1. Министерство науки и высшего образования Российской Федерации
2. Санкт-Петербургский Политехнический Университет Петра Великого
3. —

Институт компьютерных наук и технологий

Высшая школа искусственного интеллекта

Отчёт по лабораторной работе «Наивный байесовский классификатор»

по дисциплине «Машинное обучение»

Выполнила

студентка гр. 3540201/20302 Обидина А.И.

Проверил

профессор Уткин Л.В.

Санкт-Петербург

2022

**Задание**

1. Исследовать, 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. Сгенерировать 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

# Ход работы

## **Пункт 1**

1. Для того, чтобы выявить, как объем обучающей выборки и количество тестовых данных из набора «Крестики-нолики» влияют на точность классификации или на вероятность ошибочной классификации, был проведён циклический эксперимент. Объем обучающей выборки увеличивался с 0,1 части выборки до 0,9 части выборки с шагом 0,05 выборки. Для каждой итерации метод naiveBayes был повторен 50 раз с проверкой результатов методом predict. Точность классификации вычисляется как отношение верных предсказаний к произведению разности общего количества данных и отобранных данных на текущей итерации, а также 50. На рисунке 1 приведен полученный результат.

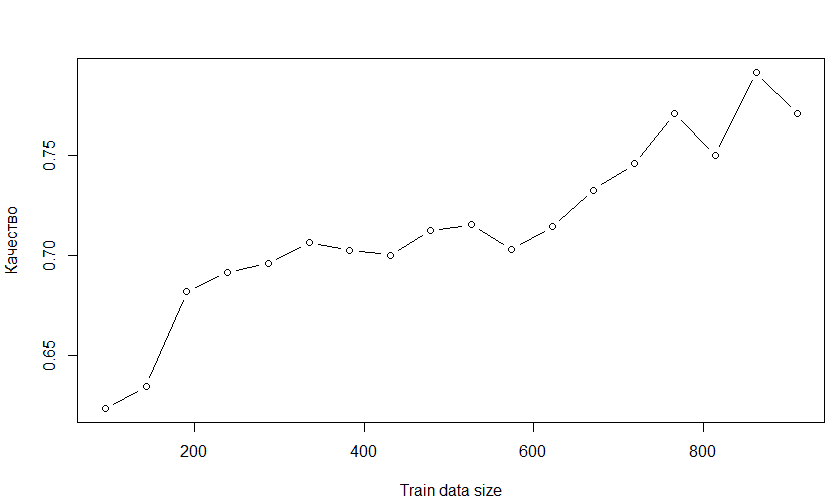


Рисунок 1 – Зависимость качества модели от размера обучающей выборки

Из полученного графика видно, что точность почти прямо пропорционально зависит от размера обучающей выборки. Наилучшая точность в данном эксперименте была достигнута при объёме обучающей выборки 862 (в этом случае точность равна 0,77). Самые лучшие показатели точности – в интервале от 0,7 до 0,9 части выборки.

1. Для того, чтобы выявить, как объем обучающей выборки и количество тестовых данных из набора «Спам» влияют на точность классификации или на вероятность ошибочной классификации, был проведён циклический эксперимент. Объем тестовой выборки увеличивался со 100 записей базы данных до 1000 с шагом в 100 записей. Для каждой итерации метод naiveBayes использован 50 раз с проверкой результатов методом predict. Точность классификации вычисляется как отношение верных предсказаний к произведению разности общего количества данных и отобранных данных, а также 50. На рисунке 2 приведен полученный результат.

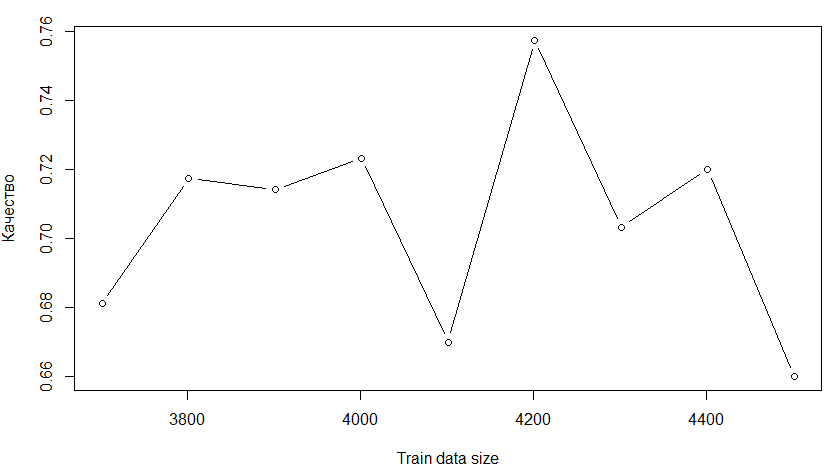


Рисунок 2 – Зависимость качества модели от размера обучающей выборки

По графику нельзя выявить точную зависимость качества классификации от объёма входных данных. При этом лучшая точность наблюдается при использовании 4201 записи в качестве обучающей выборки (точность равна 0,758).

