

Data Mining - Python

Exercise 1: Simple Preprocessing and Visualization

1.1. Python Installation

Download Anaconda and install the software on your laptop. Additional information on how to install Anaconda can be found here:

<https://docs.continuum.io/anaconda/install/>

1.2. Load and Preprocess the Students Dataset

Import the *students* data set into Python with pandas. The *students* data set is provided in ILIAS as an Excel file. Use the `read_excel` function.

1. What is the most common mark that has been given in FSS2010? To find the answer filter the examples and draw a histogram afterwards. Use the plot functionality of pandas.
2. Is there a correlation between the mark and the number of attended classes? Find the answer using a scatter plot.
3. Does this correlation hold for all students? Find the answer by aggregating the examples by student and use a scatter plot afterwards.

1.3. Visual Exploration of the Iris Dataset

The data set describes three types of Iris flowers:

- Setosa
- Virginica
- Versicolour

There are four (non-class) attributes

- Sepal width and length
- Petal width and length

Retrieve the Iris data set from the ILIAS group and put it in the folder where you started the notebook. Use different plotters to visualize and explore the data set.

1. Which attribute combination and (approximate) value ranges determine the type of Iris flower?



Answer:

Type of Iris Flower	Attribute combination and value ranges
Setosa	
Virginica	
Versicolour	