Task 1

Task 1: Build a logistic model to classify the images into malignant/benign, and write down your likelihood function, its gradient and Hessian matrix.

The variable "Diagnosis" is a binary response variable indicating if the image is coming from cancer tissue or benign cases (M = malignant, B = benign). In the following logistic regression model, the "Diagnosis" variable will be coded as 1 for malignant cases and 0 for benign cases.

Given n i.i.d. observations with p predictors, we consider a logistic regression model

$$P(Y_i = 1 \mid \mathbf{x}_i) = \frac{e^{\mathbf{x}_i^\top \beta}}{1 + e^{\mathbf{x}_i^\top \beta}},\ i = 1, \dots, n \tag{1}$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$ is the parameter vector, $\mathbf{x}_i = (1, X_{i1}, \dots, X_{ip})^\top$ is the vector of predictors in the i-th observation, and $Y_i \in \{0, 1\}$ is the binary response in the i-th observation. Let $\mathbf{y} = (Y_1, Y_2, \dots, Y_n)^\top$ denote the response vector, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times (p+1)}$ denote the design matrix. The observed likelihood of $\{(Y_1, \mathbf{x}_1), (Y_2, \mathbf{x}_2) \dots, (Y_n, \mathbf{x}_n)\}$ is

$$L(\beta; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \left[\left(\frac{e^{\mathbf{x}_i^\top \beta}}{1 + e^{\mathbf{x}_i^\top \beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{\mathbf{x}_i^\top \beta}} \right)^{1 - Y_i} \right].$$

Maximizing the likelihood is equivalent to maximizing the log-likelihood function:

$$f(\beta; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \left[Y_i \mathbf{x}_i^\top \beta - \log\left(1 + e^{\mathbf{x}_i^\top \beta}\right) \right]. \tag{2}$$

The estimates of model parameters are

$$\hat{\beta} = \arg\max_{\beta}\ f(\beta; \mathbf{y}, \mathbf{X}),$$

and the optimization problem is

$$\max_{\beta}\ f(\beta; \mathbf{y}, \mathbf{X}). \tag{3}$$

Denote $p_i = P(Y_i = 1 \mid \mathbf{x}_i)$ as given in (1) and $\mathbf{p} = (p_1, p_2, \dots, p_n)^\top$. The gradient of f is

$$\begin{aligned} \nabla f(\beta; \mathbf{y}, \mathbf{X}) &= \mathbf{X}^\top (\mathbf{y} - \mathbf{p}) \\ &= \sum_{i=1}^{n} (Y_i - p_i) \mathbf{x}_i \\ &= \begin{pmatrix} \sum_{i=1}^{n} (Y_i - p_i) \\ \sum_{i=1}^{n} (Y_i - p_i) X_{i1} \\ \vdots \\ \sum_{i=1}^{n} (Y_i - p_i) X_{ip} \end{pmatrix}. \end{aligned}$$

Denote $w_i = p_i(1-p_i) \in (0,1)$ and $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$. The Hessian matrix of f is given by

$$\begin{aligned}
\nabla^2 f(\beta; \mathbf{y}, \mathbf{X}) &= -\mathbf{X}^\top \mathbf{W} \mathbf{X} \\
&= -\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i^\top \\
&= -\begin{pmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i X_{i1} & \cdots & \sum_{i=1}^n w_i X_{i1} \\ \sum_{i=1}^n w_i X_{i1} & \sum_{i=1}^n w_i X_{i1}^2 & \cdots & \sum_{i=1}^n w_i X_{i1} X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n w_i X_{ip} & \sum_{i=1}^n w_i X_{ip} X_{i1} & \cdots & \sum_{i=1}^n w_i X_{ip}^2 \end{pmatrix}.
\end{aligned}$$

Next, we show that the Hessian matrix $\nabla^2 f(\beta; \mathbf{y}, \mathbf{X})$ is a negative-definite matrix if $\mathbf{X}$ has full rank.

