BUSINESS ACADEMY AARHUS

Exercises. Machine Learning. Week 7.

Sila, March 19th, 2025.

Exercise 1.

In groups of 2-3.

How can you spot an anomaly in a dataset

(or: might be able to spot) ? What can you do? What should you do?

Exercise 2.

In groups of 2-3.

If your model is underfitting or overfitting what might be able to make the model better? What could you try out?

Exercise 3:

In groups of 2-3.
Why is data compression useful in machine learning?

Is data compression always a good thing? Can you use (PCA) dimensionality reduction to get fewer features, and therefore make your model less likely to overfit?

Exercise 4.

Machine learning pipelines – examples. Steps to go through (in real life).

In groups of 2-3.

(in very, very broad terms):

- What kind of steps to you think we would need in a machine learning pipeline (given pictures, and ?) to get a car to learn how to stay on the road (say with markings/stripes on the road to indicate the sides of the road).

- Use your imagination: What kind of steps (in very, very broad steps) do you think an ML system should go through to learn the car system to read signs near the road? To spot people on the road?

Exercise 5.

In groups of 2-3.

K-Means recap exercise:

A major challenge in unsupervised learning is evaluating whether the algorithm has learned something useful.

Usually we apply unsupervised learning to data that doesn't contain any labels, so we don't know what the right output should be. Therefore it can be somewhat tricky to tell whether a model is "good". Often, the only way to evaluate is to inspect manually.

Luckily, the iris dataset is relatively wellknown, and therefore well suited to test unsupervised learning algorithms on – and improve our intuitions about these unsupervised learning algorithms.

See:

https://en.wikipedia.org/wiki/Iris_flower_data_set

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor), where we are given length and width of sepals and petals.

Using the file tutorial1kmeans.py (day 4) as inspiration, here we will return to the unsupervised algorithm kmeans, and try that out on the iris dataset.

Try to build clusters of sizes 2,3,4 and 5 on ***both sepal and petal datasets***.

Question:

Look at the code, iris_kmeans.py, on Canvas.

This code is initialized to k=2 and petal-length and -width, which you must manually adjust to the other values.

a) Adjust k to better value.
b) Adjust to the sepal dataset, what do you find? What k-value works?

I.e.

Build clusters for both sepal and petal datasets.

What do you find? Number of clusters for best fit?

Exercise 6.

In groups 2-3.

For a lot of machine learning applications it helps to be able to visualize your data.

The Iris dataset is 4 dimensional, and therefore not so easy to visualize and completely grasp….

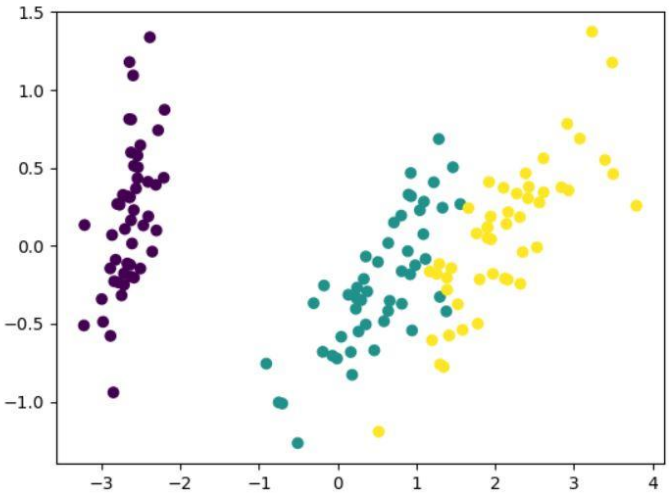
But you can use PCA to reduce the 4 dimensional data into 2 or 3 dimensions so that you can plot and hopefully understand the data better.

Download iris_kmeans_pca.py from Canvas.

- Walk through the steps of this implementation of Principal Component Analysis (PCA), and make sure that you understand the code…
- Explain to each other what is happening in the code…

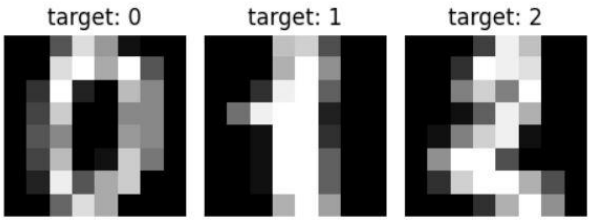
- Make small changes to the code, in order to improve your understanding.

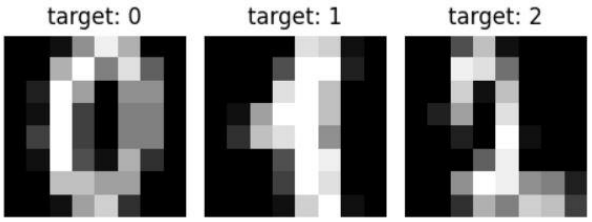
- Experiment with the number of clusters that kmeans looks for. Are the results surprising? What number of clusters give the best result?



Exercise 7.

In groups 2-3,





In this exercise we will try to classify the sklearn “load_digits“-dataset using a random forest classifer.

- See the code in Week8_mnist.py. Make sure that you understand the code.

- Then try to improve the accuracy, still using the random forest classifier. How good can you get?

- Looking at the confusion matrix, which number ist he hardest to recognize correctly?

- Try another technique. Try with a decision Tree! What is the best accuracy you can get?