Learning outcomes

After solving these exercises, you should be able to understand the following:

1. Applying the Random Forest and Adaboost algorithms to solve classification problems.
2. Applying stacking techniques
3. Interpreting the results generated from each algorithm in R.
4. Comparison of the model performance in terms of precision, recall and accuracy

Random Forest: Hepatitis Dataset

The hepatitis dataset has 20 variables and 155 records. Use "target" as target variable. Goal is to determine whether the person will be detected with hepatitis or not.

1. Import the data into R

```
hepatitis <- read.table('hepatitis.txt', header=F, dec='.',
                        col.names=c('target','age','gender','steroid','antivirals',
                        'fatigue','malaise','anorexia','liverBig',
                        'liverFirm','spleen','spiders','ascites',
                        'varices','bili','alk','sgot','albu','protime',
                        'histology'),
                        na.strings=c('?'), sep=',')
```

2. Study dataset

```
str(hepatitis)
table(hepatitis$target)
str(hepatitis$target) # 1: Die (+ve) ; 2: Live (-ve)
#Convert 1s and 2s into 1s and 0s
hepatitis$target= ifelse(hepatitis$target==1,0,1 ) # 0: Die (+ve); 1: Live (-ve)
```

3. Convert all categorical attributes into factors
4. Fill the missing values using knnImputation.
5. Split dataset into train and test
6. Build the classification model using randomForest
7. Build the classification model using randomForest

```
library(randomForest)
hepatitis_rf <- randomForest(target ~ ., data=trainR, ntree=30,mtry = 4)
```

8. View results and understand important attributes

```
print(hepatitis_rf)
hepatitis_rf$predicted
hepatitis_rf$importance
```

9. View results and understand important attributes

```
plot (directly prints the important attributes)
varImpPlot(hepatitis_rf)
```

10. Predict on Train and Test datasets
11. Calculate precision, recall and accuracy

Adaboost: Universal Bank Dataset

The Universal Bank dataset has 14 variables and 5,000 records. Use "Personal.Loan" as target variable.

1. Import the data into R
2. Drop ID & ZIP Code
3. Separate numerical and categorical attributes
4. Convert all categorical attributes into factors (except the target variable)
5. Dummify all categorical attribute
6. Descritize all numerical attributes (optional; you can pick what to bin)
7. Combine categorical and numerical attributes
8. Standardize the combined dataset
9. Add the target variable back into the dataset
9. Split the data into train, test and evaluation data sets
10. Build the classification model using Adaboost:

```
library(ada)
x = subset(train_data, select = -Personal.Loan)
y = as.factor(train_data$Personal.Loan)
```

```
a = subset(test_data, select = -Personal.Loan)
b = as.factor(test_data$Personal.Loan)
model = ada(x, y, iter=20, loss="logistic") #20 Iterations
```

11. Predict the values using model on test data sets.

```
pred = predict(model, a); pred
```

12. Calculate precision, recall and accuracy

```
result <- table(pred, b);result # 0(-ve) and 1(+ve)
```

13. Experiment with different number of iterations and find the best.