Workshop 1.2

Advanced Analytics and Applications [AAA]

Information Systems for Sustainable Society (is3) | WiSo Faculty | Univ.-Prof. Dr. Wolfgang Ketter | 4/12/23

I Multiple Choice

II Calculation

III Programming

Multiple Choice Questions

Question 1:

Astronomers have been cataloguing distant objects in the sky using long-exposure CCD images. The objects need to be labeled as star, galaxy, nebula etc. The data is highly noisy, and the images are very faint. The cataloguing can take decades to complete. Which Method would you recommend the physicists use to automate the cataloguing process, and improve its effectiveness?

- ☐ Clustering
- ☐ Classification
- ☐ Regression

Multiple Choice Questions

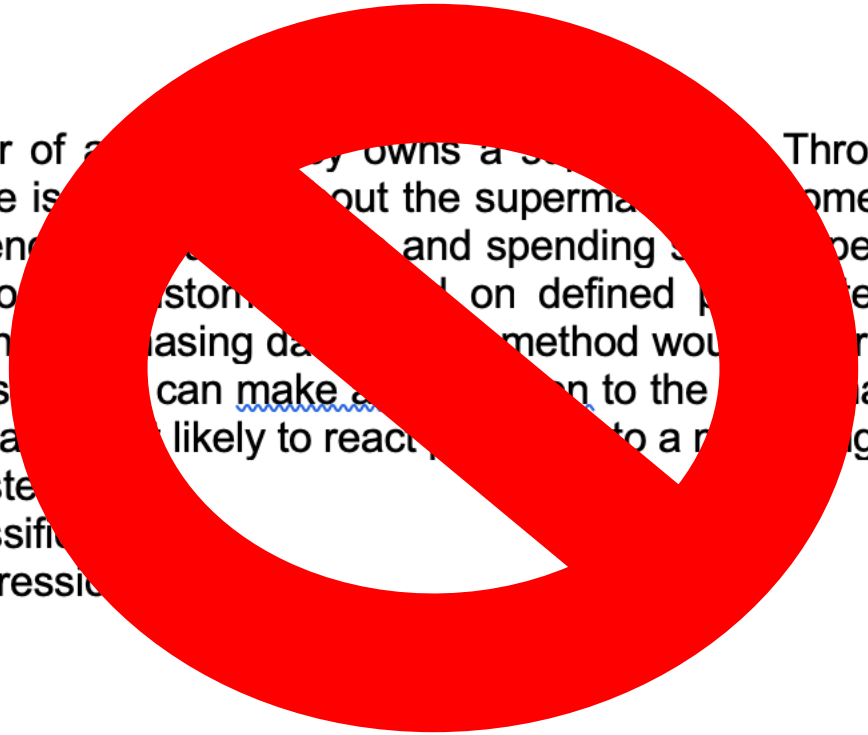
Question 1:

Astronomers have been cataloguing distant objects in the sky using long-exposure CCD images. The objects need to be labeled as star, galaxy, nebula etc. The data is highly noisy, and the images are very faint. The cataloguing can take decades to complete. Which Method would you recommend the physicists use to automate the cataloguing process, and improve its effectiveness?

- ☐ Clustering
- ☐ Classification
- ☐ Regression

Multiple Choice Questions

Question 2:



A customer of a … owns a … Through membership cards, there is … out the superma… …omers like Customer ID, age, gen… … and spending … …pending Score was assigned to … …ston… … on defined … …ers like customer behavior an… …asing da… …method wou… …recommend using so the cons… … can make a… …n to the … …arket which of his customers a… … likely to react … …o a r… …g campaign?

- ☐ Cluste…
- ☐ Classifi…
- ☐ Regressi…

Question 3:

A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company. Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away. Which method would you recommend using to model the probability of attrition where the results can be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay?

- ☐ Clustering
- ☐ Classification
- ☐ Regression

Question 3:

A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company. Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away. Which method would you recommend using to model the probability of attrition where the results can be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay?

