Лабораторная работа №1

Наивный Байесовский классификатор

Выполнил

студент гр. 33504/2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Лелюхин Д. О.

Руководитель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Селин И. А.

**Формулировка Задания**

1. Исследуйте, как объем обучающей выборки и количество тестовых данных, влияет на точность классификации или на вероятность ошибочной классификации в примере крестики-нолики и примере о спаме e-mail сообщений.

2. Сгенерируйте 100 точек с двумя признаками X1 и X2 в соответствии с нормальным распределением так, что первые 50 точек (class -1) имеют параметры: мат. ожидание X1 равно 10, мат. ожидание X2 равно 14, среднеквадратические отклонения для обеих переменных равны 4. Вторые 50 точек (class +1) имеют параметры: мат. ожидание X1 равно 20, мат. ожидание X2 равно 18, среднеквадратические отклонения для обеих переменных равны 3. Построить соответствующие диаграммы, иллюстрирующие данные. Построить байесовский классификатор и оценить качество классификации.

3. Разработать байесовский классификатор для данных **Титаник (Titanic dataset) -** <https://www.kaggle.com/c/titanic>

Исходные обучающие данные для классификации – в файле Titanic\_train.csv

Данные для тестирования – в файле Titanic\_test.csv

Использовать функцию read.csv для чтения данных из csv-файлов.

**Классы:**

survival Выжил (0 = No; 1 = Yes)

**Признаки:**

pclass Класс каюты (1 = 1st; 2 = 2nd; 3 = 3rd)

name Имя

sex Пол

age Возраст

sibsp Число братьев-сестер/муж-жена на борту

parch Число родителей/детей на борту

ticket Номер билета

fare Стоимость билета

cabin Каюта

embarked Порт посадки (C = Cherbourg; Q = Queenstown; S = Southampton)

**Специальные отметки**:

Pclass: 1st ~ Верхний; 2nd ~ Средний; 3rd ~ Нижний

Age – в годах; дробный, если возраст меньше одного года

### Выполнение 1-ого задания:

#### Код программы на языке R:

#-------------------1------------------------

library(e1071)

# импортируем данные в R

# установить параметр stringsAsFactors = TRUE,

# так как все данные - категориальные

A\_raw <- read.table("Tic\_tac\_toe.txt", sep = ",", stringsAsFactors = TRUE)

# число строк в базе

n <- dim(A\_raw)[1]

# Создан фрейм, который можно просмотреть, используя str(A\_raw).

# Имеется 9 столбцов признаков V1-V9 и V10 (класс) и

# все имеют один и тот же тип Factor.

# 2 ###########################################################

# Создание обучающей и тестирующей выборки

# Скажем, имеем n примеров в исходной выборке,

# используем 80% для обучения и оставшиеся - для тестирования.

# Устанавливаем базу генерации случайных чисел и рандомизируем выборку

set.seed(12345)

A\_rand <- A\_raw[ order(runif(n)), ]

# разделим данные на обучающие и тестирующие

for (i in seq(0, 1, by = 0.05))

{

nt <- as.integer(n\*i)

A\_train <- A\_rand[1:nt, ]

A\_test <- A\_rand[(nt+1):n, ]

# Можно убедиться, какой имеется процент каждого

# класса V2 в обучающей и тестирующей выборке

prop.table(table(A\_train$V10))

prop.table(table(A\_test$V10))

# 3 ############################################################

# Используем Наивный Байесовский классификатор из пакета e1071

# A\_classifier <- naiveBayes(A\_train[,-10], A\_train$V10)

# Другой вариант классификатора

A\_classifier <- naiveBayes(V10 ~ ., data = A\_train)

# 4 ############################################################

# Теперь оценим полученную модель:

A\_predicted <- predict(A\_classifier, A\_test)

# Используем table для сравнения прогнозируемых значений с тем, что есть

t <- table(A\_predicted, A\_test$V10)

print(t)

