Artificial Intelligence and Data Analytics for Engineers

Exercise 3

Introduction

Solving exercises is not mandatory, and exercises will not be graded or corrected by the lecture team. However, we strongly advise you to do as many exercises as possible to prepare for the final exam. Some of the tasks will be discussed with a presentation of the solutions during the exercise session on Thursday. If you want to do exercises at home, we suggest installing Anaconda (https://www.anaconda.com/distribution/ - you will want to download the Python 3.7 version).

In case you have any questions, feel free to send us an e-Mail to: aidae@ima-ifu.rwth-aachen.de.

Task 0) Setting up your Python environment

Since you created a Python environment last week, we only need to enable the environment from now on. This needs to be done **every time** you open a new terminal, for instance after a system restart or if you accidentally closed the Terminal. Open a terminal and enter:

```
source activate aidae
```

Now you are ready to start programming in Python.

Task 1) Initial Data exploration

In this task you will explore the Combined Cycle Power Plant data set. The goal of your exploration is to get to know the fundamental properties of the data: What is the format of the data? Which relations probably exist between certain values? Is there any distribution pattern? All that will be required to make the right decisions with regard to which machine learning techniques to apply, and how to interpret the results later on. All analysis tasks will require that you load the data into a Pandas data frame using Python.

a) In the first task we will have a look at columns of the input data separately from each other. First step is to understand the meaning behind each variable and the type of data that the dataset contains. List all the variables on the dataset and verify which type of data does each column contain.
b) Following up the first step, Plot separate histograms of the data in the columns AP, RH and V using matplotlib/Pandas. Then add reference lines for a Gaussian distribution to the diagrams. What assumptions can you make regarding the underlying data from this graphical overview.
Hint: https://matplotlib.org/gallery/statistics/histogram_features.html
c) Follow up the exploration and create a scatter matrix containing all the variables. What can be seen in the resulting plot?

Task 2) Data sanity and missing values

Real world data tends to be a mess. Maybe the data was collected by some overworked secretary copying values from a piece of paper into the computer, sometimes messing up values. Maybe the data was automatically collected by a machine, which was programmed by somebody who was not as competent as one would hope, and thus the data contains insensible values caused by a bug. Or maybe the collection method of your data just inherently leads to missing values sometimes. For further processing, you want your data to be

consistent with regard to the format and completeness, to avoid having to deal with inconsistencies in every further step.

a) The data contains the columns V, AP, and RH, are supposed to be numerical values related to different physical variables. Verify on each column if there are missing or insensible values. Act accordingly and filter said values.
b) The data contains the column PE which is supposed to be the Power output of a combined cycle power plant. When obtaining this data, the end goals is to try to predict the output of the power plant based on the rest of the working parameters mentioned in the dataset. Some times in the field, people misunderstand orders and end up logging incorrect data. Filter the categorical data, reformat and plot the resulting values. Are there values that seem insensible? Remove the rows containing this insensible data.
c) The column AT stands for ambient temperature. The person in charge of obtaining this variable tried to use an OCR system over one of the older sensors existing in the plant. This led to undesired results and a mix between the letter "o" and the number 0. Please correct said error and transform the column into floats. Subsequently, verify the sanity of the column and visualize the results.

Task 3) further exploration and outliers

One of the most important tasks previous to the data analysis is the preparation. This is an iterative process that sometimes has unexpected challenges. Not all the data preparation steps are standard and some of them are case specific. Thus, we must return to visualize the data to better understand its state.

a) Now we want to have a look whether some of the columns are related to each other. While there are statistical tests to verify specific relations, we just want to get an overview right now. For this we will again visualize the data. Create two matplotlib scatter plots: one with the data of columns PE and AT and another plot with the data of columns V and PE. Put the data of columns AT and V on the x-axis, while the data of PE is visualized on the y-axis. Which relation between AT and PE (and V and PE) would you assume, if any? If you would be able to predict the values of PE, what would that mean regarding the corresponding values in AT and V?
b) In certain scenarios it might be clever to split the data set and have a look at both subsets independently, to find properties that hold for the subset, but not for the entire data set. In our case, it makes sense to split depending on column TCN. Using Pandas, split the data set according to the value of TCN, which is the technician who acquired the data. You should end up with multiple separate data frames, each containing the data of one specific technician. Now have a look at the correlation between columns AT and PE by creating a scatter plot for each of the sub data sets. What does that tell you regarding the influence of the criterion? What could have happened?
c) Once the proper transforms have been done, you can see that there are still multiple outliers in the dataset. Use the scikit-learn implementation of Isolation Forest to separate the anomalies. Visualize them and verify if they must be removed or not.

Task 4) Encoding

Always remember that private information in datasets is a high security risk. The wrong management of said information can generate undesired liabilities.

a) Encode the names of the technicians that have participated in the data acquisitions.