less</td>
<td>less</td>
<td>less</td>
<td>0</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>less</td>
<td>more</td>
<td>less</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>less</td>
<td>less</td>
<td>less</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>more</td>
<td>more</td>
<td>less</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>⋮</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>more</td>
<td>more</td>
<td>less</td>
<td>3</td>
<td>No</td>
</tr>
</tbody>
</table>

We can find the target value (nNB) of the student who failed the exam from the observation sample (scored less than half in the quiz and less than half in the midterm and less than half in an assignment).

Instantiating the Naive Bayes equation to fit the current task, the target value \(v_{NB}\) is given by:

\[
v_{NB} = \arg\max_{i} P(v_j) P(p(a_i|v_j))
\]

\[
= \arg\max_{i} P(\text{failed})
\]

\[
P(\text{Quiz}=\text{less}|\text{failed}=\text{yes})
\]

\[
P(\text{Midterm}=\text{less}|\text{failed}=\text{yes})
\]

\[
P(\text{Assignment}=\text{less}|\text{failed}=\text{yes})
\]

The probabilities of different target values can be estimated, based on their frequencies over the 74 examples.

\[
P(\text{Failed} = \text{yes}) = 6/74 = 0.09
\]

\[
P(\text{Failed} = \text{no}) = 68/74 = 0.91
\]

The conditional probabilities are estimated. For example, those for Quiz = less are:

\[
P(\text{Quiz} = \text{less}|\text{Failed} = \text{yes}) = 6/6 = 1
\]
P(Quiz = more|Failed = yes) = 0/68 = 0
P(Quiz = less|Failed = no) = 20/68 = 0.294
P(Quiz = more|Failed = no) = 48/68 = 0.705

We calculate \( \nu \) as follows:

\[
P(\text{Failed}=\text{no})P(\text{Quiz}=\text{less}|\text{no}) \\
P(\text{Mid}=\text{less}|\text{no})P(\text{Assign}=\text{less}|\text{no}) \\
P(\text{Absent}=3|\text{no})
\]

\[
= 0.000594
\]

\[
P(\text{Failed}=\text{yes})P(\text{Quiz}=\text{less}|\text{yes}) \\
P(\text{Mid}=\text{less}|\text{yes})P(\text{Assign}=\text{less}|\text{yes}) \\
P(\text{Absent}=3|\text{yes})
\]

\[
= 0.028153
\]

According to the questions, and using these estimated probabilities of these joint distributions, the answer to this question is, “the student will pass this SC-2212 course”.

Thus, the NBC assigns the target value \( \text{Failure} = \text{yes} \) to this new instance, based on the probability estimates learned from the training data.

By normalizing \( P(h|D) \), the target value is \( \text{yes} \), given the observed attribute values, this probability is:

\[
\frac{0.028153}{0.000594+0.028153} = 0.97
\]

We can interpret that the chance of the student who will fail SC-2212 by the given conditions that a student for whom the quiz < half, Mid < half, Assignment < half and absent = 3 is 0.97

**Bayesian Network Method**

We can construct the Bayesian Network by analyzing those available data sets.

Finding correlation coefficient (R2) of those attributes allows us to find the relationship between each attribute. The attribute value and target value pairings have changed many times to find the highest R2 or most reliable relationship. Graph 1 is one of the examples of R2.

**Graph 1.** The number of R² of the relationship between student who get quiz less than half VS total scores (Midterm+Assignment)

**Fig. 1.** The number of R² of the relationship between the student whose quiz is less than half VS total scores (Midterm + Assignment)

Finding the correlation coefficient of each pair of data. The BN, is created by using R2, shown in the figure below:

The relationship of node ‘Absent’ and ‘Failed’ is eliminated because of the weak
relationship. Also, the relationships between attributes are very weak, indicating that each attribute is conditionally independent to the others, and therefore we have to separate the data into three networks. This will help in reducing the size of the domain representation and the complexity of the inference problem.

Identifies the variable using Bayes theorem:

\[ H: \{\text{failed}(\text{yes, no}), \neg\text{failed}(\text{yes, no})\} \]
\[ D_1: \{\text{Quiz}<\text{Half}(\text{true, false})\} \]
\[ D_2: \{\text{Midterm}<\text{Half}(\text{true, false})\} \]
\[ D_3: \{\text{Assignment}<\text{Half}(\text{true, false})\} \]

Table 2 shows the number of students who scored less than half in the quiz, midterm and assignment.