}

#### Результаты работы программы:

A\_predicted negative positive

negative 0 0

positive 332 626

A\_predicted negative positive

negative 117 117

positive 198 479

A\_predicted negative positive

negative 133 104

positive 163 463

A\_predicted negative positive

negative 137 117

positive 138 423

A\_predicted negative positive

negative 119 104

positive 139 405

A\_predicted negative positive

negative 99 64

positive 149 407

A\_predicted negative positive

negative 96 73

positive 132 370

A\_predicted negative positive

negative 88 68

positive 118 349

A\_predicted negative positive

negative 80 63

positive 115 317

A\_predicted negative positive

negative 71 59

positive 105 292

A\_predicted negative positive

negative 66 51

positive 91 271

A\_predicted negative positive

negative 62 51

positive 79 240

A\_predicted negative positive

negative 50 43

positive 74 217

A\_predicted negative positive

negative 44 39

positive 62 191

A\_predicted negative positive

negative 34 38

positive 53 163

A\_predicted negative positive

negative 28 28

positive 44 140

A\_predicted negative positive

negative 22 20

positive 40 110

A\_predicted negative positive

negative 17 13

positive 30 84

A\_predicted negative positive

negative 13 7

positive 18 58

A\_predicted negative positive

negative 8 7

positive 7 26

- - - - - -

#### Код программы на языке R:

library(kernlab)

library(e1071)

data(spam)

for(size in seq(20, 4581, by = 500))

{

idx <- sample(1:dim(spam)[1], size)

spamtrain <- spam[-idx, ]

spamtest <- spam[idx, ]

model <- naiveBayes(type ~ ., data = spamtrain)

t <- table(predict(model, spamtest), spamtest$type)

print(t)

}

#### Результаты работы программы:

nonspam spam

nonspam 9 0

spam 9 2

nonspam spam

nonspam 191 11

spam 134 184

nonspam spam

nonspam 336 21

spam 286 377

nonspam spam

nonspam 531 31

spam 373 585

nonspam spam

nonspam 660 33

spam 558 769

nonspam spam

nonspam 825 55

spam 687 953

nonspam spam

nonspam 1038 64

spam 799 1119

nonspam spam

nonspam 1070 55

spam 1070 1325

nonspam spam

nonspam 1155 69

spam 1281 1515

nonspam spam

nonspam 769 177

spam 1976 1598

## Вывод

В результате, мы получаем что, при уменьшении объема обучающей выборки уменьшается количество угаданных результатов.

## Выполнение 2-ого задания:

#### Код программы на языке R:

library(e1071)

x1\_1 <- rnorm(50, mean = 10, sd = 4)

x1\_2 <- rnorm(50, mean = 20, sd = 3)

x2\_1 <- rnorm(50, mean = 14, sd = 4)

x2\_2 <- rnorm(50, mean = 18, sd = 4)

plot(x1\_1, x2\_1, pch=21, xlim=c(0, 30),ylim=c(0, 30))

points(x1\_2,x2\_2, pch=22)

x1 <- c(x1\_1, x1\_2)

x2 <- c(x2\_1, x2\_2)

class<-c(rep('-1',50),rep('1',50))

t<-data.frame(x1, x2, class, stringsAsFactors = TRUE)

for (i in seq(20,80,by=10))

{

idx<-sample(1:dim(t)[1],20)

train<-t[-idx, ]

test<-t[idx, ]

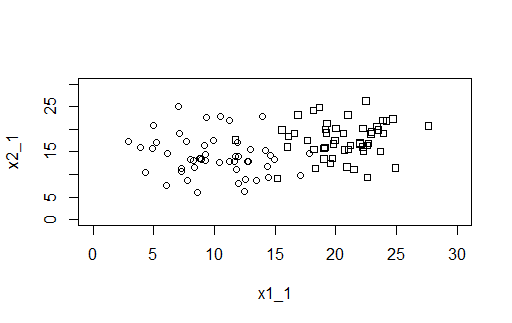
model<-naiveBayes(train[,-3],train$class)

t\_predicted<-predict(model, test)

print(table(t\_predicted,test$class))

}

#### Результаты работы программы:



t\_predicted -1 1

-1 9 1

1 0 10

t\_predicted -1 1

-1 14 0

1 0 6

t\_predicted -1 1

-1 8 2

1 0 10

t\_predicted -1 1

-1 11 0

1 0 9

t\_predicted -1 1

-1 8 0

1 1 11

t\_predicted -1 1

-1 10 0

1 0 10

t\_predicted -1 1

-1 8 1

1 0 11

## Вывод

Объем выборки почти не оказывает влияние на вероятность ошибочной классификации в данном примере.

