Practice

Data set: cpuperform.csv

Create an OLS regression model to predict the relative CPU performance (*prp*) based on the following variables: *myct, mmin, mmax, cach, chmin, chmax*. Validate your model using both the validation set method and the k-fold cross-validation method.

Data set: education.csv

Create an OLS regression model to predict the expenditure on public education (*expend*) using the following predictors: *urban, income* and *teen*. Validate your model with the validation set approach. (Retain 30-35 cases for the training set and the others for the test set.)

Data set: winequality.csv

Your task is to find the best predictors for the wines quality (*quality*) from the following 11 variables: *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates* and *alcohol*. To that effect, use all of the following techniques:

- best subset selection regression
- forward and backward stepwise regression
- ridge regression
- lasso regression
- PLS regression

Identify the model that provides the best prediction accuracy in the test set.

Data set: housedata.csv

You are supposed to find the best predictors for a house price (*price*) out of the following variables: *bedrooms, bathrooms, sqft_living, sqft_lot, floors, grade, sqft_basement* and *old*. Use all of the following techniques:

- best subset selection regression
- forward and backward stepwise regression
- ridge regression
- lasso regression
- PLS regression

Discover the model that ensures the best prediction accuracy in the test set.

Data set: bostonhousing.csv

You have to predict the median house value (*medv*) using the following variables: *crim, zn, indus, nox, rm, age, dis, rad, tax, ptratio* and *lstat.* Identify the model with the highest prediction accuracy using these methods:

- best subset selection regression
- forward and backward stepwise regression
- ridge regression
- lasso regression
- PLS regression

Data set: credit.csv

Your task is to predict the customers credit score (*rating*) knowing the following variables: *age, income, cars, education* and *carloans*. Use the following machine learning techniques:

- logistic regression
- naïve Bayes estimation
- neural networks

Which technique gives us the best prediction accuracy in the test set?

Data set: winequality.csv

Try to predict the type of wine (*type*) based on the following variables: *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates* and *alcohol*. Use these prediction techniques:

- logistic regression
- lasso logistic regression
- linear discriminant analysis
- quadratic discriminant analysis
- naïve Bayes estimation
- k nearest neighbor
- support vector machine
- neural networks

What is the greatest prediction accuracy (in the test set) that you can get?

Data set: directmail.csv

You have to predict whether a customer would respond to a direct mail campaign or not (*previous*). Your predictors are: *age, income, education, reside, gender* and *children*. Use all the following techniques:

- logistic regression
- lasso logistic regression
- linear discriminant analysis
- quadratic discriminant analysis
- naïve Bayes estimation
- k nearest neighbor
- support vector machine
- neural networks

Indicate the technique that ensures the best prediction accuracy in the test set.

Data set: vehicles.csv

Your task is to predict the vehicle type (*type*) knowing the values of the following variables: *engine, horse, weight, length* and *fuelcap*. The machine learning techniques that you must use are:

- logistic regression
- linear discriminant analysis
- quadratic discriminant analysis

- naïve Bayes estimation
- k nearest neighbor
- support vector machine
- neural networks

Which method gives the highest prediction accuracy in the test set?

Data set: ulcer_recurence.csv

Predict whether a patient presents ulcer (*result*), with the following variables: *age, duration* and *visit*. Use the following prediction methods:

- k nearest neighbor
- support vector machine
- neural networks

Data set: forestfires.csv

Predict whether a fire would take place (*fire*), based on the following predictors: *temp, RH, wind* and *rain*. Use the following techniques:

- k nearest neighbor
- support vector machine
- neural networks

Data set: education.csv

You have to predict the expenditure on public education (*expend*) based on the following predictors: *urban, income* and *teen* and using regression trees.

Data set: credit_card.csv

Predict the amount spent by a credit card owner (*spent*) using regression trees. The predictors are: *gender, card, type* and *items*.

Data set: winequality.csv

Predict wines type (*type*) using classification trees. Use the boosting technique to choose the best predictors from the following: *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates* and *alcohol*.

Data set: credit.csv

Your task is to predict the customers credit score (*rating*) using classification trees. The independent variables are: *age, income, cars, education* and *carloans*.

Data set: directmail.csv

You have to predict whether a customer would respond to a direct mail campaign or not (*previous*) using classification trees. The predictors are: *age, income, education, reside, gender* and *children*.

Data set: youngpeoplesurvey.csv

Perform a principal component analysis using the first 31 variables in this data set (from *Healthyeating* to *Charity*). Reduce these variables to a small number of relevant factors and try to give a “name” or a “label” to each factor, based on the variables that are correlated with it.

Data set: education.csv

Group the 50 states into classes through a hierarchical cluster analysis, using the variables *urban, income, teen* and *expend*. Which would be the optimal number of classes? Give a relevant name to each class (cluster).

Data set: bostonhousing.csv

Group the observations in this dataset in classes, using a k-means cluster analysis. The clustering variables should be all the variables in the dataset, except *chas*. (You can try with various numbers of clusters, to see which of them gives the best results).

Data set: cpuperform.csv

Group the observations in this dataset in classes, using a k-means cluster analysis. The clustering variables should be all the variables in the dataset, except *vendor* and *erp*.