<table>
<thead>
<tr>
<th></th>
<th>Q&lt;\text{H}</th>
<th>M&lt;\text{H}</th>
<th>A&lt;\text{H}</th>
<th>\neg\text{Q}&lt;\text{H}</th>
<th>\neg\text{M}&lt;\text{H}</th>
<th>\neg\text{A}&lt;\text{H}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\neg\text{Failed}</td>
<td>7</td>
<td>29</td>
<td>1</td>
<td>41</td>
<td>32</td>
<td>60</td>
</tr>
<tr>
<td>\text{Failed}</td>
<td>6</td>
<td>12</td>
<td>6</td>
<td>20</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

The probabilities in Table 3 were calculated from counting the frequency of the training data. The data were separated into three groups, represented by different colors, according to the networks diagram.

Table 3. The probabilities in the table below were calculated from counting the frequency of the training data. The data was separated into three groups, represented by different colors, according to the networks diagram:

<table>
<thead>
<tr>
<th></th>
<th>Q&lt;\text{H}</th>
<th>\neg\text{Q}&lt;\text{H}</th>
<th>M&lt;\text{H}</th>
<th>\neg\text{M}&lt;\text{H}</th>
<th>A&lt;\text{H}</th>
<th>\neg\text{A}&lt;\text{H}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\neg\text{Failed}</td>
<td>0.09</td>
<td>0.55</td>
<td>0.39</td>
<td>0.43</td>
<td>0.01</td>
<td>0.81</td>
</tr>
<tr>
<td>\text{Failed}</td>
<td>0.08</td>
<td>0.27</td>
<td>0.16</td>
<td>0.01</td>
<td>0.08</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Therefore, we get a probability of for Group 1:

\[ P(\text{failed}) = 0.18 \]
\[ P(\neg\text{failed}) = 0.82 \]
\[ P(\text{Q}<\text{H}|\text{failed}) = 0.46 \]
\[ P(\neg(\text{Q}<\text{H})|\text{failed}) = 0.54 \]
\[ P(\text{Q}<\text{H}|\neg\text{failed}) = 0.33 \]
\[ P(\neg(\text{Q}<\text{H})|\neg\text{failed}) = 0.67 \]

Therefore, we get a probability of for Group 2:

\[ P(\text{failed}) = 0.18 \]
\[ P(\neg\text{failed}) = 0.82 \]
\[ P(\text{M}<\text{H}|\text{failed}) = 0.16 \]
\[ P(\neg(\text{M}<\text{H})|\text{failed}) = 0.01 \]
\[ P(\text{M}<\text{H}|\neg\text{failed}) = 0.39 \]
\[ P(\neg(\text{M}<\text{H})|\neg\text{failed}) = 0.43 \]

Therefore, we get a probability of for Group 3:

\[ P(\text{failed}) = 0.18 \]
\[ P(\neg\text{failed}) = 0.82 \]
\[ P(\text{A}<\text{H}|\text{failed}) = 0.08 \]
\[ P(\neg(\text{A}<\text{H})|\text{failed}) = 0.01 \]
\[ P(\text{A}<\text{H}|\neg\text{failed}) = 0.81 \]
\[ P(\neg(\text{A}<\text{H})|\neg\text{failed}) = 0.19 \]

Focusing on finding the most probable hypothesis \( h \in H \) given the observed data D. The maximally probable hypothesis is called a "maximum a posteriori" (MAP) hypothesis, denoted by:

\[ h_{\text{MAP}} = \arg\max_{h \in H} P(D|h) P(h) \]

\[ P(\text{M}<\text{H}|\text{failed}) P(\text{failed}) = (0.16) \]
\[ (0.18) = 0.0288 \]
\[ P(\text{M}<\text{H}|\neg\text{failed}) P(\neg\text{failed}) = (0.39) \]
\[ (0.82) = 0.3198 \]

Therefore, \( h_{\text{MAP}} = \neg\text{failed} \) (passed). It tells that the most probable hypothesis is \( \neg\text{failed} \) given the observed data D. The posterior probabilities can be determined by normalizing.

\[ \frac{0.3198}{0.3198 + 0.0288} = 0.92 \]

We can see that the most probable hypothesis is “not failed--passed” given the observed data, D. We should use this result to test with the new instance. Unfortunately, we have only four attributes and each attributes value is represented by only two discrete
values. Therefore, all of the cases can be calculated without testing the new instances.