- ☐ Clustering
- ☐ Classification
- ☐ Regression

I Multiple Choice

II Calculation

III Programming

Calculation

Question 1: MSE

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8			
2	4	3,4			
3	5	4			
4	4	4,6			
5	5	5,2			

Calculation

Question 1: MSE

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8	4	1,44	0,64
2	4	3,4	0	0,36	0,36
3	5	4	1	0	1
4	4	4,6	0	0,36	0,36
5	5	5,2	1	1,44	0,04

Calculation

Question 1: MSE

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8	4	1,44	0,64
2	4	3,4	0	0,36	0,36
3	5	4	1	0	1
4	4	4,6	0	0,36	0,36
5	5	5,2	1	1,44	0,04

a) Calculate MSE using the values from the above table (y = true value, $\hat{y}$ = predicted value)

Calculation

Question 1: MSE

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8	4	1,44	0,64
2	4	3,4	0	0,36	0,36
3	5	4	1	0	1
4	4	4,6	0	0,36	0,36
5	5	5,2	1	1,44	0,04

a) Calculate MSE using the values from the above table (y = true value, $\hat{y}$ = predicted value)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{2{,}4}{5} = 0{,}48$$

Calculation

Question 1: MSE

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8	4	1,44	0,64
2	4	3,4	0	0,36	0,36
3	5	4	1	0	1
4	4	4,6	0	0,36	0,36
5	5	5,2	1	1,44	0,04

b) When is MSE useful?

MSE can represent the difference between the actual observations and the observation values predicted by the model

Puts large emphasis on large deviations

Calculation

Question 2: R^2

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8	4	1,44	0,64
2	4	3,4	0	0,36	0,36
3	5	4	1	0	1
4	4	4,6	0	0,36	0,36
5	5	5,2	1	1,44	0,04

a) Calculate R-Squared using the values from the above table (y = true value, $\hat{y}$ = predicted value)

Calculation

Question 2: R^2

x	y	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
1	2	2,8	4	1,44	0,64
2	4	3,4	0	0,36	0,36
3	5	4	1	0	1
4	4	4,6	0	0,36	0,36
5	5	5,2	1	1,44	0,04

a) Calculate R-Squared using the values from the above table (y = true value, $\hat{y}$ = predicted value)

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{2.4}{6}$$
$$= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{3,6}{6} = 0.6$$

Question 2: R^2

b) When is R-Squared useful? Compare it to MSE.