**Код:**

install.packages("e1071")

library(e1071)

first\_data <- read.table("D:\\Tic\_tac\_toe.txt", sep = ",", stringsAsFactors = TRUE)

first\_data

strings\_num<-dim(first\_data)[1]

rand\_data<-first\_data[order(runif(strings\_num)), ]

train<-c()

prediction<-c()

acc<-c()

start<-0.1

finish<-1

for\_one\_volume<-function(incr, idx, fl){

pr<-0

tr<-0

for(i in seq(1, repeats\_num)){

if (fl == 0) {

test\_data<-rand\_data[(incr+1):strings\_num, ]

train\_data<-rand\_data[1:incr, ]

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

A\_predicted<-predict(A\_classifier, test\_data)

res <- table(A\_predicted, test\_data$V10)

pr <- pr + res[1] + res[4]

}

else{

test\_data <- spam[idx,]

train\_data <- spam[-idx,]

model <- naiveBayes(type ~ ., data = train\_data)

predict(model, test\_data)

res <- table(predict(model, test\_data), test\_data$type)

predict(model, test\_data, type = "raw")

tr <- dim(train\_data)[1]

tr

pr <- pr + res[1] + res[4]

pr

}

}

return(list(tr,pr))

}

#tic-tac-toe

repeats\_num<-50

s<-start

k<-1

while (s<finish) {

incr<-as.integer(strings\_num\*s)

train[k]<-incr

tr\_pr = as.numeric(for\_one\_volume(incr,0,0))

prediction[k]<-tr\_pr[2]

acc[k] <- prediction[k] / (repeats\_num\*(strings\_num-incr))

s<-s+0.05

k<-k+1

}

train

acc

plot(train, acc, type = "b", xlab = "Train data size", ylab = "Качество")

#spam

library(kernlab)

library(e1071)

data(spam)

train<-c()

prediction<-c()

acc<-c()

spam[0:1,]

start<-100

finish<-1000

repeats\_num<-50

k<-1

s<-start

while (s<finish) {

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

test\_data <- spam[idx,]

tr\_pr<-as.numeric(for\_one\_volume(0,idx,1))

train[k]<-tr\_pr[1]

prediction[k]<-tr\_pr[2]

acc[k] <- prediction[k] / (repeats\_num\*dim(test\_data)[1])

s<-s+100

k<-k+1

}

plot(train, acc, type = "b", xlab = "Train data size", ylab = "Качество")

## **Пункт 2**

Были сгенерированы 100 точек с двумя признаками X1 (зеленого цвета) И X2 (синего цвета) в соответствии с нормальным распределением.

