exercise_02

November 20, 2023

1 Exercise 2

1.1 General

a. Define Knowledge Discovery in Databases!

[ ]:

b. What is the difference between data mining and knowledge discovery (based on the definitions used in the lecture slides)?

[ ]:

c. List four typical methods applied in the context of KDD applications and briefly describe them!

[ ]:

d. Welche Geschäftsziele werden typischerweise durch KDD unterstützt?

[ ]:

e. Nennen und erklären Sie kurz die drei Charakterisitiken für Daten mit der [Gartner 2012] Big Data definiert!

[ ]:

1.2 CRISP-DM

a. Nennen Sie die sechs Phasen der CRISP-DM Methodologie und beschreiben Sie stichpunktartig, was dort geschieht!

[ ]:

b. Wie hängt die Vorverarbeitungsphase konzeptuell mit den anderen Phasen zusammen?

[ ]:

c. Was sollte in der Evaluation-Phase erfolgen?

[ ]:

1.3 Regression: House Prices (1)

In this set of exercises, we will work on a task that aims to predict housing prices from variables describing that describe those homes. The task is based on a Kaggle Competition. Here is a more detailed description of the task:

> Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.
>
> With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

A detailed description of the data can be found on Kaggle. Additional information is included in the file `data_description.txt`.

1.3.1 Load the "House Prices" dataset using `pandas`

1) Use the function `read_csv` from the `pandas` library to read the CSV file `train.csv`.
2) Make sure the `Id` column is set as the index (see `index_col`).
3) Store the data in the variable `data_houses`.

[ ]:

1.3.2 Try to fit a model

1) Define input variables `X` and the output variable `y`.
2) Try to fit a `LinearRegression` model from scikit-learn.
3) Why did fitting the model not work?
4) What should be done before trying to fit a model.

[ ]:

1.3.3 Preliminary data inspection

1. Which variables are numbers? (**hint**: use `select_dtypes` and `numpy`'s data type `number`)
2. Which variables contain empty values? How many NAs are there for each variable? (**hint**: you can use the `pandas` function `isna`)
3. How many variables (in percent) do we loose if we only use numeric variables?
4. How many variables (in percent) do we loose if we drop all variables that have NAs?
5. How many samples do we loose (in percent) if we drop all samples that contain NAs?
6. How many variables do we have left if we only use numeric variables and drop all variables with NAs?

[ ]:

1.3.4 Try to fit the model again

1. Select only numeric variables and drop all variables with NAs
2. Try to fit the model on this "clean" data

3. Predict the house prices based on the cleaned data
4. Calculate the mean and standard devitation of the absolute difference of your predictions to the real prices

[ ]:

1.3.5 Optimize

1. Try different regularization methods (ridge, or Lasso).
2. Try normalizing the features (use `StandardScaler` and `make_pipeline`).
3. Do the result change?

[ ]:

1.3.6 Introspection

1. Discuss whether the predictions you had are good or bad.

[ ]:

2. Do you think, our evaluation of the model is OK?

[ ]:

2. Discuss whether dropping features, as we did is a good idea or not.

[ ]:

3. Look at the data again (and its description), and check whether including the variables as numeric input features makes sense.

[ ]:

4. What should we have done before fitting a model?

[ ]:

5. What would be our next steps?

[ ]:

1.3.7 BONUS

Try to optimize your predictions.