1. **Название и номер задания:**

Задание НБК - Решение задачи классификации с помощью наивного Байесовского классификатора.

1. **Постановка задачи:**

1. Выберите данные для задачи классификации (можно использовать тот же набор данных, к которому Вы применяли метод KNN).

2. Решите задачу с помощью наивного Байесовского классификатора; если число признаков = 2, то визуализуйте данные).

3. Постройте кросс-валидационную таблицу, сделайте вывод о точности решения задачи классификации.

4. Задайте нескольких новых данных, покажите соответствующие точки на графике (выделите их другим цветом).

5. Определите класс для новых данных..

1. **Студент:**

Санамян Артак

1. **Номер группы:**

09-715

1. **Ссылка на источник данных:**

<https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/>

1. **Постановка задачи**

Задача классификации — формализованная задача, в которой имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется выборкой. Классовая принадлежность остальных объектов неизвестна.

Для демонстрации маркетинговой модели RFMTC (модифицированная версия RFM) в исследовании была использована донорская база данных центра обслуживания переливания крови в городе Синь-Чу на Тайване. Центр передает автобус службы переливания крови в один университет в городе Синь-Чу, чтобы собирать кровь, которую сдают примерно каждые три месяца. Для построения модели FRMTC произвольно выбрали 748 доноров из базы данных доноров. Эти данные 748 доноров, каждый из которых включает Р – сколько прошло месяцев с момента последней сдачи крови, Ф – общее количество сдачи крови (раз), М – общее количество сданной крови, T – сколько месяцев прошло с момента первой сдачи крови и переменная отклика – сдача в определенный месяц (март 2017), 1 – кровь сдавалась, а 0 – означает, что кровь не сдавалась.

По исходному набору необходимо реализовать метод НБК.

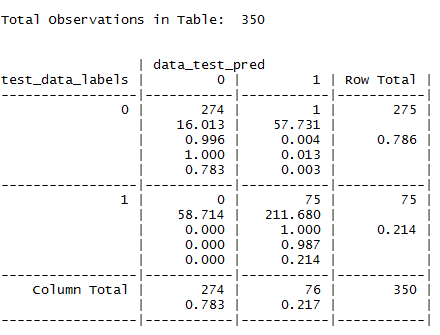
1. **Описание выполнения работы**

Разделим исходную выборку на обучающую и тестовую (необходимость этого заключается в том, что по построенной модели будет необходимо оценить истинность предсказанных значений).

Так как число атрибутов выборки больше двух, не представляется возможным привести графическую интерпретацию.

При помощи обучающей выборки, построим модель метода НБК.

Рассмотрим кросс-валидационную таблицу:



Классификатор ошибся один раз.

Сравним полученные целевые значения для тестовой выборки и истинных переменных отклика. Добавим десять вырезанных значений из тестовой выборки и сравним с предсказанными значениями:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

Истинные значения переменной отклика совпадают со значениями, вычисленными при помощи метода НКБ.

1. **Приложение**

#очищаем рабочее пространство

rm (list=ls())

#считываем данные

A=scan("file.txt")

Data\_null=matrix(A,ncol=5,byrow=TRUE)

dims = dim(Data\_null)

#задаем количество столбцов и строк

n = dims[1]

m = dims[2]

#размерность выборок

n\_data = 350

n\_train = 398

#перемешиваем

set.seed(1983)

Data=Data\_null[order(runif(748)),]

train\_data = Data[1:n\_train,]

train\_data\_labels = train\_data[,5]

train\_data\_labels<-factor(train\_data\_labels)

test\_data = Data[(n\_train+1):n,]

#формируем тестовые данные

test\_data\_labels = test\_data[,5]

#подключаем библиотеку

library(e1071)

#строим модель

my\_classifier <- naiveBayes(train\_data,train\_data\_labels)

data\_test\_pred <- predict(my\_classifier, test\_data)

#подключаем библиотеку

library("gmodels")

#строим таблицу

CrossTable(x = test\_data\_labels, y = data\_test\_pred, prop.chisq=TRUE)