RAZAVET Mael 13/02/2014

Data Mining – Exercise 5

**Question1:**

The predicted class here is the sex of a person, so either “Male” or “Female”.

The two parameters that we need for this data to be classified using a Naïve Bayes classifier model are the prior probability distribution and the likelihood function because this dataset contains discrete attributes.

**Question2 :**

HairEyeColorDF <- as.data.frame(HairEyeColor)

HEC <- HairEyeColorDF[rep(row.names(HairEyeColorDF), HairEyeColorDF$Freq), 1:3]

To compute the prior probability distribution, I used this command :

table(HEC$Sex)

returns

Male Female

279 313

The goal of this command is to display the frequency table for the two class labels we want to predict.

P(sex=Male)=279/(279+313)= 0.4712838

P(sex=Female)=313/(279+313)= 0.5287162

The likelihood function aims at classifying instances and is noted: P(x|Ci) with x a single instance and Ci a class label (Female or Male). So, I will need the contingency table for each conditional probability.

table(HEC$Hair, HEC$Sex) returns

Male Female

Black 56 52

Brown 143 143

Red 34 37

Blond 46 81

P(Hair=Black|Sex=Male)=56/279= 0.2007168

P(Hair=Brown|Sex=Male)=143/279= 0.5125448

P(Hair=Red|Sex=Male)=34/279= 0.1218638

P(Hair=Blond|Sex=Male)=46/279= 0.1648746

P(Hair=Black|Sex=Female)=52/313= 0.1661342

P(Hair=Brown|Sex=Female)=143/313= 0.456869

P(Hair=Red|Sex=Female)=37/313= 0.1182109

P(Hair=Blond|Sex=Female)=81/313= 0.2587859

table(HEC$Eye, HEC$Sex) returns

Male Female

Brown 98 122

Blue 101 114

Hazel 47 46

Green 33 31

P(Eye=Brown|Sex=Male)=98/279= 0.3512545

P(Eye=Blue|Sex=Male)=101/279= 0.3620072

P(Eye=Hazel|Sex=Male)=47/279= 0.1684588

P(Eye=Green|Sex=Male)=33/279= 0.1182796

P(Eye=Brown|Sex=Female)=122/313= 0.3897764

P(Eye=Blue|Sex=Female)=114/313= 0.3642173

P(Eye=Hazel|Sex=Female)=46/313= 0.1469649

P(Eye=Green|Sex=Female)=31/313= 0.09904153

**Question 3:**

NB=function(df, class){

DF <- table(df[,class])

res <- list()

for(attribute in names(df)){

if(attribute != class)

{

x <- table(df[,class], df[,attribute])

res[[attribute]] <- prop.table(x, 1)

}

}

return (list(prop.table(DF), res))

}

NB(HEC, 'Sex') returns

[[1]]

Male Female

0.4712838 0.5287162

[[2]]

[[2]]$Hair

Black Brown Red Blond

Male 0.2007168 0.5125448 0.1218638 0.1648746

Female 0.1661342 0.4568690 0.1182109 0.2587859

[[2]]$Eye

Brown Blue Hazel Green

Male 0.35125448 0.36200717 0.16845878 0.11827957

Female 0.38977636 0.36421725 0.14696486 0.09904153

naiveBayes(Sex ~ ., data = HEC)

returns

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

Male Female

0.4712838 0.5287162

Conditional probabilities:

Hair

Y Black Brown Red Blond

Male 0.2007168 0.5125448 0.1218638 0.1648746

Female 0.1661342 0.4568690 0.1182109 0.2587859

Eye

Y Brown Blue Hazel Green

Male 0.35125448 0.36200717 0.16845878 0.11827957

Female 0.38977636 0.36421725 0.14696486 0.09904153

This confirms the results I found in question 2.

**Question 4:**

If one of the features were zero, then the conditional probability for this feature would be equal to zero as the frequency of this feature equals to 0.

