**Задание 8**

*Построение и применение однофакторной*

*логистической регрессионной модели*

Найдите реальные данные для задачи однофакторной бинарной классификации и выполните для них следующее:

1. Найдите коэффициенты логистической регрессионной модели.
2. Постройте логистическую кривую и облако точек.
3. Выполните прогноз переменной отклика для нескольких новых объектов и нескольких объектов обучающей выборки.
4. Проиллюстрируйте на графике результаты прогнозирования.
5. Сделайте выводы о качестве логистической регрессионной модели.
6. Результаты оформите в виде Word-файла и прикрепите его здесь (Весь R-код поместите в конец файла как приложение).

Были использованы данные исследования 3,168 образцов голоса мужчин и женщин. Сразу же скажем, что в качестве переменной отклика мы выберем пол человека.

Источник данных: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.

Данные содержат следующие характеристики:

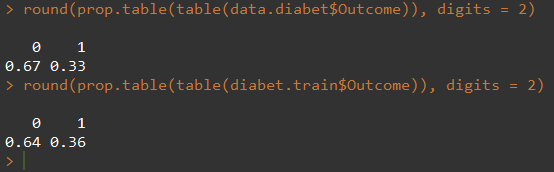
1. meanfreq — средняя частота звука, кГЦ;
2. sd — стандартное отклонение частоты;
3. median — медиана частоты, кГЦ;
4. Q25 — 25% квартиль, кГЦ;
5. Q75 — 75% квартиль, кГЦ;
6. IQR — интерквартильный размах, кГЦ;
7. skew — коэффициент ассиметрии;
8. kurt — коэффициент эксцесса;
9. sp.ent — спектральная энтропия;
10. sfm — спектральная плоскостность энтропия Винера;
11. mode — мода частоты;
12. centroid — центроид частоты;
13. meanfun — среднее основной частоты;
14. minfun — минимум основной частоты; measured across acoustic signal
15. maxfun — максимум основной частоты; measured across acoustic signal
16. meandom — average of dominant frequency measured across acoustic signal
17. mindom — minimum of dominant frequency measured across acoustic signal
18. maxdom — maximum of dominant frequency measured across acoustic signal
19. dfrange — range of dominant frequency measured across acoustic signal
20. modindx — modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
21. label — male or female

В качестве факторов будем использовать характеристики 1-8, характеристика 9 — переменная отклика.

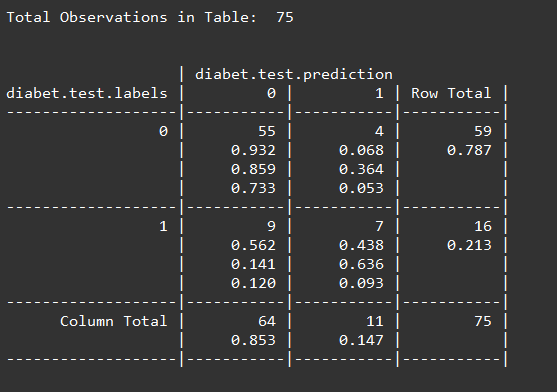
После изучения существующей информации о данных, оказалось, что присутствуют некорректные наблюдения. К примеру, уровень кровяного давления составляет или же индекс массы тела равен . После удаления записей с некорректными значениями, объем выборки составил **392** наблюдения.

Исходная выборка была разделена обучающую и тестовую, объемами **317** и **75** наблюдений соответственно. Для сохранения репрезентативности обучающей выборки с помощью функции sample () был сгенерирован вектор индексов длиной в 317 (объем обучающей выборки) элементов, состоящий из целых чисел от 1 до 392 (объем исходной выборки).

Соотношения классов в исходной и обучающей выборках почти одинаковы:



Приведем кросс-валидационную таблицу для анализа эффективности метода **kNN**:



Заметим, что метод ошибочно **определил** наличие диабета у **4** обследуемых и ошибочно **не определил** наличие диабета у **9** обследуемых. от общего числа прогнозов — ошибочны, следовательно, метод точен на .

Результаты прогноза для пользовательских значение приведены [здесь](#_Результаты_прогноза_пользовательски).

Скрипт программы на R:

data.diabet <- read.csv("data/diabetes.csv",

header = TRUE,

sep = ",")

na.fail(data.diabet)

# отберем только те наблюдения, которые не содержат логических ошибок

data.diabet <- data.diabet[data.diabet$Glucose > 0 &

data.diabet$BloodPressure > 0 &

data.diabet$SkinThickness > 0 &

data.diabet$Insulin > 0 &

data.diabet$BMI > 0,]

# определим объемы тестовой и обучающей выборок

data.size <- length(data.diabet[, 1])

train.size <- 317

test.size <- data.size - train.size

# определим индексы наблюдений тестовой и обучающей выборок

train.indices <- sample(1:data.size, train.size)

test.indices <- setdiff(1:data.size, train.indices)

diabet.train <- data.diabet[train.indices,]

# проверим соотношения классов в исходной и обучающей выборках

round(prop.table(table(data.diabet$Outcome)), digits = 2)

round(prop.table(table(diabet.train$Outcome)), digits = 2)

# запомним отклики для обучающей и тестовой выборок

diabet.train.labels <- diabet.train[, 9]

diabet.train <- diabet.train[, -9]

diabet.test <- data.diabet[test.indices,]

diabet.test.labels <- diabet.test[, 9]

diabet.test <- diabet.test[, -9]

# предскажем значение отклика для тестовой выборки

library("class")

diabet.test.prediction <- knn(train = diabet.train,

test = diabet.test,

cl = diabet.train.labels,

k = round(sqrt(train.size)))

# проанализируем полученные результаты

library("gmodels")

CrossTable(diabet.test.labels, diabet.test.prediction, prop.chisq = FALSE)

# спрогнозируем нзначение переменной-отклика для произвольных значений

my.diabet.test <- **matrix**(nrow = 2, ncol = 8)

my.diabet.test[1,] <- c(2, 76, 70, 23, 67, 34.7, 0.23, 22)

my.diabet.test[2,] <- c(0, 90, 88, 26, 80, 31.4, 0.63, 25)

my.diabet.test <- as.data.frame(my.diabet.test)

colnames(my.diabet.test) <- colnames(diabet.train)

my.diabet.test.prediction <- knn(train = diabet.train,

test = my.diabet.test,

cl = diabet.train.labels,

k = round(sqrt(train.size)))

my.diabet.test.prediction

Приложение

# Результаты прогноза пользовательских значений

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
| 2 | 76 | 70 | 23 | 67 | 34.7 | 0.23 | 22 | 0 |
| 0 | 90 | 88 | 26 | 80 | 31.4 | 0.63 | 25 | 0 |