## Выполнение 3-го задания:

#### Код программы на языке R:

#подключаем библиотеку e1071, содержащую naiveBayes

library(e1071)

#Считываем данные для обучения из csv-файла

train <- read.csv("Titanic\_train.csv")

#Считываем тестовые данные из csv-файла

test <- read.csv("Titanic\_test.csv")

#Вычисляем условные апостериорные вероятности категориальных переменных по Байесу

BayesTitanic <- naiveBayes(as.factor(Survived)~., train)

#str(BayesTitanic)

# возвращает либо те значения, которые модель fit предсказывает

# на основе исходных данных, либо те значения, которые модель

# предсказывает для новых, заданных пользователем данных

BayesPrediction<-predict(BayesTitanic, test)

#str(BayesPrediction)

# выводит обобщенную информацию об объекте а;

# набор статистических параметров, описывающих а,

summary(BayesPrediction)

# создаем таблицу данных

output<-data.frame(test$PassengerId, BayesPrediction)

#str(output)

# объединяет свои аргументы в одну матрицу или таблицу данных по столбцам,

colnames(output)<-cbind("PassengerId","Survived")

# записываем результаты в csv-файл

write.csv(output, file = 'Rushton\_Solution.csv', row.names = F)

#### Результаты работы программы:

|  |
| --- |
| PassengerId,"Survived" |
| 892,"0" |
| 893,"0" |
| 894,"0" |
| 895,"0" |
| 896,"0" |
| 897,"0" |
| 898,"1" |
| 899,"0" |
| 900,"1" |
| 901,"0" |
| 902,"0" |
| 903,"0" |
| 904,"1" |
| 905,"0" |
| 906,"1" |
| 907,"1" |
| 908,"1" |
| 909,"0" |
| 910,"0" |
| 911,"1" |
| 912,"0" |
| 913,"0" |
| 914,"1" |
| 915,"1" |
| 916,"1" |
| 917,"0" |
| 918,"1" |
| 919,"0" |
| 920,"1" |
| 921,"0" |
| 922,"0" |
| 923,"0" |
| 924,"1" |
| 925,"0" |
| 926,"1" |
| 927,"1" |
| 928,"1" |
| 929,"1" |
| 930,"0" |
| 931,"0" |
| 932,"0" |
| 933,"1" |
| 934,"0" |
| 935,"0" |
| 936,"1" |
| 937,"0" |
| 938,"1" |
| 939,"0" |
| 940,"1" |
| 941,"1" |
| 942,"1" |
| 943,"0" |
| 944,"0" |
| 945,"1" |
| 946,"1" |
| 947,"0" |
| 948,"0" |
| 949,"1" |
| 950,"0" |
| 951,"1" |
| 952,"0" |
| 953,"0" |
| 954,"0" |
| 955,"1" |
| 956,"1" |
| 957,"1" |
| 958,"1" |
| 959,"0" |
| 960,"1" |
| 961,"1" |
| 962,"1" |
| 963,"0" |
| 964,"0" |
| 965,"0" |
| 966,"1" |
| 967,"1" |
| 968,"0" |
| 969,"1" |
| 970,"0" |