For example, if there were no red haired men in the dataset, then I would have:

P(Hair=Red|Sex=Male)=0

To remedy this, we can implement the Laplacian correction, which consists in adding one more instance for each pair of values for this feature.

For example, if there were no red haired men in the dataset, we would add one more instance to each Red-value pair, which that we would add one instance for a red haired man, one more instance for brown haired men, one more instance for black haired men and one more instance for blond haired men.

To do that, we need to have a large dataset, with a large number of instances because adding these one more instances to each pair, has to stay negligible compared to the rest of the dataset.

**Question 5:**

If I recall the dataset irisMissing from exercise 3, only the attribute Sepal.Width had missing values.

First of all, I need to train the Naïve Bayes Model on this irisMissing dataset in order to compute the parameters of the model that will help to classify the class of this attribute (having missing values).

To do that, I need to extract the instances that don’t have any missing value in this column and from this new dataset, I need to discretize the values of Sepal.Width as it is a continuous variable. I will discrete the values into 3 equal-width bins.

The three intervals are:

[2,2.8]

(2.8,3.6]

(3.6,4.4]

Note that, beforehand, I discretized the three other numerical attributes in order to avoid overfitting.

The goal is now to predict the interval in which the missing value should be, according to the Naïve Bayes model built beforehand.

So, I extract each instance having a missing value in order to predict these instances.

model <- klaR::NaiveBayes(Sepal.Width ~ ., data = i) #i is the data frame containing only the instances with no missing values

predict(model, iMissing) # iMissing is the data frame containing only the missing values to predict

This “predict” function returns the following labels for the instances:

$class

11 12 23 33 40 47 54

(3.6,4.4] (3.6,4.4] (3.6,4.4] (3.6,4.4] (3.6,4.4] (3.6,4.4] [2,2.8]

63 83 88 102 108 117 129

[2,2.8] [2,2.8] [2,2.8] (2.8,3.6] (2.8,3.6] (2.8,3.6] (2.8,3.6]

147 150

(2.8,3.6] (2.8,3.6]

Levels: [2,2.8] (2.8,3.6] (3.6,4.4]

$posterior

[2,2.8] (2.8,3.6] (3.6,4.4]

11 2.742476e-05 0.3535977 6.463749e-01

12 2.742476e-05 0.3535977 6.463749e-01

23 2.742476e-05 0.3535977 6.463749e-01

33 2.742476e-05 0.3535977 6.463749e-01

40 2.742476e-05 0.3535977 6.463749e-01

47 2.742476e-05 0.3535977 6.463749e-01

54 7.799611e-01 0.2200389 1.967403e-09

63 8.866965e-01 0.1133035 4.217657e-10

83 8.866965e-01 0.1133035 4.217657e-10

88 8.866965e-01 0.1133035 4.217657e-10

102 4.729164e-01 0.5203469 6.736701e-03

108 2.767318e-01 0.7068430 1.642520e-02

117 4.729164e-01 0.5203469 6.736701e-03

129 4.729164e-01 0.5203469 6.736701e-03

147 4.729164e-01 0.5203469 6.736701e-03

150 4.729164e-01 0.5203469 6.736701e-03

The important information are stated in first within the variable $class. It is stating for each instance to be predicted (for example, the instance 11 was a missing value for the feature Sepal.Width), in which bin the prediction would be, according to the Naïve Bayes model (the instance 11 should be between 3.6 and 4.4).

Here is the entire irisMissing data frame:

