Quiz 2

Instructions:

- In this quiz you are asked to use the **R** package glmnet to analyze a dataset, and prepare an "R Markdown" script to automatically generate a PDF report as described below.
- This is a take home quiz due Sunday February 4 at 11:59pm.
- Submit your quiz on LEARN as follows:
  1. Submit *only* one file in R Markdown format: **uwname-quiz2.Rmd**, where **uwname** is your UW username.
  2. To upload this file to LEARN, navigate to Assessments/Dropbox, then click on Quiz 2.
- You may work on this quiz in groups. However, you **must include the names of all collaborators** in the Comments field during the LEARN file upload process.
- Organized, well-commented, and efficient code is required for full marks.

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset consists of 30 features of cell nuclei extracted from 569 digitized images of benign and malignant breast tumors. The data is contained in the file **wdbc.csv**, which can be imported into **R** as follows:

```
tumor <- read.csv("wdbc.csv")
dim(tumor)
```

```
[1] 569  32
```

```
colnames(tumor)
```

```
 [1] "id"        "diag"      "radM"      "textM"     "perimM"
 [6] "areaM"     "smoothM"   "compactM"  "concM"     "cptsM"
[11] "symM"      "fracM"     "radSE"     "textSE"    "perimSE"
[16] "areaSE"    "smoothSE"  "compactSE" "concSE"    "cptsSE"
[21] "symSE"     "fracSE"    "radW"      "textW"     "perimW"
[26] "areaW"     "smoothW"   "compactW"  "concW"     "cptsW"
[31] "symW"      "fracW"
```

In addition to the patient ID (variable id) and diagnosis (variable diag; M = malignant, B = benign), the 30 real-valued cell nuclei features are of the form feature{M/SE/W}, where the suffix stands for mean, standard error, and worst along the following ten nuclei characteristics:

- rad: radius (mean of distances from center to points on the perimeter).
- text: texture (standard deviation of gray-scale values).

- `perim`: perimeter.
- `area`: area.
- `smooth`: smoothness (local variation in radius lengths).
- `compact`: compactness (perimeter^2 / area - 1).
- `conc`: concavity (severity of concave portions of the contour).
- `cpts`: concave points (number of concave portions of the contour).
- `sym`: symmetry.
- `frac`: fractal dimension (“coastline approximation” - 1).

The purpose of this quiz is twofold:

1. **Learn to use the `glmnet` package for variable selection with GLM models.**

   Recall that documentation on any **R** function can be obtained via e.g., the command `?glm`. In addition, the full list of functions in the package can be obtained via `help(package = "glmnet")`. Finally, many packages provide tutorials known as “vignettes”. You can check whether a package has tutorials with `vignette(package = "glmnet")`. In this case, the relevant tutorial can be accessed with `vignette("glmnet_beta")`.

2. **Learn to use the package `rmarkdown` for automatically generating well-organized and well-formated reports.**

   In fact, the R Markdown (`.Rmd`) file you are submitting is not the report itself, but rather a human-readable file which can be converted to a PDF document using the **R** command `rmarkdown::render("uwname-quiz2.Rmd")`. An in-depth tutorial on `rmarkdown` can be found here. To make sure your `Rmd` file converts to a PDF document, see instructions here.

Q1. Using the **R** package `glmnet`, determine the important predictors of malignant tumors using a penalized logistic regression. The logistic regression model is

$$y_i \mid \boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho_i), \qquad \rho_i = \text{logit}^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{1}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})},$$

where y is the diagnostic (`diag`) binary response variable, and $\boldsymbol{x}$ is a vector of 31 predictors (including the intercept term).

The logistic regression model results in a loglikelihood function $\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X})$. The penalty function on $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{30})$ is Elastic Net, such that for fixed α and λ, the penalized likelihood estimator is

$$\tilde{\boldsymbol{\beta}} = \operatorname*{arg\,max}_{\boldsymbol{\beta}} \left[\ell(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) - \lambda \sum_{j=1}^{30} (1-\alpha)\beta_j^2 + \alpha|\beta_j| \right].$$

Note: the intercept term is *not* penalized. Also, by default each of the covariates is scaled to have a standard deviation of 1. However, the relevant option in `glmnet` does this automatically for you, and conveniently reports estimates of $\boldsymbol{\beta}$ on the original scale.

