Statistical Learning, LAB 1

Xijia Liu

Task 1: About Proceptron Algorithm

In this task, we implement the Perceptron algorithm in R.

Task 1.1: Write your own program to implement the algorithm in R

Task 1.2: Use the R code 'Task1-DataGeneration.R' to generate the data. Apply your function of perceptron on the data and visualize the decision boundary in a plot.

Tips: The perceptron model can be represented as $Sign\left(\mathbf{w}^\top \mathbf{x}\right)$, where $\mathbf{x} = (1, x_1, \ldots, x_p)^\top$, $\mathbf{w} = (w_0, w_1, \ldots, w_p)^\top$

Task 2: About Ridge Regression

In this task, we implement Ridge Regression and train a predictive model with the 'Boston' data. **Task 2.1**: Implement Ridge Regression (RR) in R. Write an R function 'Rreg'. The inputs should include

- 'trainX': a data matrix containing feature variables.
- 'trainY': a numeric vector containing the target variable.
- 'lambda': a scalar to specify the shrinkage parameter in KRR.

Task 2.2: Apply function 'Rreg' (results of Task 3.1) to 'Boston' data

- Set the random seed as '2023', Draw a random sample with 400 observations (use build-in function 'sample') from the Boston data as the training data set. Use the rest observation as the testing data.
- Target variable: 'medv' (median value of owner-occupied homes in 1000s.)
- Feature variables: all remaining variables
- Candidate values for tuning parameter: $0, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1$
- Apply 10-fold-cross-validation on the training set and select the 'best' tuning parameter. Root mean square error can be used as the metric of performance.
- Estimate the performance of the model with the 'best' tuning parameter with the testing set.

Task 3: About Logistic Regression with LASSO Penalty

Statistical analyses of high-dimensional omics data, e.g. methylation data or gene-expression data, from patients with brain cancer, can address several problems. A central problem is to derive a classifier that allows us to predict which sub-type of brain cancer the patient has. In this task, we will try to solve this kind of problem with Logistic Regression with a LASSO penalty.
Description of the data: Gene-expression data from 226 patients diagnosed with brain cancer were observed. The data were generated by an RNA-sequencing technique (Illumina HiSeq 2000 RNA Sequencing Version 2) conducted on tumor samples from the patients. For each patient, the gene-expressions were measured on 20,532 genes. The pre-processed gene expression measurements are stored in the file named '*GeneExpressionData.txt*'. This file consists of 20,532 rows, corresponding to the different genes, and 226 columns representing the patients. In addition to the gene-expression data we have additional data linked to the patients, so-called

'*meta-data*'. For each patient, several additional variables were observed, for example: gender, age, survival data (dead or alive 5 years after diagnosis), and sub-type (the last variable 'V31', two sub-types 'IDHmut-codel' and 'IDHmut-non-codel'). **Task**: Train a penalized logistic regression classifier to predict the subtype of brain cancer with the raw gene expression data.

- Set the random seed as '2023', Draw a random sample with 181 observations from the data as the training data set. Use the rest observation as the testing data.
- Train a logistic regression with LASSO penalty with the training set.
- 10-fold cross-validation method is recommended.
- Use default candidate lambda values by 'cv.glmnet' function

Report your best model with a list of selected gene variables and estimate the accuracy and kappa statistics of your final model.