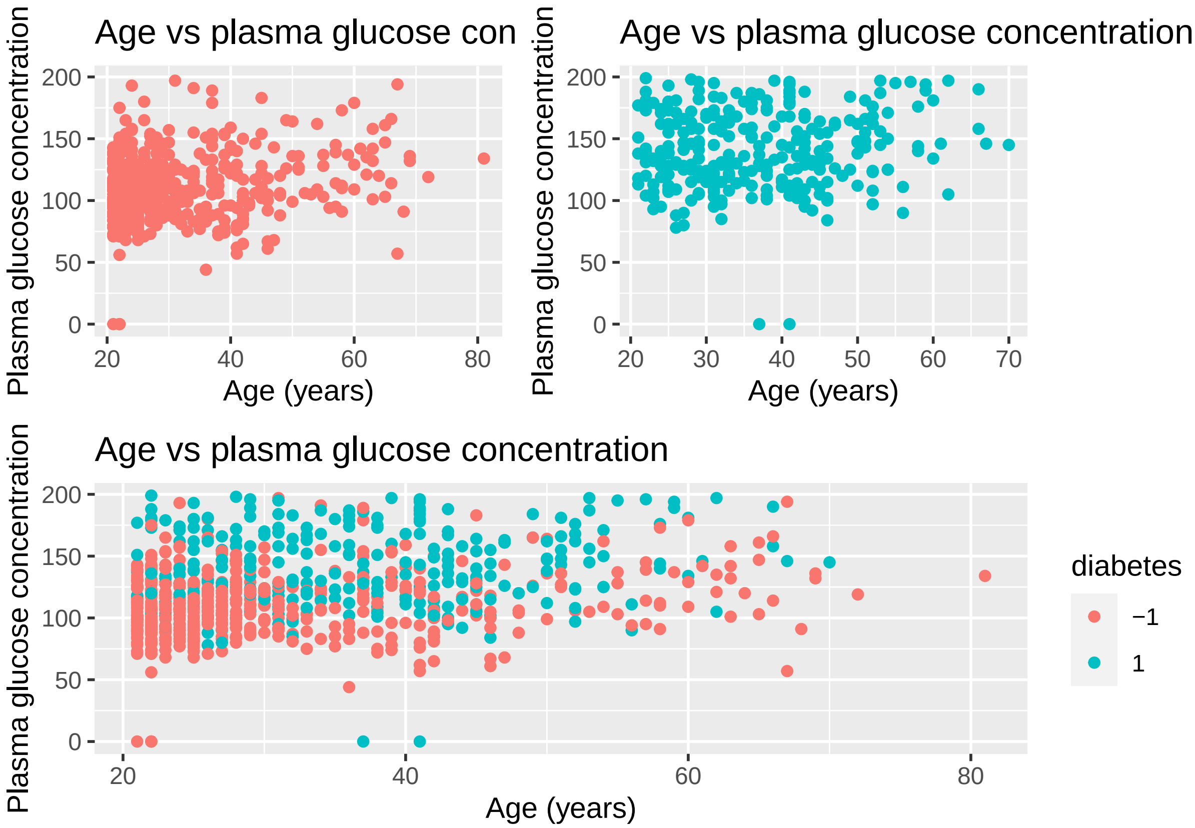
Lab01 report

2022-11-08

Assignment 3: logistic regression

3.1: Plot

Plot of age vs glucose plasma concentration:



3.2: Logistic regression model with diabetes as target and glucose plasma concentration and age as features

Probabilistic model:

$$g(\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}^T \boldsymbol{x}}} = \frac{e^{(-5.912449 + 0.03564404 x_1 + 0.02477835 x_2)}}{1 + e^{(-5.912449 + 0.03564404 x_1 + 0.02477835 x_2)}}$$

Training misclassification error (percent):

```
#> [1] 26.3
```

Appendix: All code for this report

```
knitr::opts_chunk$set(comment = "#>",
                       echo = FALSE)
#------------------------------------------------------------------------------#
#- Set libraries
#------------------------------------------------------------------------------#
library(tidyverse)
library(ggplot2)
library(patchwork)
library(scales)
library(tinytex)
#------------------------------------------------------------------------------#
#- 0) Read in data and divide into training, validation and test
#------------------------------------------------------------------------------#

file_in <- "C:/Users/kerstin/Documents/LiU/ML/Labs/Lab01/Data/pima-indians-diabetes.csv"
data_in <- read_csv(file_in, col_names = FALSE)

spec(data_in)

names(data_in) <- c(
  "pregnant_num",
  "glucose_pl",
  "bp_dia",
  "triceps_skin_thick",
  "insulin_serum",
  "bmi",
  "diabetes_pedigree",
  "age",
  "diabetes")

#- Set factors
data_1 <- mutate(data_in,diabetes=ifelse(diabetes<1,-1,1),
                 diabetes=as.factor(diabetes))



data <- data_1

ylim_min <- min(data$glucose_pl)
ylim_max <- max(data$glucose_pl)

hex <- hue_pal()(2)

p1 <- data %>% ggplot() + aes(x=age,y=glucose_pl, colour=diabetes)+ geom_point() +
  labs(title="Age vs plasma glucose concentration",
       y="Plasma glucose concentration", x="Age (years)") +
  ylim(ylim_min, ylim_max)

p2 <- data %>% filter(diabetes==-1) %>% ggplot() + aes(x=age,y=glucose_pl, colour=diabetes)+
  geom_point(colour= "#F8766D") +
  labs(title="Age vs plasma glucose concentration",
       y="Plasma glucose concentration", x="Age (years)") +
  ylim(ylim_min, ylim_max)

p3 <- data %>% filter(diabetes==1) %>% ggplot() + aes(x=age,y=glucose_pl, colour=diabetes)+
  geom_point(colour= "#00BFC4") +
  labs(title="Age vs plasma glucose concentration",
       y="Plasma glucose concentration", x="Age (years)") +
  ylim(ylim_min, ylim_max)

(p2 + p3)/p1

train=data%>%select(glucose_pl,age,diabetes)

m1=glm(diabetes~., train, family = "binomial")
prob=predict(m1, type="response")
pred=ifelse(prob>0.5, 1, -1)


cm_train <- table(train$diabetes, pred)

(err_train_perc <- round(100*((dim(train)[1]-sum(diag(cm_train)))/dim(train)[1]),2))
```