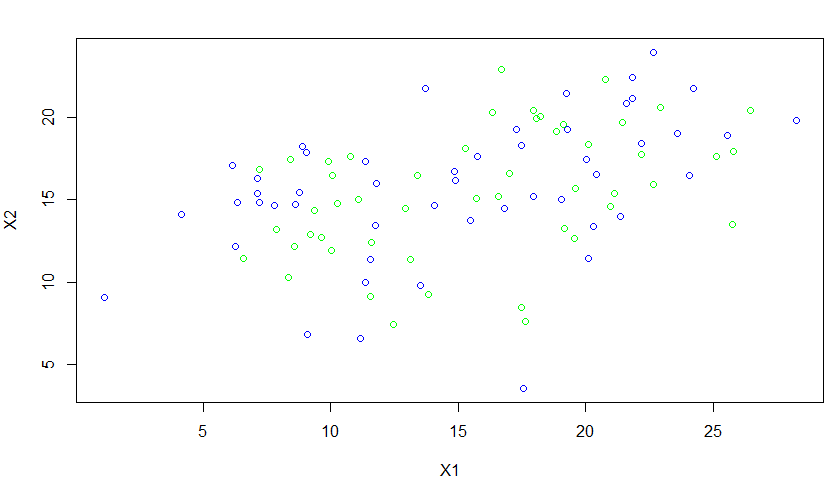


Рисунок 3 – Сгенерированные данные

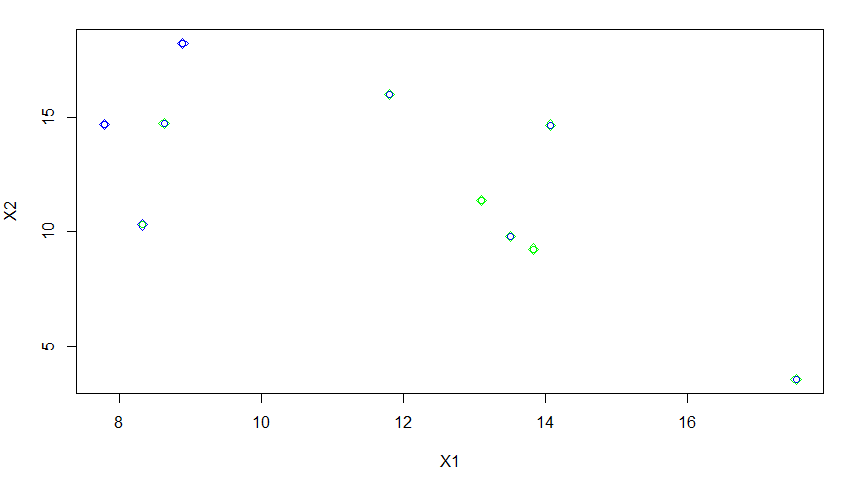


Рисунок 4 – Верно и неверно классифицированные данные



При проведении экспериментов точность классификации равна 0,95.

**Код:**

library(e1071)

n<-100

train\_part<-0.8

m\_x11=10

m\_x12=20

m\_x21=14

m\_x22=18

s1=4

s2=3

X1<-c(rnorm(n/2,m\_x11,s1),rnorm(n/2,m\_x12,s2))

X2<-c(rnorm(n/2,m\_x21,s1),rnorm(n/2,m\_x22,s2))

type <- c(rep(-1, n/2), rep(1, n/2))

X1\_X2\_type <- data.frame(X1, X2, type) #объединение

X1\_X2\_type\_rand <- X1\_X2\_type[order(runif(n)), ] #перемешивание

nt <- as.integer(n\*train\_part)

train\_data <- X1\_X2\_type\_rand[1:nt, ]

test\_data <- X1\_X2\_type\_rand[(nt+1):n, ]

A\_classifier <- naiveBayes(type ~ ., data=train\_data)

A\_predicted <- predict(A\_classifier, test\_data)

result <- table(A\_predicted, test\_data$type)

acc <- (result[1]+result[4])/(n-nt)

plot(X1\_X2\_type$X1,X1\_X2\_type$X2,

col=ifelse(X1\_X2\_type$type==-1,"green","blue"),

xlab="X1",ylab="X2")

plot(test\_data$X1,test\_data$X2,

col=ifelse(test\_data$type==-1,"green","blue"),

xlab="X1",ylab="X2")

points(test\_data$X1,test\_data$X2,

pch=5,col=ifelse(A\_predicted==-1,"green","blue"),

xlab="X1",ylab="X2")

result

acc

## **Пункт 3**

При расчете точности байесовского классификатора для данных Titanic (параметры - класс, пол, возраст, число братьев-сестер/муж-жена на борту, число родителей/детей на борту, номер, стоимость билета, каюта, порт посадки; предварительная обработка данных не была осуществлена) были получены следующие результаты:



Точность классификатора: 75,36%.

**Код:**

A\_test = read.csv(file = 'D:\\All\_Labs\\test.csv', sep=',', header=TRUE, na.strings='NA', stringsAsFactors=F)

A\_train = read.csv(file = 'D:\\All\_Labs\\train.csv', sep=',', header=TRUE, na.strings='NA', stringsAsFactors=F)

A\_train[0,]

A\_train

A\_classifier <- naiveBayes(A\_train[,-12], as.factor(A\_train$Survived))

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

result <- table(A\_predicted, A\_test$Survived)

acc = (result[1] + result[4])/dim(A\_test)[1]

result

acc

## **Вывод**

При выполнении данной лабораторной работы были разработаны байесовские классификаторы для нескольких наборов данных и оценены их точности классификации, было проанализировано влияние объема выборки на точность классификации (в общем случае не влияет, а в некоторых даже ухудшает точность). Так же был рассмотрен случай нормально распределенных данных – в этом случае классификатор достаточно точно провел классификацию.