Tutorial Business Analytics

Tutorial 5 - Solution

Exercise 5.1

The following table contains empirical values about your past decisions whether or not to play.

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

a) Calculate the rule set using the 1-Rule classification. Would you play if it is windy with high humidity, a sunny outlook and cool temperature? Utilize the 1-Rule classification to determine the answer.

b) Would you play if it is windy with high humidity, a sunny outlook and cool temperature? Determine the answer using Naïve Bayes classification.

c) Would you play if the evidence of outlook from b) were changed to overcast? Determine the answer using Naïve Bayes classification.

Solution

a) The 1-Rule algorithm predicts a new dataset on the basis of error evaluations regarding the training set by identifying the attribute with the smallest error.

Step 1: Enumerate all possible attribute values and their frequency for each class, in this case Play = yes/no and calculate (overall) errors for each attribute.

Step 2: Choose a single attribute with the smallest error rate for prediction.

Outlook	Play = yes	Play = no	Error
Sunny	2	3	2/5
Overcast	4	0	0/4
Rainy	3	2	2/5
Σ	**9**	**5**	**(2+0+2)/(5+4+5)=4/14**

Temperature	Play = yes	Play = no	Error
Hot	2	2	2/4
Mild	4	2	2/6
Cool	3	1	1/4
Σ	**9**	**5**	**(2+2+1)/(4+6+4)=5/14**

Humidity	Play = yes	Play = no	Error
High	3	4	3/7
Normal	6	1	1/7
Σ	**9**	**5**	**(3+1)/(7+7)=4/14**

Windy	Play = yes	Play = no	Error
False	6	2	2/8
True	3	3	3/6
Σ	**9**	**5**	**(2+3)/(8+6)=5/14**

In this case, both attributes “Outlook” and “Humidity” reveal an error rate of 4/14. You can randomly choose either of these attributes for your rule set.

For new data set “windy=true, humidity=high, outlook=sunny and temperature=cool” we therefore have two variants:

Random selection A) Outlook=sunny → Play=no
Random selection B) Humidity=high → Play=no
→ With either of the selections, the prediction for the new data set is “Play=no”.

Problems of 1-Rule:

- Uses only a single attribute for classification
- How to deal with missing values, e.g. values for “Outlook” or “Humidity”?
- How to deal with numeric values? → Solution: Discretization of numerical values, but increase of class complexity

b) Preconditions for Naïve Bayes:

- All attributes independent from each other
- All attributes are equally important

Quick recap on Naïve Bayes:

1. Enumerate all possible attribute characteristics and their frequency for each class, in this case Play = yes/no.
2. Identify zero-frequency problem, resolve if needed.
3. Calculate prior probability $Pr[h_i]$ and likelihood $Pr[e_i \mid h_i]$.
4. Calculate posterior probability $Pr[h_i \mid E]$ using Bayes' theorem.

Play	
Yes	**9/14**
No	**5/14**
Σ	**1**

Outlook	**Play = yes**	**Play = no**
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5
Σ	**1**	**1**

Temperature	**Play = yes**	**Play = no**
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5
Σ	**1**	**1**

Humidity	**Play = yes**	**Play = no**
High	3/9	4/5
Normal	6/9	1/5
Σ	**1**	**1**

Windy	**Play = yes**	**Play = no**
False	6/9	2/5
True	3/9	3/5
Σ	**1**	**1**

The next step encompasses the classification of a given data tuple. Let:
$E = \{e_1$ (outlook=sunny), e_2 (temperature=cool), e_3 (humidity=high), e_4 (windy=true)$\}$ and $h \in$ {Play=yes, Play=no}:

Probabilities	**h_1 (Play=yes)**	**h_2 (Play=no)**
$Pr[e_1 \mid h_i]$	2/9	3/5
$Pr[e_2 \mid h_i]$	3/9	1/5
$Pr[e_3 \mid h_i]$	3/9	4/5
$Pr[e_4 \mid h_i]$	3/9	3/5
$Pr[h_i]$	9/14	5/14