(a) Much like `lars`, `glmnet` can be used to produce the entire solution path $\tilde{\beta}(\lambda)$ for $\lambda \in (0, \infty)$. Plot the entire solution path for the penalized logistic regression with the WDBC data for $\alpha = 0.5$.

(b) For fixed α, `glmnet` helps you select the optimal value of λ by a procedure called K-fold cross-validation, which works like this:

i. Randomly divide the full dataset into training and testing samples $(\boldsymbol{y}_{\text{train}}, \boldsymbol{X}_{\text{train}})$ and $(\boldsymbol{y}_{\text{test}}, \boldsymbol{X}_{\text{test}})$. The exact way this is done will be specified momentarily.

ii. Using only the training data to estimate $\tilde{\boldsymbol{\beta}}_{\text{train}}(\lambda)$, the predicted probability of a new tumor with covariate $\boldsymbol{x}_\star$ being malignant is

$$\hat{\rho}_{\text{train}}(\boldsymbol{x}_\star, \lambda) = \operatorname{logit}^{-1}\Big(\boldsymbol{x}_\star' \tilde{\boldsymbol{\beta}}_{\text{train}}(\lambda)\Big).$$

iii. A cross-validation (CV) method for estimating the optimal value of λ is to minimize the out-of-sample negative-loglikelihood, or “Deviance” criterion:

$$\hat{\lambda} = \arg\min_{\lambda} - \sum_{i=1}^{n_{\text{test}}} \Omega\Big(y_{i,\text{test}} \mid \hat{\rho}_{\text{train}}(\boldsymbol{x}_{i,\text{test}}, \lambda)\Big), \tag{1}$$

where $\Omega(y \mid \rho) = y \log \rho + (1 - y) \log(1 - \rho)$ is the log-PDF of $y \sim \text{Bernoulli}(\rho)$.

iv. In order to determine the training and test sets, K-fold cross-validation randomly divides the whole sample into K disjoint groups, and solves (1) K times, using each group once as the test set with the remaining $K - 1$ groups used for training.

Plot the results of a 15-fold cross-validation on the Deviance metric (1) as a function of $\log(\lambda)$, next to the plot of the full solution path (i.e., the second plot should not overwrite the first).

(c) The most common CV estimate of λ is not in fact the minimum value of (1), but one standard error larger than it. Let’s call this value $\hat{\lambda}_{\text{1se}}$. The idea is to regularize the model a bit more (i.e., larger penalty), without straying too far from the minimum value in (1). Add a vertical line at $\hat{\lambda}_{\text{1se}}$ to the plot in a **Q1(a)**.

(d) Nicely display the non-zero penalized regression estimates $\tilde{\boldsymbol{\beta}}(\hat{\lambda}_{\text{1se}})$.

Q2. In the file **`uwname-quiz2.Rmd`**, create an R Markdown script to automatically generate a report presenting the results of **Q1**. The generated report must display all your **R** code, and consist of the following sections:

- *Data and Model:* Give a brief description of the dataset and write down the logistic regression model along with the penalized likelihood estimator. Feel free to copy-paste parts from the Quiz itself; the point here is for you to practice Markdown formatting. Most of it is very straightforward, except *perhaps* the math symbols which are encoded using LaTeX notation. A short tutorial covering just about all the math notation you’ll ever need can be found here.
- *Results:* This consists of two subsections:

- ⋆ *Penalized Likelihood*, which produces the two plots created in **Q1(a-b-c)**. Briefly explain what you are plotting using 1-2 sentences.
- ⋆ *Variable Selection*, which produces the output of **Q1(d)**. Provide 1-2 sentences summarizing which features of tumor cell nuclei are most predictive of a malignant growth.

  Note: It is poor form to display a 1-column matrix in a report, as it wastes a lot of vertical space. Convert to a named vector or 1-row matrix instead, or take a look at how to create `rmarkdown` tables.

Hint: `rmarkdown` is a powerful tool for combining **R** input/output into a legible document. However, it can take several tries to get the **R** code correctly commented and its output looking right, which takes far longer through the `rmarkdown::render` mechanism. Therefore, I rarely use `rmarkdown` as a first step. Instead, create a file called **`quiz2-test.R`** (which you won't turn in) in which you get the **R** part looking right. Then, create **`uwname-quiz2.Rmd`** and copy-paste the **R** code into the relevant sections.