| 971,"1" |
| 972,"0" |
| 973,"1" |
| 974,"1" |
| 975,"0" |
| 976,"0" |
| 977,"0" |
| 978,"0" |
| 979,"1" |
| 980,"1" |
| 981,"1" |
| 982,"0" |
| 983,"0" |
| 984,"1" |
| 985,"0" |
| 986,"1" |
| 987,"0" |
| 988,"1" |
| 989,"0" |
| 990,"1" |
| 991,"0" |
| 992,"1" |
| 993,"0" |
| 994,"0" |
| 995,"0" |
| 996,"1" |
| 997,"0" |
| 998,"0" |
| 999,"0" |
| 1000,"0" |
| 1001,"1" |
| 1002,"0" |
| 1003,"1" |
| 1004,"1" |
| 1005,"0" |
| 1006,"1" |
| 1007,"0" |
| 1008,"0" |
| 1009,"0" |
| 1010,"1" |
| 1011,"0" |
| 1012,"1" |
| 1013,"0" |
| 1014,"1" |
| 1015,"0" |
| 1016,"0" |
| 1017,"1" |
| 1018,"0" |
| 1019,"0" |
| 1020,"1" |
| 1021,"0" |
| 1022,"0" |
| 1023,"1" |
| 1024,"0" |
| 1025,"0" |
| 1026,"0" |
| 1027,"0" |
| 1028,"1" |
| 1029,"0" |
| 1030,"0" |
| 1031,"0" |
| 1032,"0" |
| 1033,"1" |
| 1034,"1" |
| 1035,"0" |
| 1036,"1" |
| 1037,"0" |
| 1038,"0" |
| 1039,"0" |
| 1040,"0" |
| 1041,"0" |
| 1042,"1" |
| 1043,"0" |
| 1044,"0" |
| 1045,"0" |
| 1046,"0" |
| 1047,"0" |
| 1048,"1" |
| 1049,"1" |
| 1050,"0" |
| 1051,"0" |
| 1052,"1" |
| 1053,"1" |
| 1054,"1" |
| 1055,"0" |
| 1056,"0" |
| 1057,"1" |
| 1058,"0" |
| 1059,"0" |
| 1060,"1" |
| 1061,"0" |
| 1062,"0" |
| 1063,"0" |
| 1064,"0" |
| 1065,"1" |
| 1066,"0" |
| 1067,"1" |
| 1068,"1" |
| 1069,"1" |
| 1070,"1" |
| 1071,"1" |
| 1072,"0" |
| 1073,"1" |
| 1074,"1" |
| 1075,"0" |
| 1076,"1" |
| 1077,"0" |
| 1078,"1" |
| 1079,"0" |
| 1080,"0" |
| 1081,"0" |
| 1082,"0" |
| 1083,"0" |
| 1084,"0" |
| 1085,"0" |
| 1086,"0" |
| 1087,"0" |
| 1088,"1" |
| 1089,"0" |
| 1090,"0" |
| 1091,"0" |
| 1092,"1" |
| 1093,"0" |
| 1094,"1" |
| 1095,"1" |
| 1096,"0" |
| 1097,"0" |
| 1098,"0" |
| 1099,"0" |
| 1100,"1" |
| 1101,"0" |
| 1102,"0" |
| 1103,"0" |
| 1104,"0" |
| 1105,"1" |
| 1106,"0" |
| 1107,"1" |
| 1108,"1" |
| 1109,"1" |
| 1110,"1" |
| 1111,"0" |
| 1112,"1" |
| 1113,"0" |
| 1114,"1" |
| 1115,"0" |
| 1116,"1" |
| 1117,"1" |
| 1118,"0" |
| 1119,"1" |
| 1120,"0" |
| 1121,"0" |
| 1122,"1" |
| 1123,"1" |
| 1124,"0" |
| 1125,"0" |
| 1126,"1" |
| 1127,"0" |
| 1128,"0" |
| 1129,"0" |
| 1130,"1" |
| 1131,"1" |
| 1132,"1" |
| 1133,"1" |
| 1134,"1" |
| 1135,"0" |
| 1136,"0" |
| 1137,"1" |
| 1138,"1" |
| 1139,"0" |
| 1140,"0" |
| 1141,"0" |