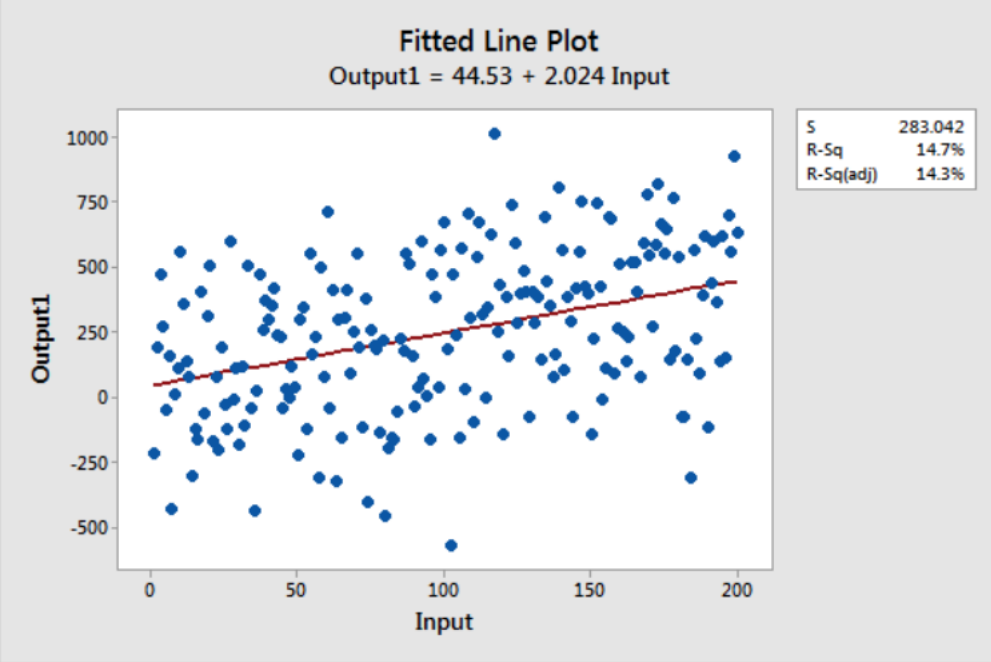
Question 2: R^2

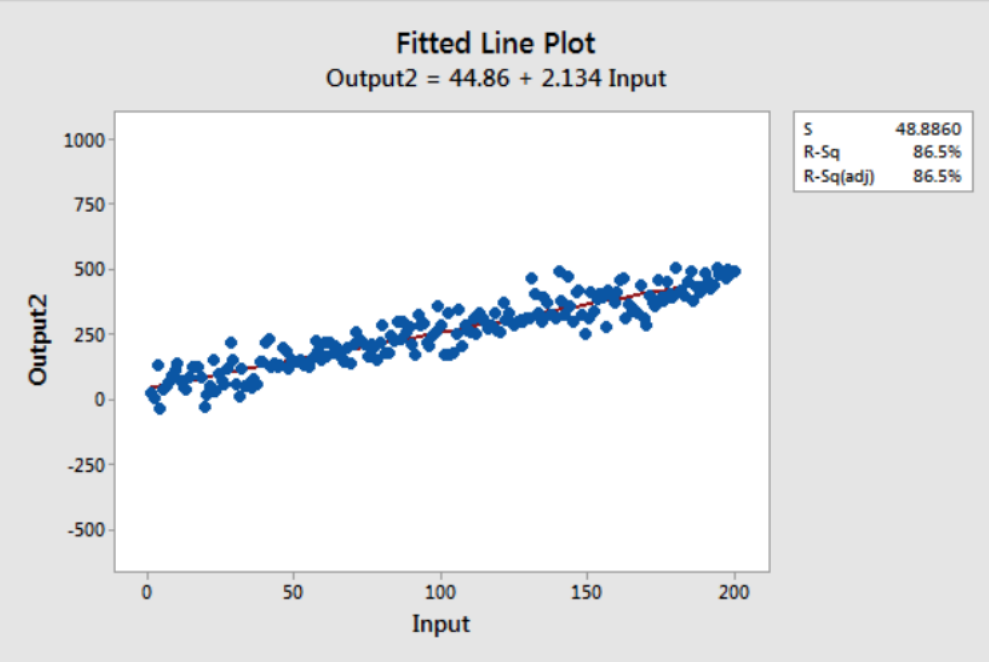
b) When is R-Squared useful? Compare it to MSE.

R-squared represents the fraction of variance of response variable captured by the regression model rather than the MSE which captures the residual error.

Calculation

Question 2: R^2





Universität zu Köln

I Multiple Choice

II Calculation

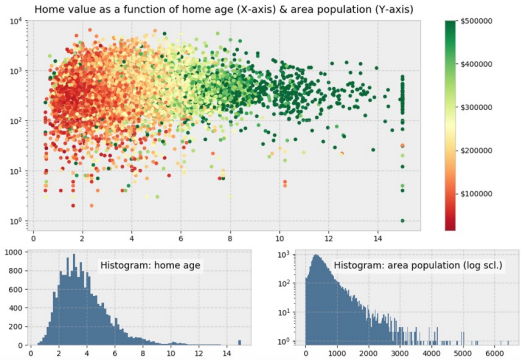
III Programming

Question 3: Working with Pandas

a) (Pandas) Create a python notebook that loads the *employee_survey_data.csv* into a Data Frame and find out the following stats:
   (1) Dimensionality: _____
   (2) Sum of rows with missing values: _____
   (3) Number of employees most satisfied with their jobs: _____
   Tip: Find out max value. Filter employees by max value. Sum number of employees with max job satisfaction.

Calculation

Question 3: Plotting with MatplotLib



(Plotting) In this task, we want to learn how create a plot using the python library matplotlib with multiple axes in one figure. To do so, create a Jupyter Notebook, then download and initialize the California housing data set.

