Exercise 1: Bagging [only for lecture group A]

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 expected quadratic loss of the prediction $f^{[M]}(\mathbf{x})$ of a size-M ensemble ($M \in \mathbb{N}$).

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

Hint: Use the law of total expectation ("Verschiebungssatz der Varianz"[1]) stating $\mathbb{E}(Z^2) \geq (\mathbb{E}(Z))^2$ for a random variable Z.

Exercise 2: Classifying `spam`

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

   (R Hint: access the corresponding task `?mlr3::mlr_tasks_spam`)
   (Python Hint: read `spam.csv`)

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? [only for lecture group B]

   (R Hint: Use `rpart.plot()` from the package `rpart.plot` to visualize the trees.)
   (Python Hint: Use `from sklearn.tree import plot_tree` to visualize the trees.)

c) 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.

d) 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.

   (R Hint: use an adequate variable importance filter as described in
   `https://mlr3filters.mlr-org.com/#variable-importance-filters`.)
   (Python Hint: choose an adequate importance measure as described in
   `https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html`)

Exercise 3: 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.

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

a) find the terminal node of each tree the observations are placed in,

b) compute the observations' pairwise proximities, and

[1] $\mathsf{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 \iff \mathbb{E}(Z^2) = \mathsf{Var}(Z) + (\mathbb{E}(Z))^2$, where $\mathsf{Var}(Z) \geq 0$ by definition.

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

R Hint: 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

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
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

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

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