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

1 5.1 3.5 1.4 0.2 setosa

2 4.9 3 1.4 0.2 setosa

3 4.7 3.2 1.3 0.2 setosa

4 4.6 3.1 1.5 0.2 setosa

5 5.0 3.6 1.4 0.2 setosa

6 5.4 3.9 1.7 0.4 setosa

7 4.6 3.4 1.4 0.3 setosa

8 5.0 3.4 1.5 0.2 setosa

9 4.4 2.9 1.4 0.2 setosa

10 4.9 3.1 1.5 0.1 setosa

11 5.4 (3.6,4.4] 1.5 0.2 setosa

12 4.8 (3.6,4.4] 1.6 0.2 setosa

13 4.8 3 1.4 0.1 setosa

14 4.3 3 1.1 0.1 setosa

15 5.8 4 1.2 0.2 setosa

16 5.7 4.4 1.5 0.4 setosa

17 5.4 3.9 1.3 0.4 setosa

18 5.1 3.5 1.4 0.3 setosa

19 5.7 3.8 1.7 0.3 setosa

20 5.1 3.8 1.5 0.3 setosa

21 5.4 3.4 1.7 0.2 setosa

22 5.1 3.7 1.5 0.4 setosa

23 4.6 (3.6,4.4] 1.0 0.2 setosa

24 5.1 3.3 1.7 0.5 setosa

25 4.8 3.4 1.9 0.2 setosa

26 5.0 3 1.6 0.2 setosa

27 5.0 3.4 1.6 0.4 setosa

28 5.2 3.5 1.5 0.2 setosa

29 5.2 3.4 1.4 0.2 setosa

30 4.7 3.2 1.6 0.2 setosa

31 4.8 3.1 1.6 0.2 setosa

32 5.4 3.4 1.5 0.4 setosa

33 5.2 (3.6,4.4] 1.5 0.1 setosa

34 5.5 4.2 1.4 0.2 setosa

35 4.9 3.1 1.5 0.2 setosa

36 5.0 3.2 1.2 0.2 setosa

37 5.5 3.5 1.3 0.2 setosa

38 4.9 3.6 1.4 0.1 setosa

39 4.4 3 1.3 0.2 setosa

40 5.1 (3.6,4.4] 1.5 0.2 setosa

41 5.0 3.5 1.3 0.3 setosa

42 4.5 2.3 1.3 0.3 setosa

43 4.4 3.2 1.3 0.2 setosa

44 5.0 3.5 1.6 0.6 setosa

45 5.1 3.8 1.9 0.4 setosa

46 4.8 3 1.4 0.3 setosa

47 5.1 (3.6,4.4] 1.6 0.2 setosa

48 4.6 3.2 1.4 0.2 setosa

49 5.3 3.7 1.5 0.2 setosa

50 5.0 3.3 1.4 0.2 setosa

51 7.0 3.2 4.7 1.4 versicolor

52 6.4 3.2 4.5 1.5 versicolor

53 6.9 3.1 4.9 1.5 versicolor

54 5.5 [2,2.8] 4.0 1.3 versicolor

55 6.5 2.8 4.6 1.5 versicolor

56 5.7 2.8 4.5 1.3 versicolor

57 6.3 3.3 4.7 1.6 versicolor

58 4.9 2.4 3.3 1.0 versicolor

59 6.6 2.9 4.6 1.3 versicolor

60 5.2 2.7 3.9 1.4 versicolor

61 5.0 2 3.5 1.0 versicolor

62 5.9 3 4.2 1.5 versicolor

63 6.0 [2,2.8] 4.0 1.0 versicolor

64 6.1 2.9 4.7 1.4 versicolor

65 5.6 2.9 3.6 1.3 versicolor

66 6.7 3.1 4.4 1.4 versicolor

67 5.6 3 4.5 1.5 versicolor

68 5.8 2.7 4.1 1.0 versicolor

69 6.2 2.2 4.5 1.5 versicolor

70 5.6 2.5 3.9 1.1 versicolor

71 5.9 3.2 4.8 1.8 versicolor

72 6.1 2.8 4.0 1.3 versicolor

73 6.3 2.5 4.9 1.5 versicolor

74 6.1 2.8 4.7 1.2 versicolor

75 6.4 2.9 4.3 1.3 versicolor

76 6.6 3 4.4 1.4 versicolor

77 6.8 2.8 4.8 1.4 versicolor