The calculations of all cases are listed below:

\[
P(Q<H|\text{failed}) P(\text{failed}) = (0.46) (0.18) = 0.0828
\]

\[
P(Q<H|\sim\text{failed}) P(\sim\text{failed}) = (0.33) (0.82) = 0.2706
\]

Therefore, \(h_{\text{MAP}} = \sim\text{failed} \) (passed). It signifies that the most probable hypothesis is \(\sim\text{failed} \) given the observed data, \(D\). The posterior probabilities can be determined by normalizing:

\[
\frac{0.2706}{0.2706 + 0.0828} = 0.77
\]

\[
P(Q>H|\text{failed}) P(\text{failed}) = (0.54) (0.18) = 0.0972
\]

\[
P(Q>H|\sim\text{failed}) P(\sim\text{failed}) = (0.67) (0.82) = 0.5494
\]

Therefore, \(h_{\text{MAP}} = \sim\text{failed} \) (passed). It signifies that the most probable hypothesis is \(\sim\text{failed} \) given the observed data, \(D\). The posterior probabilities can be determined by normalizing:

\[
\frac{0.5494}{0.5494 + 0.0972} = 0.85
\]

\[
P(M>H|\text{failed}) P(\text{failed}) = (0.01) (0.18) = 0.0018
\]

\[
P(M>H|\sim\text{failed}) P(\sim\text{failed}) = (0.43) (0.82) = 0.3526
\]

Therefore, \(h_{\text{MAP}} = \sim\text{failed} \) (passed). It signifies that the most probable hypothesis is \(\sim\text{failed} \) given the observed data, \(D\). The posterior probabilities can be determined by normalizing:

\[
\frac{0.3526}{0.3526 + 0.0018} = 0.99
\]

\[
P(A<H|\text{failed}) P(\text{failed}) = (0.08) (0.18) = 0.0144
\]

\[
P(A<H|\sim\text{failed}) P(\sim\text{failed}) = (0.01) (0.82) = 0.0082
\]

Therefore, \(h_{\text{MAP}} = \text{failed} \) (passed). It signifies that the most probable hypothesis is \(\text{failed} \) given the observed data, \(D\). The posterior probabilities can be determined by normalizing:

\[
\frac{0.0144}{0.0144 + 0.0082} = 0.64
\]

\[
P(A>H|\text{failed}) P(\text{failed}) = (0.09) (0.18) = 0.0162
\]

\[
P(A>H|\sim\text{failed}) P(\sim\text{failed}) = (0.81) (0.82) = 0.6642
\]

Therefore, \(h_{\text{MAP}} = \sim\text{failed} \) (passed). It signifies that the most probable hypothesis is \(\sim\text{failed} \) given the observed data, \(D\). The posterior probabilities can be determined by normalizing:

\[
\frac{0.6642}{0.6642 + 0.0162} = 0.97
\]

**Other Machine Learning Methods Consideration**

**Ruled-based**

One of the most important aspects of BN, based on semantic networks, is its ability to represent default values for categories.

For example: the target concept contains: `studentMarks( (Quiz < Half) , failed = yes)`

(direct value)

`studentMarks( (Quiz < Half) , failed = no)`

(contradiction)

In a strictly logical KB, this would be a
contradiction. However, in a semantic network, the assertion that studentMarks((Quiz < Half),
failed = yes) has only default status; that is assumed that studentMarks((Quiz < Half), yes),
unless this is contradicted by more specific information. The default semantic is enforced
naturally by the inheritance algorithm, because it follows link upwards from the object itself,
(Quiz<Half) in this case, and stops as soon as it finds a value. We say that the default is
overridden by the more specific value.

The other reason is the properties of the rule-based system is simply not appropriate for
uncertain reasoning.

For example:

<table>
<thead>
<tr>
<th>P(A)</th>
<th>P(B)</th>
<th>P(AB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(H1) = 0.5</td>
<td>P(H1 ^ H1) = 0.5</td>
<td></td>
</tr>
<tr>
<td>P(H) = 0.5</td>
<td>P(T1) = 0.5</td>
<td>P(H1 ^ T1) = 1.0</td>
</tr>
<tr>
<td></td>
<td>P(H2) = 0.5</td>
<td>P(H1 ^ T2) = 0.75</td>
</tr>
</tbody>
</table>

Let:

H1 be the event that a fair coin flip comes up heads,
T1 be the event that the coin comes up tails on that same flip,
H2 be the event that the coin comes up heads on that second flip,

It can be seen that all three events have the same probability, 0.5, and so a truth-
functional system must assign the same belief to the disjunction of any two of them. But, the
probability of the disjunction depends on the events themselves, and not just on the
probabilities.

### Decision Tree

Because this Bayesian method provides a more flexible approach to learning than algorithms that completely eliminates a hypothesis if it is found to be inconsistent with any single example.

The Bayesian method does this by allowing the observed training example to decrease or increase the estimated probability that a hypothesis is correct.

### Summary

This paper is part of the subject, Machine Learning (SC-4388), that shows how to apply
the BN knowledge to solve the problem on how to construct and calculate the probability
from any queries. The time constraint is dealt with by using time management. The
limitation of the information is solved by adjusting the raw data, then finding the best
results to use in forming the networks.

### References

Prentice Hall, New York, NY, USA.
Markus, L. 2004. (Website) Implementation of Naive Bayesian Classifiers in Java.
<http://ai.ijs.si/Mezi/pedagosko/markuslang-seminar.doc>