Machine Learning

Problem Set 10/11

Hesam Montazeri
Fereshteh Fallah
Mozhgan Mozaffari Legha

Khordad 14, 1399
(June 3, 2020)

Problem 1: Review part
Write your reviews for the whiteboard notes and the slides of the lectures of the past two weeks. Write down all formulas and explain in detail each step of the derivations, if applicable.

Problem 2: Conceptual questions
[ISL] chapter 8: questions 1, 3, 4, 5; programming question 8

Problem 3: Programming: bootstrapped confidence interval
Your task is to study logistic regression model on a bioinformatics dataset of your choice. Compare regression coefficient bootstrapped confidence intervals with those of built-in implementation in R.

Problem 4: Programming: Liquid Biopsy for Cancer Diagnosis
Early detection of cancer types is a major challenge in cancer diagnosis. Cancer is primarily diagnosed by clinical presentation, radiology, biochemical tests, pathological analysis and molecular profiling of tumor tissue. The emerging non-invasive blood-based *liquid biopsies* provides a promising alternative diagnostic tool to cancer care. In this exercise, you will investigate whether gene expression profiles of *tumor-educated blood platelets* (TEPs) can be used for identification of six cancer types namely breast, hepatobiliary, colorectal cancer, glioblastoma, pancreatic, and non-small cell lung cancer as well as healthy samples. Your tasks are given below:

(a) Explore the data.

(b) **Two-class classification:** the second task is to determine the presence of cancer, of any type, in the input sample according to the gene expression profile. Assume the response variable $Y = 1$ denotes a cancer sample and $Y = 0$ indicates a healthy donor. Explore the applications of logistic regression, LDA, SVM (linear kernel and at least two non-linear kernels), KNN, classification tree, random forest, and boosted trees on this problem. Use McNemar's test for comparing top three classifiers

(c) **Multi-class classification:** the third task is to perform a multi-class classification where the response variable Y can take seven categories namely six cancer types and the *healthy donor* category. Explore logistic regression, softmax regression, LDA, SVM (linear kernel and at least two non-linear kernels), KNN, classification tree, random forest, and boosted trees for multi-class problem. If needed use one-versus-one and one-versus-all approaches.

Provide a sound statistical analysis (cross validation, confidence intervals, etc). If necessary, avoid the overfitting by performing feature selection or regularization. Assess the performance of your method by appropriate measures (accuracy, sensitivity, specificity, ROC curve, ...).

We encourage discussing the problems with other students, however, similarity between solutions is not allowed. (**Important**) Studying any online or previous solutions, no matter to what extent, is strictly forbidden and is considered as a violation of the academic honor code. Submit your solutions by Khordad 22, 1399.