Problemset (1),
Foundations of Machine learning,
Winter 2025

Maximilian Kasy

In this problem, you are asked to run some supervised learning algorithms in Python. (In the following practice sessions, you can use either R or Python, depending on your preferences.) Your code should run from start to end in one execution, producing all the output. Output and discussion of findings should be integrated in a Jupyter Notebook. Figures and tables should be clearly labeled and interpretable. The findings should be discussed in the context of the theoretical results that we derived in class.

You might want to consult the following references, as you solve these exercises:

- Chapter 5 of Python Data Science Handbook
- The documentation of scikit-learn.

1. Load the iris data set, using the seaborn package:

```
import seaborn as sns
iris = sns.load_dataset('iris')
X = iris.drop('species', axis=1)
y = iris['species']
```

2. Split the Iris dataset into training and testing sets using *train_test_split*.
3. Train a *LogisticRegression* model on the training set.
4. Evaluate the model's performance on the test set using *accuracy_score*.
5. Perform 5-fold cross-validation on the Iris dataset using *cross_val_score* with a *KNeighborsClassifier*, using 3 nearest neighbors.
6. Use *GridSearchCV* to find the best hyperparameter (number of nearest neighbors) for this classifier, using only the training data.
7. Compute and plot the confusion matrix for the optimal classifier that your found, using the test data.