Then, try to render a plot similar to the one depicted in Figure 1. In the first figure, we can want to render a scatter plot of “age” (Home Age) and “pop” (Population). In the second plot, we want to visualize a histogram of “age” (Home Age). Finally, we want to render a plot a histogram of the “pop” (Population). Follow the template from Figure 2 with 3 axis objects, and by using subplots (for more information, read the docs for matplotlib)

Matplotlib: Motivation for this workshop

Extensive documentation – The documentation of matplotlib is extensive and in detail BUT also hard for beginners.

Digging through the code is tedious and time-consuming (+70.000 lines).

Thus, we will cover important concept of matplotlib without commencing an "Indiana Jones" like adventure through the codebase.

Source: https://matplotlib.org/3.2.1/contents.html

Matplotlib: History

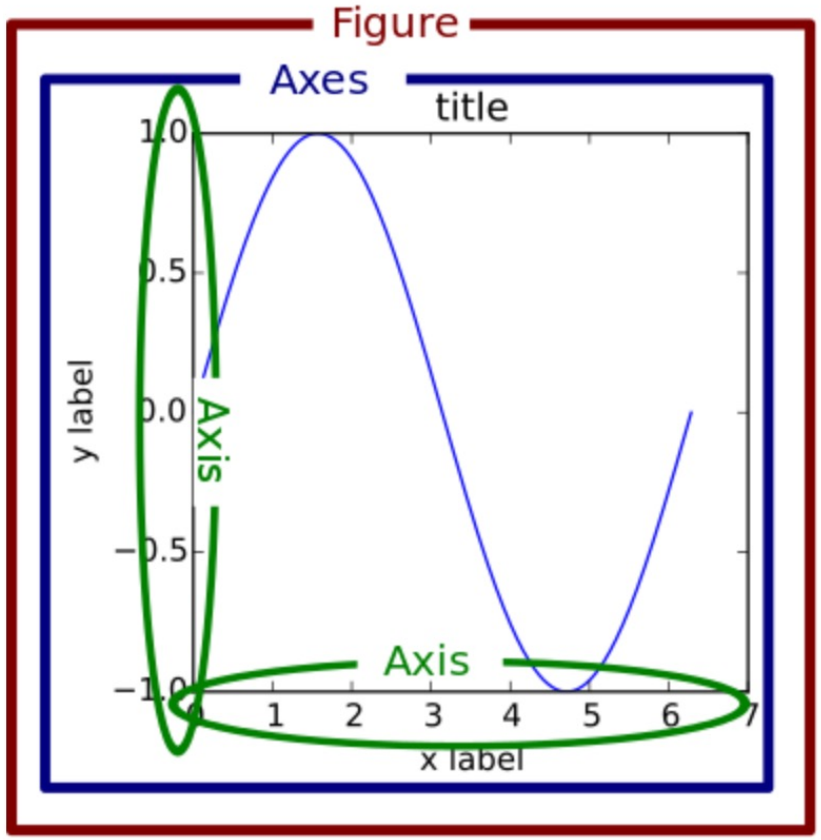
John D. Hunter (1968 – 2012) was an American neurobiologist and the original author of matplotlib.

