1. **Название и номер задания:**

KNN1: Решение задачи классификации методом kNN

1. **Постановка задачи:**

1. Найдите реальные данные для задачи классификации. Зафиксируйте свой выбор в приведённом ниже объекте "Wiki".

2. Перечислите факторы (и укажите единицы их измерения) и поясните смысл переменной отклика.

3. Если число факторов m <= 2, то визуализируйте данные (постройте облако точек).

4. Реализуйте метод kNN для найденных данных.

5. Оцените точность полученного решения с помощью метода кросс валидации. (Скриншот кросс-валидационнной таблицы. Обоснование репрезентативности обучающей выборки.)

6. Задайте нескольких новых данных. Если m <= 2, то покажите соответствующие точки на графике (выделите их другим цветом).

7. Определите значение переменной отклика (номер класса) 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 kNN.

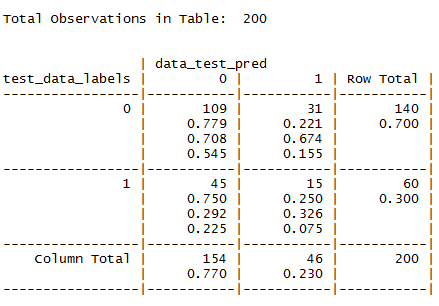
1. **Описание выполнения работы**

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 kNN.

Рассмотрим кросс-валидационную таблицу:



Классификатор ошибся в 76 случаях из 200, что составляет 38%. В реальных задачах метод kNN не приемлем для данного набора.

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 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

Истинные значения переменной отклика не совпадают со значениями, вычисленными при помощи метода НКБ в двух из десяти случаев.

1. **Приложение**

#отчищаем рабочее пространство

rm (list=ls())

#считываем данные

A=scan("file.txt")

#преобразовываем в матрицу

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

#объем тестовой выборки

n\_data = 300

#объем выборки

n\_train = 100

#перемешиваем выборку

set.seed(1983)

#выбираем данные

Data=Data\_null[order(runif(n\_data)),]

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

#вырезаем результат

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

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

#вырезаем результат

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

#подключаем библиотеку

library("class")

#применяем метод

data\_test\_pred <- knn(train = train\_data,

test = test\_data,

cl = train\_data\_labels,

k = round(sqrt(n\_train)))

library("gmodels")

#строим кросс таблицу

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