Universität Bielefeld

CITEC / Faculty of Technology
Ambient Intelligence Group
Dr. Thomas Hermann

Sekretariat: S. Strunk
Office: CITEC-3.311
106-6891

Exercises of the lecture "Introduction to Data Mining"

WS 2018/2019 Exercise sheet 1

- Introduction -

The GZI offers Python and SciPy Tools, which you can use in order to solve the following tasks. Python is a mighty, yet easy to use programming language. Also it is very suitable to implement komplex processes fast and comprehensibly. Scipy is a scientific library for Python.

More information as well as a diversity of introductions examples and tutorials on Python and SciPy can be found at http://www.python.org (http://www.python.org) and http://www.scipy.org (http://www.scipy.org).

If you have problems solving Problems 1.1 and 1.2 or are programming in Python for the first time, please consider doing the additional problems first. These are not relevant for the exam though.

Problem 1.1, Data representation

In the data mining and machine learning communities datasets are commonly represented as matrices. The rows of the matrix usually correspond to the data points and the columns correspond to the dimensions. If available the first column contains a class label.

Your task is to create a normally distributed two-dimensional dataset of 100 data points, where the first dimension is defined by

$$\mu_1 = 0, \quad \sigma_1 = 1$$

and the second dimension by

$$\mu_2 = 0.1, \quad \sigma_2 = 1$$

```
In [ ]:
```

Problem 1.2, Student-t Test

Implement the t-test in form of a python-function.

a) Test the *null hypothesis*, that the *mean values* of the underlying distributions of the example data created in Problem 1.1 are equal. Can you reject the null hypothesis with a significance of 0.05? How about a significance of 0.01?

b) Create two samples X_A and X_B of $p(x) = \mathcal{N}(x; 0, 1)$ each including 10 examples and calculate

$$t = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sigma_{err}}.$$

Repeat this experiment N times (with $N > 1000$) and plot a histogram of the set of the calculated t with $\sqrt{N}$ bins. What can be observed?

For the solution of this task use `pylab.hist`!

In [ ]:

```
%pylab inline
```

Additional Problem 1.1, Scalar Product

Implement a function `scalar_prod(x, y)` which calculates the scalar product of two arrays.

In [ ]:

Additional Problem 1.2, Gaussian density function

Implement a function `gauss_df(x, mue, sigma)` that calculates a normal distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Also plot the results using `pylab.plot()`

In [ ]:

Additional Problem 1.3, Iris-Dataset

Load the Iris-Dataset[1] using Python and create a scatter plot of it with `pylab.plot()`
[1] You can find the Iris-Dataset `iris.txt` in the e-Learning space (Lernraum) of the ekVV.

In [ ]:

Thomas Hermann (thermann@techfak)	Lecture Wed 12-14, CITEC lecture hall	office building	public transportation
Ferdinand Schlatt	Tutorial Tue 12:15-13:45, H11	Universität Bielefeld	light rail 4
		Universitätsstraße 25	from Bahnhof and Jahnplatz
		33615 Bielefeld	

In [ ]: