Sim_notebook

Simon Jorstedt

2023-11-16

Question 3

We are provided data covering the onset of diabetes within a five year period for a group of individuals. The data consists of nine variables including a binary response variable indicating the presence of diabetes or not. In Figure 3.1, we plot the Plasma Glucose Concentration (PGC) against age, and color datapoints by diagnosis.

We do not think it will be easy to make the distinction using only these two variables. In Figure 3.1 we observe a large cluster of young people (ages ~20-25) that do not have diabetes, along with a significant number of outliers (among the non-diabetes people.) The people with diabetes however are much more spread out, with no smaller clusters. It appears as though people with diabetes tend to have slightly larger Plasma Glucose Concentration (PGC) values than people without diabetes. Thus it does appear as though there is some precedent for using only PGC values and Age as explanatory variables. It won't be easy or "clear-cut" though.

There is clearly some explanatory power within these two variables, but it is likely not enough to achieve a highly accurate logistic regression (classification) predictor.

```
## Warning: package 'webshot2' was built under R version 4.3.2
```

See *Machine Learning - A first course for engineers and scientists* (pp. 45-52) for a discussion on logistic regression.

3.2

We will now fit a logistic regression (classification) model using PGC and age to predict the presence of diabetes. We will initially use a classification threshold of $r = 0.5$. The model printout is

```
##
## Call:
## glm(formula = diabetes ~ pg_con + age, family = binomial(link = "logit"),
##     data = diab_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.897858   0.462450  -12.75  < 2e-16 ***
## pg_con       0.035582   0.003288   10.82  < 2e-16 ***
## age          0.024502   0.007379    3.32 0.000899 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 796.49  on 764  degrees of freedom
## AIC: 802.49
##
## Number of Fisher Scoring iterations: 4

## Misclassification error: 0.2659713
```

Mathematically, our model predictor $g(\boldsymbol{x})$ can be represented as

$$g(\boldsymbol{x}) = \frac{e^z}{1+e^z}$$

where

$$z = \boldsymbol{\theta}^T \boldsymbol{x} = (-5.8979, 0.0356, 0.0245) \cdot (1, x_{\text{pg}}, x_{\text{age}})^T$$

In Figure 3.2, we see the same data as in Figure 3.1, but datapoints are now colored by the predicted diagnosis using the classification threshold $r = 0.5$. When comparing Figure 3.1 and 3.2 we see that the quality of the classification is decent at best. The resulting predictions is visually precisely what could be expected when analysing Figure 3.1. The model essentially just takes what was stated before, about diabetic people being overrepresented among people with high PGC values. This means that there will be a high true/false pos/neg probability.

3.3

The decision boundary for the first model will be the set of points that satisfy the following equation:

$$g(\boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}^T \boldsymbol{x}}}{1+e^{\boldsymbol{\theta}^T \boldsymbol{x}}} = \frac{1}{2} \implies \boldsymbol{\theta}^T \boldsymbol{x} = 0 \implies \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 = 0 \implies x_2 = \frac{-\theta_0}{\theta_2} - \frac{\theta_1}{\theta_2} x_1.$$

With our trained parameters, this comes out to the line

$$y = ...$$

More complicated boundary. When it comes to the expanded model, the decision boundary can be found as the solution to the same kind of equation as the simpler model, but the equation is now more complicated. It can however be solved numerically using the R `polyroot` function. The equation is

$$\theta_0 + \theta_1 x_1 + \theta_2 \cdot x_2 + \theta_3 \cdot x_1^4 + \theta_4 \cdot x_1^3 x_2 + \theta_5 \cdot x_1^2 x_2^2 + \theta_6 \cdot x_1 x_2^3 + \theta_7 \cdot x_2^4 = 0$$

3.4

Comment on Figure 3.3.

3.5

```
# Create new variables z_1, ... z_5
diab_data <- diab_data %>%
  mutate(z1 = pg_con^4 * age^0,
         z2 = pg_con^3 * age^1,
         z3 = pg_con^2 * age^2,
         z4 = pg_con^1 * age^3,
         z5 = pg_con^0 * age^4)

# New model
model2 <- glm(formula = diabetes ~ pg_con + age + z1 + z2 + z3 + z4 + z5,
    family = binomial(link = "logit"),
    data = diab_data)

summary(model2)
```

```
##
## Call:
## glm(formula = diabetes ~ pg_con + age + z1 + z2 + z3 + z4 + z5,
##     family = binomial(link = "logit"), data = diab_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.279e+00  1.129e+00  -8.217  < 2e-16 ***
## pg_con       3.772e-02  9.473e-03   3.981 6.85e-05 ***
## age          1.453e-01  2.072e-02   7.014 2.32e-12 ***
## z1           1.266e-08  5.610e-09   2.257  0.02402 *
## z2          -1.760e-07  7.638e-08  -2.304  0.02122 *
## z3           8.424e-07  3.439e-07   2.450  0.01430 *
## z4          -1.682e-06  6.317e-07  -2.662  0.00776 **
## z5           8.045e-07  4.056e-07   1.983  0.04732 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 991.38  on 766  degrees of freedom
## Residual deviance: 741.07  on 759  degrees of freedom
## AIC: 757.07
##
## Number of Fisher Scoring iterations: 5
```

```
diab_data_newmodel <- diab_data %>%
  mutate(predicted_z_05 = as.integer(fitted(model2) >= 0.5))
```

```
fake_data <- data.frame(age=runif(1000, 0, 80), pg_con=runif(1000, 0, 200)) %>%
  mutate(z1 = pg_con^4 * age^0,
         z2 = pg_con^3 * age^1,
         z3 = pg_con^2 * age^2,
         z4 = pg_con^1 * age^3,
         z5 = pg_con^0 * age^4) %>%
```

```
  mutate(predicted = as.integer(predict.glm(model2, newdata = .) %>% logit() >= 0.5))

# FAKE Plot
plot_ly(type="scatter", mode="markers",
        data = fake_data,
        x = ~age,
        y = ~pg_con,
        colors = c("#1f77b4", "#ff7f0e"),
        color = ~factor(x = fake_data$predicted, labels = c("No diabetes", "Diabetes"))) %>%
  layout(title = "Fig 3.? PGC vs Age, colored by predicted diagnosis (r=0.5)") %>%

  # Add classifaction trace
  add_trace(inherit = F, type="scatter", mode="lines",
            data = class_bound,
            x = ~age,
            y = ~pgc,
            split = ~group)

# Add trace!
```