Proof. For any $(p+1)$-dimensional nonzero vector α, given that $\mathbf{X}$ has full rank, $\mathbf{X}\alpha$ is also a nonzero vector. Since $\mathbf{W}$ is positive-definite, we have

$$\begin{aligned}
\alpha^\top \nabla^2 f(\beta; \mathbf{y}, \mathbf{X}) \alpha &= \alpha^\top (-\mathbf{X}^\top \mathbf{W} \mathbf{X}) \alpha \\
&= -(\mathbf{X}\alpha)^\top \mathbf{W} (\mathbf{X}\alpha) \\
&< 0.
\end{aligned}$$

Thus, $\nabla^2 f(\beta; \mathbf{y}, \mathbf{X})$ is negative-definite. □

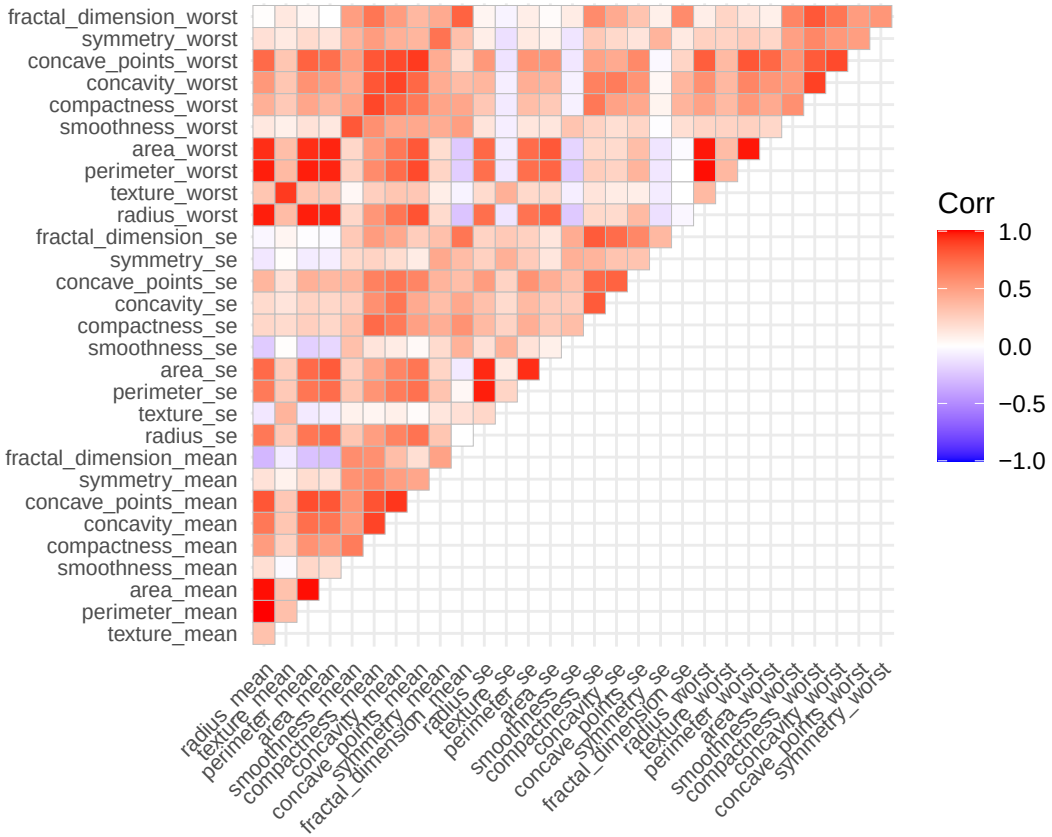
Hence, the optimization problem (3) is a well-defined problem.

Variable selection is automatically conducted by LASSO in task 3 and 4.

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggcorrplot)
cancer <- read.csv("breast-cancer.csv") %>%
  janitor::clean_names() %>%
  select(-1,-33) %>%
  mutate(diagnosis = recode(diagnosis, "M" = 1, "B" = 0))
#ID labels individual breast tissue images;
#The second column 'Diagnonsis' identifies if the image is coming from cancer tissue or benign cases (M
#The other 30 columns correspond to mean, standard deviation and the largest values (points on the tail
corr = cancer[2:31] %>%
  cor()
ggcorrplot(corr, type = "upper", tl.cex = 8)
```


fractal_dimension_worst
symmetry_worst
concave_points_worst
concavity_worst
compactness_worst
smoothness_worst
area_worst
perimeter_worst
texture_worst
radius_worst
fractal_dimension_se
symmetry_se
concave_points_se
concavity_se
compactness_se
smoothness_se
area_se
perimeter_se
texture_se
radius_se
fractal_dimension_mean
symmetry_mean
concave_points_mean
concavity_mean
compactness_mean
smoothness_mean
area_mean
perimeter_mean
texture_mean
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave_points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave_points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave_points_worst
symmetry_worst
Corr
1.0
0.5
0.0
−0.5
−1.0