Weekly Exercise - Week 5

Marjolein Fokkema

We use a real dataset on prediction of depressive disorder for this week's exercise. You can get the data from Brightspace, and read them into R as follows:

```
train <- readRDS("masq_train.Rda")
test <- readRDS("masq_test.Rda")
```

Or in Python:

```
import feather
train = feather.read_dataframe("masq_train.feather")
test = feather.read_dataframe("masq_test.feather")
```

Possible predictor variables are item scores on the Mood and Anxiety Symptom Questionnaire (`MASQ01` - `MASQ90`) and socio-demographic characteristics (`GENDER`, `Leeftijd`, `DEMOG1` - `DEMOG8`). The response variable is `D_DEPDYS`, whether the respondent has a current depressive or dysthymic disorder (0 = no, 1 = yes), as evaluated by a mental-health professional through a structured interview.

We will fit a range of penalized logistic regression models on the training dataset and compare their performance on a test dataset. You will do this using the cross-validation + test set approach.

a) Inspect multicollinearity between the numeric MASQ items. What do you expect about relative performance of lasso, ridge and elastic net regression?

b) Pick three candidate procedures from ridge, elastic net (with any $0 \leq \alpha \leq 1$), lasso, relaxed lasso. Ideally, this should be done by considering the training set (e.g., multicollinearity), and/or thinking about what would constitute a useful result (e.g., would a (non)-sparse solution be useful for reducing respondent burden of diagnostic procedures in clinical practice?), but you are allowed to just make a random choice.

c) Use library **`glmnet`** to fit the models, and select the most accurate model through 10-fold cross-validation on the training set.

Note that function `cv.glmnet` require a matrix of predictor variables, so a `data.frame` needs to be converted, first. E.g., in R:

```
x_train <- model.matrix(D_DEPDYS ~ . -1, data = train)
x_test <- model.matrix(D_DEPDYS ~ . - 1, data = test)
```

d) Compute the misclassification rate (MCR) on the test set.

e) Use the `coef` method to extract the selected variables and their coefficients from the best-performing model. From which of the following MASQ subscale were most items selected?

- Anhedonic Depression: Items 1, 14, 18, 21, 23, 26, 27, 30, 33, 35, 36, 39, 40, 44, 49, 53, 58, 66, 72, 78, 86 and 89.
- Anxious Arousal: Items 3, 19, 25, 45, 48, 52, 55, 57, 61, 67, 69, 73, 75, 79, 85, 87 and 88.
- General Distress Depression: Items 6, 8, 10, 13, 16, 22, 24, 42, 47, 56, 64 and 74.
- General Distress Anxiety: Items 2, 9, 12, 15, 20, 59, 63, 65, 77, 81 and 82.
- General Distress Mixed: Items 4, 5, 17, 29, 31, 34, 37, 50, 51, 70, 76, 80, 83, 84 and 90.