| 1142,"1" |
| 1143,"0" |
| 1144,"1" |
| 1145,"0" |
| 1146,"0" |
| 1147,"0" |
| 1148,"0" |
| 1149,"0" |
| 1150,"0" |
| 1151,"0" |
| 1152,"0" |
| 1153,"0" |
| 1154,"1" |
| 1155,"0" |
| 1156,"0" |
| 1157,"0" |
| 1158,"1" |
| 1159,"0" |
| 1160,"0" |
| 1161,"0" |
| 1162,"1" |
| 1163,"0" |
| 1164,"1" |
| 1165,"0" |
| 1166,"0" |
| 1167,"1" |
| 1168,"0" |
| 1169,"0" |
| 1170,"0" |
| 1171,"0" |
| 1172,"0" |
| 1173,"0" |
| 1174,"0" |
| 1175,"1" |
| 1176,"1" |
| 1177,"0" |
| 1178,"0" |
| 1179,"1" |
| 1180,"0" |
| 1181,"0" |
| 1182,"0" |
| 1183,"1" |
| 1184,"0" |
| 1185,"1" |
| 1186,"0" |
| 1187,"0" |
| 1188,"1" |
| 1189,"0" |
| 1190,"0" |
| 1191,"0" |
| 1192,"0" |
| 1193,"1" |
| 1194,"0" |
| 1195,"0" |
| 1196,"1" |
| 1197,"1" |
| 1198,"1" |
| 1199,"0" |
| 1200,"1" |
| 1201,"0" |
| 1202,"0" |
| 1203,"0" |
| 1204,"0" |
| 1205,"1" |
| 1206,"1" |
| 1207,"0" |
| 1208,"1" |
| 1209,"0" |
| 1210,"0" |
| 1211,"0" |
| 1212,"0" |
| 1213,"0" |
| 1214,"0" |
| 1215,"0" |
| 1216,"1" |
| 1217,"0" |
| 1218,"1" |
| 1219,"1" |
| 1220,"0" |
| 1221,"0" |
| 1222,"1" |
| 1223,"1" |
| 1224,"0" |
| 1225,"1" |
| 1226,"0" |
| 1227,"1" |
| 1228,"0" |
| 1229,"0" |
| 1230,"0" |
| 1231,"1" |
| 1232,"0" |
| 1233,"0" |
| 1234,"0" |
| 1235,"1" |
| 1236,"0" |
| 1237,"0" |
| 1238,"0" |
| 1239,"1" |
| 1240,"0" |
| 1241,"0" |
| 1242,"1" |
| 1243,"0" |
| 1244,"0" |
| 1245,"1" |
| 1246,"1" |
| 1247,"0" |
| 1248,"1" |
| 1249,"0" |
| 1250,"0" |
| 1251,"0" |
| 1252,"0" |
| 1253,"1" |
| 1254,"0" |
| 1255,"0" |
| 1256,"1" |
| 1257,"0" |
| 1258,"0" |
| 1259,"0" |
| 1260,"1" |
| 1261,"0" |
| 1262,"0" |
| 1263,"1" |
| 1264,"0" |
| 1265,"0" |
| 1266,"1" |
| 1267,"1" |
| 1268,"0" |
| 1269,"0" |
| 1270,"1" |
| 1271,"0" |
| 1272,"0" |
| 1273,"0" |
| 1274,"0" |
| 1275,"0" |
| 1276,"0" |
| 1277,"1" |
| 1278,"0" |
| 1279,"0" |
| 1280,"0" |
| 1281,"0" |
| 1282,"1" |
| 1283,"1" |
| 1284,"0" |
| 1285,"0" |
| 1286,"0" |
| 1287,"1" |
| 1288,"0" |
| 1289,"1" |
| 1290,"0" |
| 1291,"0" |
| 1292,"1" |
| 1293,"0" |
| 1294,"1" |
| 1295,"0" |
| 1296,"1" |
| 1297,"1" |
| 1298,"0" |
| 1299,"1" |
| 1300,"0" |
| 1301,"1" |
| 1302,"1" |
| 1303,"1" |
| 1304,"1" |
| 1305,"0" |
| 1306,"1" |
| 1307,"0" |
| 1308,"0" |
| 1309,"0" |