Project descriptions and tasks in classification.

Fagansvarlig: Magne H. Johnsen

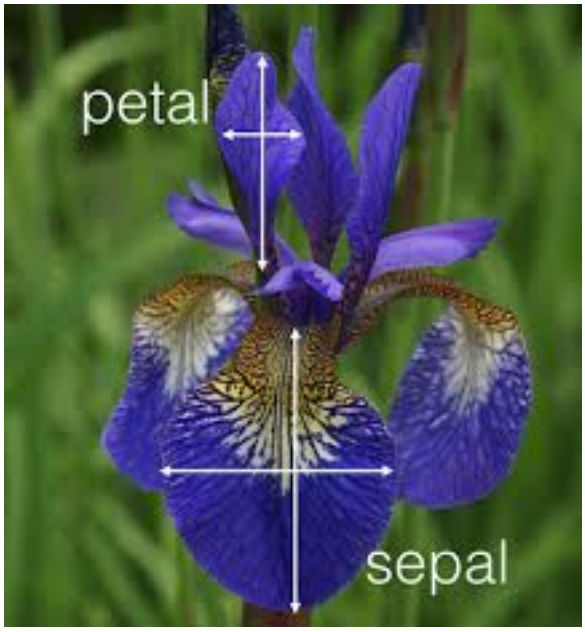
27. februar 2018

All students will be organized into groups of two. The Iris task is to be done by all groups which choose a classification task. The groups shall in addition choose one of the tasks "vowels" or "handwritten numbers". All experiments shall be implemented in Matlab or other preferred languages (as Python).

1 The Iris task

The Iris flower has several variants including three called respectively Setosa, Versicolor and Virginica, see figure 1. The flower has both large (Sepal) and små ll (Petal) leaves, see figure 2. The three mentioned variants can be discriminated by the different lenghts and widths of the petal and sepal; i.e. these four measurements are a logical choice for input features.
A database (often called "Fisher Iris data") is produced, consisting of 50 examples of each of the three variants/classes . We refer to https://en.wikipedia.org/wiki/Iris_flower_data_set for a more detailed description of the database.
The Iris task is one of a few practical problems which are close to linearly separable. Thus an error free linear classifier can be designed for the database. This first part of the project therefore has focus on the design and evaluation of a linear classifier. Further one should analyze the relative importance of each of the four feature with respect to linear separability.

Figur 1: The three Iris variants



Figur 2: Length and width for repectively petal og sepal

The task consists of two parts

1. The first part has focus on design/training and generalization.
   (a) Choose the first 30 samples for training and the last 20 samples for testing.
   (b) Train a linear classifier as described in subchapter 2.4 and 3.2. Tune the step factor α in equation 19 until the training converge.
   (c) Find the confusion matrix and the error rate for both the training and the test set.
   (d) Now use the last 30 samples for training and the first 20 samples for test. Repeat the training and test phases for this case.
   (e) Compare the results for the two cases and comment
2. The second part has focus on features and linear separability. In this part the first 30 samples are used for training and the last 20 samples for test.
   (a) Produce histograms for each feature and class. Take away the feature which shows most overlap between the classes. Train and test a classifier with the remaining three features.
   (b) Repeat the experiment above with respectively two and one features.
   (c) Compare the confusion matrixes and the error rates for the four experiments. Comment on the property of the features with respect to linear separability both as a whole and for the three separate classes.