Data Mining

ENSIA 2025-2026

Lab sheet N°5: PCA & Feature Selection

Objectives:

- Understand and apply PCA for dimensionality reduction.
- Analyze the effects of dimensionality reduction on data visualization.
- Investigate methods for selecting the most relevant features to enhance data mining models.
- Compute entropy and information gain in a binary classification context.
- Analyze feature selection techniques and their computational complexity.

Exercise 1: (20 minutes)

Name and discuss four methods that perform effective dimensionality reduction and four methods that perform effective numerosity reduction.

Exercise 2: (60 minutes)

1. Give the steps of PCA.
2. Plot the centered dataset **X** with 5 data points in a 2D plane.

   The points are (1,1), (2,2), (-1,-1), (-2,-2), (-1,1), (1,-1).
3. Calculate the covariance matrix for the previous dataset **X**.
4. Determine the eigenvalues and eigenvectors of the covariance matrix.
5. Plot the two principal components (eigenvectors) on the same 2D plane as the original data points.
6. What is the explained variance ratio for both components?
7. Deduce geometric insights or conclusions from the PCA decomposition.
8. Compute the reduced dataset, **X_reduce**, using a single principal component.
9. Find the **X_approx** dataset, which represents the original data after reversing the reduction. (**For students**)

Exercise 3: (60 minutes)

We have a binary classification problem with a dataset of 100 samples.

The distribution of the target variable ***Y*** is 60 samples with **Y=0** and 40 samples with **Y=1**.

1. Calculate the entropy of the initial dataset.

Then, for the two features (X1, X2), compute the entropy of ***Y*** given each feature.

Here are the details for each feature:

- **Feature X1:**
    - When **X1=0**, there are 30 samples (**Y=0**: 15, **Y=1**: 15).
    - When **X1=1**, there are 70 samples (**Y=0**: 45, **Y=1**: 25).
- **Feature X2:**
    - When **X2=0**, there are 40 samples (**Y=0**: 25, **Y=1**: 15).
    - When **X2=1**, there are 60 samples (**Y=0**: 35, **Y=1**: 25).

2. Using the information gain formula, calculate the information gain for each feature and rank them in descending order.
3. Prove that entropy is always positive.
4. What is the highest achievable Information Gain when the target variable can take on **n** different categories? (**For students**)

Exercise 4: (40 minutes)

1. Calculate the maximum number of models trained in forward selection with ***100*** features, and similarly for backward elimination.
2. Analyze the computational demands (complexity) for both methods in scenarios with a high number of initial features (***N***) with the goal of selecting only a small subset with (***M***) features.
3. Which method is better at finding feature interactions: forward selection or backward elimination?
4. Discuss the possibility of combining the two methods to develop an improved method (**For students**).