Work Sheet 7

Classification

Exercise 1

Exploring more exercise data.

a) The file `Examples.zip` contains the three files `Example1.csv`, `Example2.csv` and `Example3.csv`. These are sample sets. Load each of the `.csv` files.

Each *column* contains one sample of the sample set. The first row contains the class number Ω_κ. Rows two and three contain the two-dimensional feature vectors $\boldsymbol{c}$.

For each file, plot all the features with different colors for the classes.

b) A very famous sample set is *Fisher's Iris data set* [2]. The data set consists of 50 samples from each of the three flower species Iris Setosa, Iris Virginica and Iris Versicolor. Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

The file `iris-numclass.csv` (in zip file `FisherIris.zip`) contains all 150 samples of the sample set. Here each *row* contains one sample. The class number in column 1 contains the species (Iris setosa, Iris virginica and Iris versicolor). The columns 2-5 contain the measurements of each sample (sepal length, sepal width, petal length, petal width).

Plot all the features (maybe in different combinations of 2 features) with different colors for the classes.

Exercise 2

The following excercise is based on the large data set from [1], available in the file `anthrokids.csv`. It contains many body measurements from children and adolescents.

a) Split the data set into two sets: one for male and one for female persons. We will only use the columns for the age as classes Ω_κ and the height measurements as feature c.

b) Split the male and the female datasets into training and test data sets. A ratio of 2:1 is suitable.

c) Use the training data sets to estimate the parameters for all classes: Estimate the parameters of normal distributions $p(c|\Omega_\kappa)$ for different ages (classes Ω_κ) given the features c (height measurements) of this class. Calculate the prior probabilities $p(\Omega_\kappa)$ of all classes. You might want to try with classes for ages of $3, 4, \ldots, 18$.

d) Go through all the data in the testing data sets and classifiy them with the optimal Bayes classifier:

- Use the feature c (height measurement) to evaluate the normal distributions $p(c|\Omega_\kappa)$ of all possible classes.
- Make a decision for the class Ω_κ that maximizes the posterior $p(\Omega_\kappa|c)$.
- Check if the decision is correct using the known real age.

e) At the end calculate the overall recognition rates for the male and female data sets. Are you satisfied with the result?

Exercise 3

The following excercise is based on the Iris data and the data from `Examples.zip`.

a) Split each data set into training and test data.

b) Implement a Bayes classifier that estimates a two-dimensional normal distribution for each class for the data sets from `Examples.zip`. Use them to classify the test samples. Are you satisfied with the classification results?

c) Implement a Bayes classifier using normal distributions for the Iris data set. Vary the used features from using all 4 measurements per sample down to using only 1 of the measurements. Which of the possible feature combinations perform best and which worst? What conclusions do you draw from the results?

d) Look for ways to visualize how your classifiers work in the case of two-dimensional features. The important thing here is: Which areas of the feature space are assigned to which class?

Exercise 4

We will try to improve the classification results from Exercise 3.

The critical point with the previous classifier is the choice of the underlying density function. If a unimodal density does not produce satisfying results, one can turn to mixture distributions.

Replace the classifier from Exercise 3 with one that estimates a mixture distribution in training. What results do you get with this classifier? How do you choose the number of mixture components?

Exercise 5

Comparing classifier performance.

a) Implement a classifier based on the idea of Parzen estimation.

For this purpose, realize a function to compute $p(\boldsymbol{c}|\Omega_\kappa)$ according to the idea of a Parzen estimation (kernel density estimation). Your function should take as input a classified sample set, the class κ to be evaluated, the covariance matrix $\boldsymbol{\Sigma}$ and of course the feature vector $\boldsymbol{c}$.

You can use $p(\boldsymbol{c}|\Omega_\kappa)$ and $p(\Omega_\kappa)$ to perform a Bayesian classification.

b) Implement a Nearest Neighbor classifier.

Write a function that takes a classified sample set, a parameter m, and the feature vector $\boldsymbol{c}$. Your function should now search for the m nearest neighbors within the sample set to the feature vector $\boldsymbol{c}$. Use this to calculate the probabilities $p(\Omega_\kappa|\boldsymbol{c}) = \frac{m_\kappa}{m}$ by counting within these neighbors the memberships m_κ to each class.

You can use $p(\Omega_\kappa|\boldsymbol{c})$ to perform a Bayesian classification.

c) You have previously examined the data sets from *Fisher's Iris data set* [2] and from `Examples.zip`. Now use these to evaluate the performance of different classification methods. Compare the achievable classification rates for the following classifiers:

- Simple normal distribution classifier with a unimodal, multivariate normal distribution as the underlying density. This corresponds in essence to the classification exercises for the Kids' Size Problem.
- Classification based on a Parzen estimate.
- Classification with the Nearest Neighbor classifier.
- If you have other classifiers available, feel free to include and compare their results.

Which classification methods would you choose for a practical application depending on the data set?

d) Take the entire *Fisher's Iris data set* [2] and ignore the membership of the samples to the different flower species.

Use your EM algorithm to estimate a mixture distribution with 3 components from this data. How do the individual mixture components relate to the species data? For visualization, you can limit yourself to 2 dimensions (sepal length/width or petal length/width).

References

[1] STAT 3202: Group Project I. `https://daviddalpiaz.github.io/stat3202-au18/project/proj-01/proj-01-E.html`, Autumn 2018, OSU.

[2] Ronald Aylmer Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.