Exercise 9 – Random Forests

Introduction to Machine Learning

Hint: Useful libraries

R

```
# Consider the following libraries for this exercise sheet:

library(proxy)
library(mlr3)
library(rpart.plot)
library(mlr3learners)
library(data.table)
library(mlr3verse)
```

Python

```
# Consider the following libraries for this exercise sheet:

# general
import numpy as np
import pandas as pd
from scipy.spatial.distance import pdist
from scipy.sparse import dok_matrix
# plots
import matplotlib.pyplot as plt
# sklearn
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.inspection import permutation_importance
from sklearn.model_selection import train_test_split
```

Exercise 1: Bagging

Only for lecture group A

Learning goals

1. Understand benefit of bagging from a mathematical perspective
2. Solve "show that…"-type exercises
3. Handle expectations over random variables

In this exercise, we briefly revisit why bagging is a useful technique to stabilize predictions.

For a fixed observation $(\mathbf{x}, y)$, show that the expected quadratic loss over individual base learner predictions $b^{[m]}(\mathbf{x})$ is larger than or equal to the quadratic loss of the prediction $f^{[M]}(\mathbf{x})$ of a size-M ensemble.

You can consider all hyperparameters of the base learners and the ensemble fixed.

Hint

Use the law of total expectation ("Verschiebungssatz der Varianz": $\mathsf{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 \iff \mathbb{E}(Z^2) = \mathsf{Var}(Z) + (\mathbb{E}(Z))^2$, where $\mathsf{Var}(Z) \geq 0$ by definition.) stating $\mathbb{E}(Z^2) \geq (\mathbb{E}(Z))^2$ for a random variable Z.

Exercise 2: Classifying spam

Learning goals

1) Apply RF to data for prediction, OOB error estimation & feature importance computation
2) Understand how 63% probability for observations to end up in a tree comes about

Only for lecture group B

Take a look at the `spam` dataset and shortly describe what kind of classification problem this is. [only for lecture group B]

Hint

R

Access the corresponding task `?mlr3::mlr_tasks_spam`.

Python

Read spam.csv.

> Only for lecture group B

Use a decision tree to predict `spam`. Re-fit the tree using two random subsets of the data (each comprising 60% of observations). How stable are the trees?

Hint

R

Use `rpart.plot()` from the package `rpart.plot` to visualize the trees.

Python

Use `from sklearn.tree import plot_tree` to visualize the trees.

Forests come with a built-in estimate of their generalization ability via the out-of-bag (OOB) error.

i. Show that the probability for an observation to be OOB in an arbitrary bootstrap sample converges to $\frac{1}{e}$.
ii. Use the random forest learner (R: `classif.ranger`, Python: `RandomForestClassifier()` to fit the model and state the out-of-bag (OOB) error.

You are interested in which variables have the greatest influence on the prediction quality. Explain how to determine this in a permutation-based approach and compute the importance scorses for the `spam` data.

Hint

R

Use an adequate variable importance filter as described here.

Python

Choose an adequate importance measure as described here.

Exercise 3: Proximities

> Learning goals
>
> 1) Be able to make predictions from code output for RF
> 2) Compute proximities

You solve the `wine` task, predicting the `type` of a wine – with 3 classes – from a number of covariates. After training, you wish to determine how similar your observations are in terms of proximities.

The model information was created with `ranger::treeInfo()`, which assigns observations with values larger than `splitval` to the right child node in each split.

observation	alcalinity	alcohol	flavanoids	hue	malic	phenols
1	11.4	14.75	3.69	1.25	1.73	3.10
2	25.0	13.40	0.96	0.67	4.60	1.98
3	17.4	13.94	3.54	1.12	1.73	2.88

```
[1] "Tree 1:"
```

nodeID	leftChild	rightChild	splitvarID	splitvarName	splitval	terminal	prediction
0	1	2	5	phenols	1.94	FALSE	NA
1	3	4	1	alcohol	12.43	FALSE	NA
2	5	6	1	alcohol	13.04	FALSE	NA

nodeID	leftChild	rightChild	splitvarID	splitvarName	splitval	terminal	prediction
3	NA	NA	NA	NA	NA	TRUE	2
4	NA	NA	NA	NA	NA	TRUE	3
5	NA	NA	NA	NA	NA	TRUE	2
6	NA	NA	NA	NA	NA	TRUE	1

```
[1] "Tree 2:"
```

nodeID	leftChild	rightChild	splitvarID	splitvarName	splitval	terminal	prediction
0	1	2	1	alcohol	12.78	FALSE	NA
1	3	4	3	hue	0.68	FALSE	NA
2	5	6	2	flavanoids	2.18	FALSE	NA
3	NA	NA	NA	NA	NA	TRUE	3
4	NA	NA	NA	NA	NA	TRUE	2
5	NA	NA	NA	NA	NA	TRUE	3
6	NA	NA	NA	NA	NA	TRUE	1

```
[1] "Tree 3:"
```

nodeID	leftChild	rightChild	splitvarID	splitvarName	splitval	terminal	prediction
0	1	2	1	alcohol	12.79	FALSE	NA
1	3	4	5	phenols	2.01	FALSE	NA
2	5	6	5	phenols	2.28	FALSE	NA
3	NA	NA	NA	NA	NA	TRUE	2
4	NA	NA	NA	NA	NA	TRUE	2
5	NA	NA	NA	NA	NA	TRUE	3
6	NA	NA	NA	NA	NA	TRUE	1

For the following subset of the training data and the random forest model given above,

find the terminal node of each tree the observations are placed in, compute the observations' pairwise proximities, and

---

construct a similarity matrix from these proximities in `R` resp. `Python`.