$A = \prod (Pr[e_i \mid \text{Play=yes}]) \cdot Pr[\text{Play=yes}] = 2/9 \cdot 3/9 \cdot 3/9 \cdot 3/9 \cdot 9/14 = 0.0053$
$B = \prod (Pr[e_i \mid \text{Play=no}]) \cdot Pr[\text{Play=no}] = 3/5 \cdot 1/5 \cdot 4/5 \cdot 3/5 \cdot 5/14 = 0.0206$

Normalization:
$Pr[\text{Play=yes} \mid E] = A / (A+B) = 0.205$
$Pr[\text{Play=no} \mid E] = B / (A+B) = 0.795$

Answer: Based on the calculation, we predict "Play=no" for the given data tuple E to maximize the likelihood of a correct prediction.

c) Naïve Bayes calculation with a zero-frequency problem

To illustrate how to deal with the zero-frequency problem, we adapt the given data tuple E slightly. The new data tuple E' is constituted as follows:

E' = {e_1 (outlook=~~sunny~~ overcast), e_2 (temperature=cool), e_3 (humidity=high), e_4 (windy=true)}

and h ∈ {Play=yes, Play=no}.

Play	
Yes	**9/14**
No	**5/14**
Σ	**1**

Outlook	Play = yes	Play = no
Sunny	2/9	3/5
Overcast	4/9	**0/5**
Rainy	3/9	2/5
Σ	**1**	**1**

For the attribute value "outlook=overcast", we now encounter a probability value of zero for Pr[e_1 | Play=no], which would in turn yield an aggregated probability of zero for Pr[Play=no | E]. The solution is to add one to each absolute frequency in the attribute frequency tables (cf. next page), a.k.a. the *Laplacian correction*.

Temperature	Play = yes	Play = no
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5
Σ	**1**	**1**

Humidity	Play = yes	Play = no
High	3/9	4/5
Normal	6/9	1/5
Σ	**1**	**1**

Windy	Play = yes	Play = no
False	6/9	2/5
True	3/9	3/5
Σ	**1**	**1**

Tables after solving the zero-frequency problem by adding one to the count of the whole attribute value frequency table:

Play	
Yes	**9/14**
No	**5/14**
Σ	**1**

Outlook	Play = yes	Play = no
Sunny	3/12	4/8
Overcast	5/12	1/8
Rainy	4/12	3/8
Σ	**1**	**1**

In this case, +1 has been added to the count of each table cell, therefore e.g. Pr[outlook=overcast | Play=no) turned from 0/5 to (0+1)/(5+3) =1/8 due to three existing attribute values.

Temperature	Play = yes	Play = no
Hot	3/12	**3/8**
Mild	5/12	**3/8**

Cool	4/12	**2/8**
Σ	**1**	**1**

Humidity	**Play = yes**	**Play = no**
High	4/11	**5/7**
Normal	7/11	**2/7**
Σ	**1**	**1**

Windy	**Play = yes**	**Play = no**
False	7/11	**3/7**
True	4/11	**4/7**
Σ	**1**	**1**

The next step encompasses the classification of the slightly changed data tuple E'

Probabilities	**h_1 (Play=yes)**	**h_2 (Play=no)**
$Pr[e_1 \mid h_i]$	5/12	1/8
$Pr[e_2 \mid h_i]$	4/12	2/8
$Pr[e_3 \mid h_i]$	4/11	5/7
$Pr[e_4 \mid h_i]$	4/11	4/7
$Pr[h_i]$	9/14	5/14

$A = \prod (Pr[e_i \mid Play=yes]) \cdot Pr[Play=yes] = 5/12 \cdot 4/12 \cdot 4/11 \cdot 4/11 \cdot 9/14 = 0.0118$
$B = \prod (Pr[e_i \mid Play=no]) \cdot Pr[Play=no] = 1/8 \cdot 2/8 \cdot 5/7 \cdot 4/7 \cdot 5/14 = 0.0046$

Normalization:
Pr[Play=yes | E] = A / (A+B) = 0.0118 / (0.0118 + 0.0046) = 0.720
Pr[Play=no | E] = B / (A+B) = 0.0046 + (0.0118 + 0.0046) = 0.280

Answer: Based on the calculation, we predict "Play=yes" for the given data tuple E' to maximize the likelihood of a correct prediction.

