# Computer Exercise 2

Sep. 22, 2022

**Task:**

Classify patients' survival (0: survived; 1: dead) using 108 features (a mixture of numeric or binary variables) from their Intensive Care Unit (ICU) records, such as age, BMI, height, weight, heart rate, blood pressure, etc. (Detailed descriptions available in the data folder.)

**Goal:**

Write your own computer programs of Fisher Linear Discriminant (FLD) without directly importing packages of the algorithm. Find public packages for Perceptron and Logistic Regression (LR) or write your own codes. Do experiments with FLD, Perceptron and LR on the data. Make observations the learning procedures, performances and effects of optional choices on the performance.

**Data:**

Please check the "data1forEx1to4" folder for the following datasets.

| Datasets | Sample size | Feature data file | Class label file |
|---|---|---|---|
| **TrainingSet-1** | 5000 | train1_icu_data.csv | train1_icu_label.csv |
| **TrainingSet-2** | 1475 | train2_icu_data.csv | train2_icu_label.csv |
| **TestSet-1** | 1097 | test1_icu_data.csv | test1_icu_label.csv |
| **TestSet-2** | 450 | test2_icu_data.csv | test2_icu_label.csv |

Note: You may need to scale features to the same reasonable range before training. Meaning of each feature can be found in "feature_description.csv". Original data were from Kaggle (https://www.kaggle.com/c/widsdatathon2020/data).

**Experiment 1 (FLD):**

Use TrainingSet-1 to calculate the discriminant function using FLD. Apply the discriminant function on TestSet-1. Calculate the error rate.

**Experiment 2 (Perceptron):**

1) Use TrainingSet-1 to train the Perceptron classifier. Calculate the training error and cross validation error on the training set. Apply the trained Perceptron to TestSet-1. Compare the training error rate, cross-validation error rate and test error rate.
2) Use TrainingSet-2 to train the Perceptron classifier. Apply the Perceptron trained with TrainingSet-2 to TestSet-1 and TestSet-2, respectively. Compare the error rates.
3) Discuss your observations on the results.

**Experiment 3 (LR):**

1) Use TrainingSet-1 to train the classifier with Logistic Regression. Calculate the training error and cross validation error on the training set. Apply the trained classifier on TestSet-1. Calculate the test error.
2) Using Python package matplotlib to draw the ROC curve according to the test results.
3) Analyze the significance of association between each feature and patients' survival. You may need to study some materials beyond the course content by yourself (hints: you may refer to available Python packages like statsmodels).

## Experiment Report:

- Write an experiment report to describe and analyze the experiment observations (no more than 4 pages).
- Provide detailed supplementary materials that should include at least the following:
  - A readme file containing information on all supplementary files, programming environment and parameters used in the experiments (if any),
  - Source codes of your own program (TAs should be able to run the code and reproduce your experiments),
  - Links to the original source of the packages you used, and
  - Experiment result files.

**Due date: Oct. 5 (Wed.) 23:00 Beijing time**