STAT 577: Homework 3

Due: April 1 , 2022

You will be graded on completeness, correctness, and clarity of your results. You are encouraged to work in groups, but you have to write your own report.

For this assignment, use the seed number of 577. You will have to set a seed everytime you split the data, and everytime you call the `train` function.

Energy efficiency (`energy_efficiency_homework3.csv`)

This data concerns energy efficiency analysis. The data is generated using 12 different simulated building shapes. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters, resulting in 768 building shapes. The dataset comprises 768 samples and 7 predictors, aiming to predict a real valued response. The predictors are: `Relative Compactness, Surface Area,Wall Area, Overall Height, Orientation, Glazing Area, Glazing Area Distribution`, and the response is `Heating Load`. See https://archive.ics.uci.edu/ml/datasets/Energy+efficiency for more details.

Your task is to build a predictive model for `Heating Load` using the other characteristics of a building. Using the `train` function from the `caret` package, train the predictive models below. For each model, be sure to report the performance metrics.

Write a few sentences comparing the performances of the methods. Also describe the meaning of the estimated RMSE for the best performing model, speculate why the best model you find could be the best for this data, and report variable importance plot with comments on the relative importance of the building characteristics on predicting the `Heating Load` of the building.

Use 70% of the data for training and the rest as a holdout.

- The multiple linear regression model
- The Ridge regression model
- The Lasso regression model
- The Elastic Net model, set `alpha = seq(0.1,0.9,by=0.1)`
- The principal components regression model
- The partial least squares regression model
- The k-nearest neighbors regression model (try `k=1:25`)

Leukemia Data: `leukem_std.csv`

This is a microarray gene expression data orginally used in this science paper https://www.ncbi.nlm.nih.gov/pubmed/10521349. The data frame (`leukem_std.csv`) is a 72×7130 matrix, where the columns correspond to genes (predictors), with the last column being the response. The rows correspond to tissue samples (observations). The first 38 tissue samples are for ALL (Acute Lymphoblastic Leukemia), the next 9 tissue samples are AML-B (Acute Myeloid Leukemia - Cell B) and the last 25 tissue samples are AML-T (Acute Myeloid Leukemia - Cell T).

Data pre-processing: I have centered and scaled the expression profiles of each gene to have mean 0 and variance 1. Then I ordered the genes in increasing order of an $F - test$ which compares the means of the three classes. That is, the first gene in the data, first column, has the largest $F - test$ (One Way ANOVA). **Use the first 8 genes for your analysis.**

Your task is to build a classification model to predict the cancer type using the 8 gene expression profiles of the tissue. Using the `train` function from the `caret` package, train the classification models below. Use the correct classificaton rate (accuracy) as a metric to train. Report the performance metrics for each of the methods, which method performs best for this data? Can you speculate why? Also, report a variable importance plot showing relative importance of the 8 gene expression profiles in predicting the cancer type of the tissue.

Use 80% of the data for training and the rest as a holdout.

- KNN Classifier
- LDA
- Quadratic Discriminant Analysis
- Logistic regression
- Regularized logistic regression
- Linear SVMs (SVC)
- Kernel SVMs - radial kernels

WINE Quality Data (`WINE.csv`)

This data is related to red and white vinho verde wine samples, from the north of Portugal. The data contains informatio on:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality (Output variable based on sensory data, a factor with levels `high` `low`)

This is the famous wine dataset from the UCI data repository https://archive.ics.uci.edu/ml/datasets/Wine+Quality with some modifications. Namely, the quality in the original data was a score between 0 and 10. These has been coded as either high or low. See description on UCI for description of variables.

You task is to build a classification model to predict the quality of wine based on its chemical characteristics. Using the `train` function from the `caret` package, train the classification models below. **use the AUC as a metric to train and select the best model**. Report the performance metrics for each of the methods. Also, report a variable importance plot showing which chemical characteristics are most predictive of the wine quality.

Use 50% of the data for training and the rest as a holdout.

- KNN Classifier
- LDA
- Quadratic Discriminant Analysis
- Logistic regression
- Regularized logistic regression
- Linear SVMs (SVC)
- Kernel SVMs - radial kernels