Study Guide Machine Learning (Part 1, classical ML)

walter.daelemans@uantwerpen.be

Study the notebooks (both theory and code). Also look at the documentation associated with the different sklearn functions to understand better the meaning of those functions and their arguments (as far as used in the notebooks). Experiment with variations/ extensions of the code to thoroughly understand it. Also study the solutions of the homework.

If parts of the theory in the notebooks are unclear, use ChatGPT or other LLMs to get more explanation (or use more traditional sources: there exist many online tutorials, courses and blogs). Code snippets can be explained adequately by LLMs as well.

A few example exam questions

Introduction

- What factors have contributed to the success of machine learning?
- Explain the difference between supervised and unsupervised learning and provide an example of each.
- Outline the basic steps involved in a typical machine learning project.
- What ethical concerns are raised using machine learning, for example in the context of recognizing criminals based on their photograph?

Linear Regression

- Explain what the coefficients and intercept in a linear regression model are.
- Why is feature scaling necessary in machine learning?
- For a binary class ML task, explain f-score, precision and recall in terms of calculations on the confusion matrix of the task.
- What does a high R^2 (coefficient of determination) value signify about the performance of a linear regression model?
- Extend the following Python snippet so that a validation set is added (X_val and y_val) . The size of the validation set is the same size as the final train set.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
shuffle=True, random_state=42, test_size=.25)
```

Logistic Regression

- Why is it necessary to impute missing values before fitting a logistic regression model? What would happen if missing values were not handled?
- Why is accuracy not a good evaluation metric for imbalanced datasets (skewed data)?
- Adapt the following code snippet to include one-hot encoding (with OneHotEncoder() for the Embarked feature instead of using the current mapping. Why is that a better solution?

```
titanic['Embarked'] = titanic['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
```