Data Mining

ENSIA 2025-2026

Lab sheet N°6: Feature Selection and Classification

Objectives

- Introduction to Feature Selection Methods in Scikit-learn
- Learn to determine which feature selection is appropriate and how to apply it effectively
- Evaluate the computational complexity (execution time) of feature selection techniques
- Explore strategies for combining different feature selection methods
- Explore the impact of feature selection on model performance
- Understand how decision tree algorithms choose the best split based on training data

Resources

- **Fetch OpenML datasets:** **fetch_openml — scikit-learn 1.7.2 documentation**
- **Forest Covertype dataset**: **The Forest Cover Type Dataset — scikit-learn 1.7.2 documentation**
- **Feature Selection:**
  - **1.13. Feature selection — scikit-learn 1.7.2 documentation**
  - **Feature Selection Methods in Scikit Learn | Medium**
  - **SelectKBest — scikit-learn 1.7.2 documentation**
- **Sequential Feature Selection:**
  - **SequentialFeatureSelector: Forward and Backward feature selection - mlxtend**
  - **SequentialFeatureSelector — scikit-learn 1.7.2 documentation**
- **Genetic algorithms:**
  - **sklearn-genetic-opt**
  - **GAFeatureSelectionCV — sklearn genetic opt 0.12.0 documentation**

Notebooks

- **Non-guided Notebook:** Feature_selection_Non_Guided_.ipynb

Part 1: Feature Selection with Scikit-learn (90 minutes)

- Activate **DM_ENV** and install the required Python libraries: **mlxtend** and **sklearn-genetic-opt**
- Alternatively, students who find problems setting up the environment can use Google Colab
- Fill in the gaps and write the missing code in the **non-guided** Jupyter Notebook file
- Take a look at the provided resources (documentation, tutorial, reference card) for more info

Part 2: Exercises on the chapter Classification (90 minutes)

Exercise 1:

Consider the training examples shown in the table for a binary classification problem.

Instance	a1	a2	a3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

1. What is the entropy of this training set with respect to the class attribute?
2. Compute the information gain for **a1** and **a2**
3. Compute the information gain for **a3** (Exercise1_continuous_split.xlsx)
4. What is the best split among **a1**, **a2**, and **a3** according to the information gain?
5. What is the best split between **a1** and **a2** according to the misclassification error rate?
6. What is the best split between **a1** and **a2** according to the Gini index?

Exercise 2:

Consider the dataset below for a binary class problem.

1. Calculate the information gain when splitting the dataset on **A** and **B**.
   Which attribute would the decision tree induction algorithm choose?
2. Calculate the gain in the Gini index when splitting the dataset on **A** and **B**.
   Which attribute would the decision tree induction algorithm choose?
3. In the lecture, we have shown that entropy and the Gini index are both monotonically increasing in the range [0, 0.5] and decreasing in the range [0.5, 1].
   Can information gain and Gini index gain favor different attributes? Explain.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

Exercise 3 (For students):

Consider splitting a parent node **P** in two child nodes **C1** and **C2** using some test condition. The composition of labeled training instances at every node is summarized in this table:

	P	C1	C2
Class 0	7	3	4
Class 1	3	0	3

1. Calculate the Gini index and misclassification error rate of the parent node **P**.
2. Calculate the weighted Gini index of the child nodes.

   Would you consider this attribute test condition if the Gini index is used as the impurity measure?
3. Calculate the weighted misclassification rate of the child nodes.

   Would you consider this attribute test condition if the misclassification rate is used as the impurity measure?