Exercise 5.2 - Bayesian Networks

Consider the following Boolean random variables related to the state of a given car:

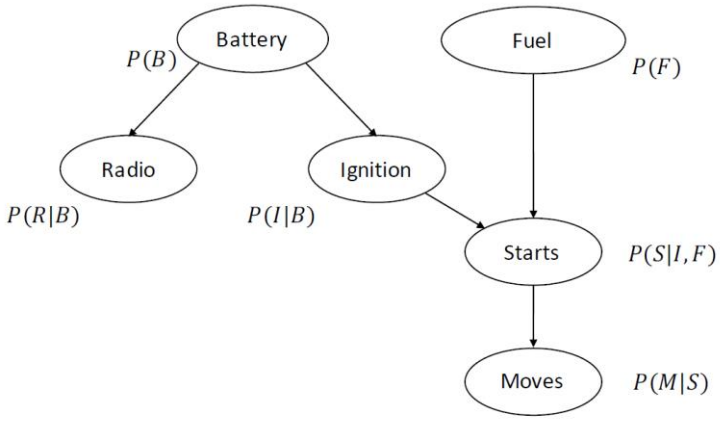
- Battery (B): is the battery charged?
- Fuel (F): is fuel tank empty?
- Ignition (I): does the ignition system work?
- Moves (M): does the car move?
- Radio (R): can the radio be switched on?
- Starts (S): does the engine fire?

a) Represent the joint probability density function using a Bayesian network.

b) Rewrite the probabilities using the chain rule, after defining a proper set of causal relationships between the variables.

Solution

a) Bayesian network:



The state of the battery and of the fuel tank can be seen as the “root causes”. They can also be considered independent on each other. Whether the radio works or not directly depends only on the battery state. The same for the ignition system, whose functioning can be considered independent on that of the radio, given the state of the battery. The state of the fuel tank, together with the ignition system, directly influences whether the engine fires. Finally, it can be assumed that whether the car moves or not directly depends only on the engine state.

b) Accordingly, the considered random variables can be sorted from the “root causes” to the “end effects" as follows: Fuel, Battery, Radio, Ignition, Starts, and Moves.

Their joint probability can then be factorized through the chain rule as follows:

$$\Pr[F,B,R,I,S,M] = \Pr[F]\cdot\Pr[B|F]\cdot\Pr[R|B,F]\cdot\Pr[I|R,B,F] \cdot\Pr[S|I,R,B,F]\ \cdot\Pr[M|S,I,R,B,F].$$

The conditional independence assumptions described above finally lead to simplify the joint probability as follows:

$$\Pr[F,B,R,I,S,M] = \Pr[F]\cdot\Pr[B]\cdot\Pr[R|B]\cdot\Pr[I|B]\cdot\Pr[S|I,F]\cdot\Pr[M|S].$$

Exercise 5.3

Note: Use R to solve this exercise(Exercise 5.3_R-template.R).

Load the training data ("loan-train.csv") and the test data("loan-test.csv") into R. Proceed by typing *names(train)* to print the attribute names to the console.

```
library(tidyverse)
train = read_csv("loan-train.csv")
test  = read_csv("loan-test.csv")
names(train)
```

The dependent binary variable *"loan"* indicates whether an installment loan has been repaid without any issues. Except for the attribute *"age"*, which is numerical, all other independent variables are categorical. You will find further information on what each variable tells us in the file *"variables.rtf"*. We want to create a prediction model using Naïve Bayes.
Note: Not all functions you need are given in the exercise definition, check the provided R-template script for them.

a) Transform the independent attribute *"age"* into a categorical attribute by placing the values into buckets.
   Why are categorical variables preferable when using Naïve Bayes? What problems can occur with numerical data?

b) For using Naïve Bayes functions, import the *"e1071"* library first.

```
library(e1071)
```

   Iterate through the independent attributes to find the most suitable attribute for the 1-rule classification. What attribute would you use for a 1-rule classification?

c) Create a prediction model using the Naïve Bayes classifier and apply it on the test-dataset. Build a confusion matrix and determine the model's error rate.

Solution: Exercise 5.3_R-Script.R