78 6.7 3 5.0 1.7 versicolor

79 6.0 2.9 4.5 1.5 versicolor

80 5.7 2.6 3.5 1.0 versicolor

81 5.5 2.4 3.8 1.1 versicolor

82 5.5 2.4 3.7 1.0 versicolor

83 5.8 [2,2.8] 3.9 1.2 versicolor

84 6.0 2.7 5.1 1.6 versicolor

85 5.4 3 4.5 1.5 versicolor

86 6.0 3.4 4.5 1.6 versicolor

87 6.7 3.1 4.7 1.5 versicolor

88 6.3 [2,2.8] 4.4 1.3 versicolor

89 5.6 3 4.1 1.3 versicolor

90 5.5 2.5 4.0 1.3 versicolor

91 5.5 2.6 4.4 1.2 versicolor

92 6.1 3 4.6 1.4 versicolor

93 5.8 2.6 4.0 1.2 versicolor

94 5.0 2.3 3.3 1.0 versicolor

95 5.6 2.7 4.2 1.3 versicolor

96 5.7 3 4.2 1.2 versicolor

97 5.7 2.9 4.2 1.3 versicolor

98 6.2 2.9 4.3 1.3 versicolor

99 5.1 2.5 3.0 1.1 versicolor

100 5.7 2.8 4.1 1.3 versicolor

101 6.3 3.3 6.0 2.5 virginica

102 5.8 (2.8,3.6] 5.1 1.9 virginica

103 7.1 3 5.9 2.1 virginica

104 6.3 2.9 5.6 1.8 virginica

105 6.5 3 5.8 2.2 virginica

106 7.6 3 6.6 2.1 virginica

107 4.9 2.5 4.5 1.7 virginica

108 7.3 (2.8,3.6] 6.3 1.8 virginica

109 6.7 2.5 5.8 1.8 virginica

110 7.2 3.6 6.1 2.5 virginica

111 6.5 3.2 5.1 2.0 virginica

112 6.4 2.7 5.3 1.9 virginica

113 6.8 3 5.5 2.1 virginica

114 5.7 2.5 5.0 2.0 virginica

115 5.8 2.8 5.1 2.4 virginica

116 6.4 3.2 5.3 2.3 virginica

117 6.5 (2.8,3.6] 5.5 1.8 virginica

118 7.7 3.8 6.7 2.2 virginica

119 7.7 2.6 6.9 2.3 virginica

120 6.0 2.2 5.0 1.5 virginica

121 6.9 3.2 5.7 2.3 virginica

122 5.6 2.8 4.9 2.0 virginica

123 7.7 2.8 6.7 2.0 virginica

124 6.3 2.7 4.9 1.8 virginica

125 6.7 3.3 5.7 2.1 virginica

126 7.2 3.2 6.0 1.8 virginica

127 6.2 2.8 4.8 1.8 virginica

128 6.1 3 4.9 1.8 virginica

129 6.4 (2.8,3.6] 5.6 2.1 virginica

130 7.2 3 5.8 1.6 virginica

131 7.4 2.8 6.1 1.9 virginica

132 7.9 3.8 6.4 2.0 virginica

133 6.4 2.8 5.6 2.2 virginica

134 6.3 2.8 5.1 1.5 virginica

135 6.1 2.6 5.6 1.4 virginica

136 7.7 3 6.1 2.3 virginica

137 6.3 3.4 5.6 2.4 virginica

138 6.4 3.1 5.5 1.8 virginica

139 6.0 3 4.8 1.8 virginica

140 6.9 3.1 5.4 2.1 virginica

141 6.7 3.1 5.6 2.4 virginica

142 6.9 3.1 5.1 2.3 virginica

143 5.8 2.7 5.1 1.9 virginica

144 6.8 3.2 5.9 2.3 virginica

145 6.7 3.3 5.7 2.5 virginica

146 6.7 3 5.2 2.3 virginica

147 6.3 (2.8,3.6] 5.0 1.9 virginica

148 6.5 3 5.2 2.0 virginica

149 6.2 3.4 5.4 2.3 virginica

150 5.9 (2.8,3.6] 5.1 1.8 virginica