Source: https://www.anaconda.com/distribution/

Matplotlib: The object hierarchy

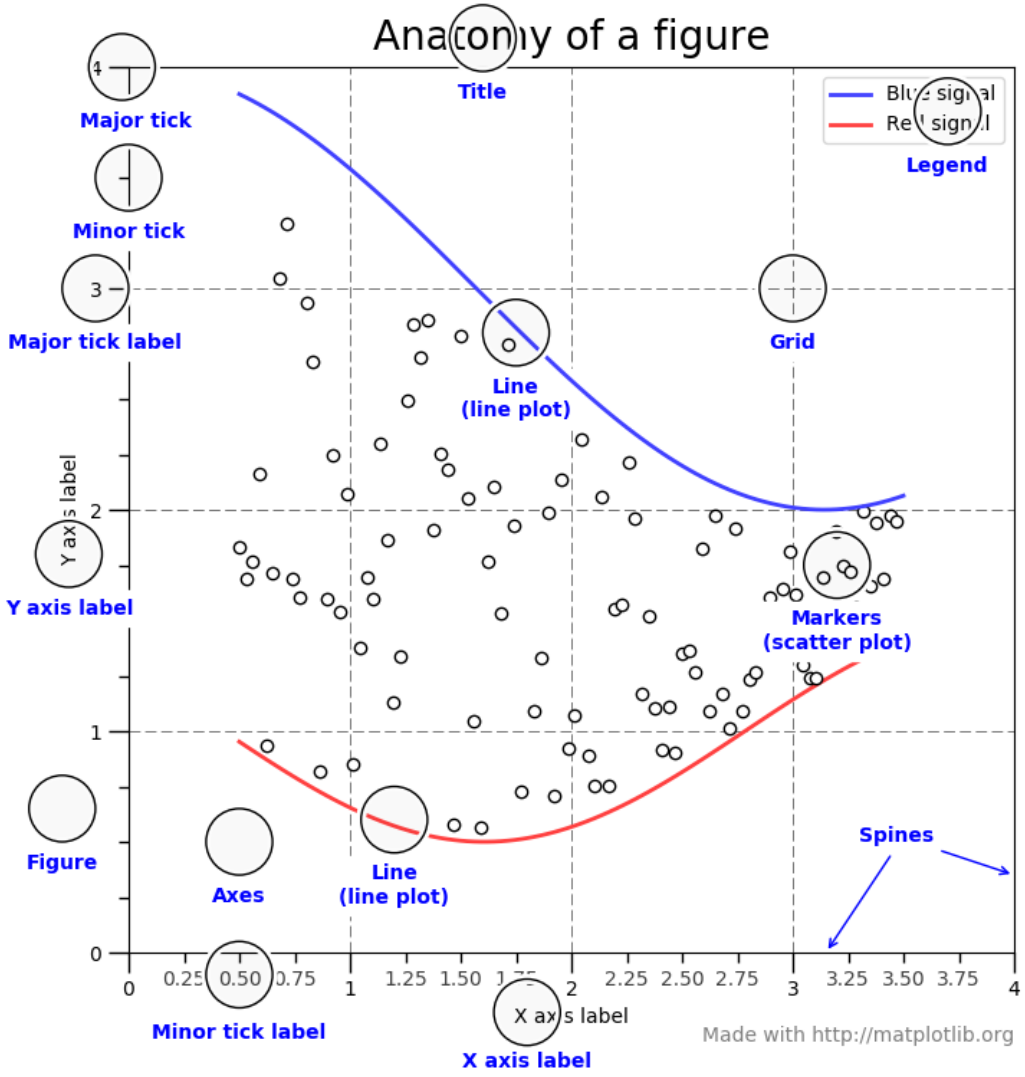
```
plt.plot([1, 2, 3])
```

Create tree-like object hierarchy



Source: https://alchetron.com/John-D-Hunter /

Matplotlib: Object hierarchy under the hood



Source: https://matplotlib.org/examples/showcase/anatomy.html

Matplotlib: Stateful versus Stateless Approach

```
import matplotlib.pyplot as plt
```

```
plt.plot([1, 2, 3])
plt.title(„Hello World“)
plt.ylabels(„Emissions)
```

Seems like a stateful "enviroment“ but in reality it just mimics stateful behavior.

Source: https://matplotlib.org/examples/showcase/anatomy.html

Matplotlib: Stateful versus Stateless Approach

Python >>>

```
# matplotlib/pyplot.py
>>> def plot(*args, **kwargs):
...     """An abridged version of plt.plot()."""
...     ax = plt.gca()
...     return ax.plot(*args, **kwargs)

>>> def gca(**kwargs):
...     """Get the current Axes of the current Figure."""
...     return plt.gcf().gca(**kwargs)
```

Applies for almost any function in pyplot

https://realpython.com/python-matplotlib-guide/

Matplotlib: Stateful versus Stateless Approach

Bottom line: Matplotlib interface is object-oriented and thus is built upon the stateless approach

https://realpython.com/python-matplotlib-guide/

Contact

For general questions and enquiries on **research**, **teaching**, **job openings** and new **projects** refer to our website at www.is3.uni-koeln.de

For specific enquiries regarding this course contact us by sending an email to the **IS3 teaching** address at is3-teaching@wiso.uni-koeln.de