Data Analytics / Data Mining and Machine Learning

Lab: Classification

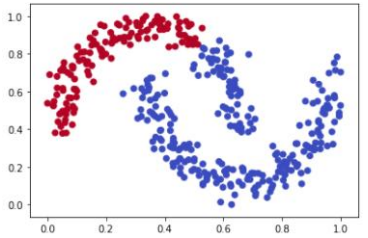
Lab: Classification

Data sets for the following exercises

Option 1) make_moons

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html#sklearn.datasets.make_moons

create with: `data, labels = make_moons(n_samples=200, noise=0.1, random_state=123)`



Option 2) wine data set (see https://scikit-learn.org/stable/datasets/toy_dataset.html)

- Chemical analysis to determine the origin of wines using the „wine“ data set.
- number of instances: 178
- number of features: 13
- number of „classes“: 3 different origins of Italian wine
- **features**: Alcohol ; Malic acid ; Ash ; Alcalinity of ash ; Magnesium ; Total phenols ; Flavanoids ; Nonflavanoid phenols ; Proanthocyanins ; Color intensity ; Hue ; OD280/OD315 of diluted wines ; Proline
- one column „class“: with the types of wine {1, 2, 3}

Lab: Classification

Exercise 1: Classification using k-nearest neighbors.

Alternative: if you prefer not to write programs, experiment here:

https://www.ml-and-vis.org/eduml

https://playground.tensorflow.org

a) Discuss: Do we need to scale the wine data set in order to classify it using k-NN?

b) Classify the wine data set using the nearest neighbor classifiers (1-NN, i.e. k=1).

- train and classify
- print the confusion matrix and the accuracy

c) Repeat the classification for different values of k (3, 5, 7,...) and compare the results.

d) Discuss the results with your neighbors.

Lab: Classification

Exercise 2: Classification using decision tree, SVMs, MLP (MLPClassifier).

a) Discuss: Do we need to scale the data set in order to classify it using decision trees? What about the other classifiers?

b) Classify the data set.

- train and classify
- print the confusion matrix and the accuracy

c) Optional for decision tree: Plot the tree, see
https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html

Which feature is at the root of the tree and what does that mean?

d) Discuss the results with your neighbours.

Lab: Classification

Additional Exercise: Free experimentation, learning by doing... „own data set“

- Do you have a favourite structured data set (matrix-like) at hand?
- Classify the data and discuss with other students