Due date: Wednesday, November 10, 2021

1 Introduction to python

We have compiled a set of very basic python functionalities in a jupyter notebook that you can use to familiarize yourself with both python and jupyter notebooks. To go through the tutorial, start `jupyter-notebook` and open the file `python-intro.ipynb`.

2 Linear regression with scikit-learn (10 pt)

The jupyter notebook `linear-regresssion.ipynb` contains an example of a least square linear regression using `scikit-learn`. The steps include:

1. creating a dataset
2. randomly dividing the dataset into a test and training set
3. fitting using the training set
4. visualization
5. evaluation of the predictive power

Go through the notebook and familiarize yourself with the different steps.

2.1 Task 1 – Random numbers

Random numbers are used to create the dataset and to divide the dataset into a test and training set. Explain what the functions `np.random.randn(npoints)` and the class `sklearn.model_selection.ShuffleSplit` do and how are they used in the workflow.

2.2 Task 2 – Model

Write down the linear model used in the regression. How are the parameters that correspond to the weights retrieved in the python model? Which python function is used to make predictions with the fitted model?

2.3 Task 3 – Coefficient of determination

The function `score` returns the coefficient of determination R^2. Write down the equation to compute R^2. What does a value of $R^2 = 1$ mean? What does a value of $R^2 = 0$ mean?

2.4 Task 4 – Dataset

Create 3 different datasets with $d_{\max} = \{1, 10, 20\}$ and fit a linear regression model for each of these datasets. Compute R^2, MAE, and MSE for each fit. How do the values change as a function of $d_{\max}$? Explain the observed trend.

3 Diabetes dataset (10 pt)

Scikit-learn provides a number of toy datasets to work with. The Diabetes dataset contains 442 samples where the disease progression is recorded as a function of 10 different features. Details about the dataset can be found at https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset. An example of how the dataset is loaded is given in the jupyter notebook `linear-regresssion.ipynb`.

3.1 Task 1 – Fitting

For each of the 10 attributes, perform the 5 steps outlined in section 2 and report your results. (The 'creation' of the dataset in this case corresponds to loading the respective data). Feel free do write a loop over the 10 attributes. Make sure the axes of the plots are properly labeled.

3.2 Task 2 – Data preparation

The values of the 10 features have been centered around the mean and scaled by the standard deviation multiplied by the number of samples. Why have the data been pre-processed in this way?

3.3 Task 3 – Model prediction

How well does the linear regression model work for each of the 10 features? Which of the attributes can be reasonably fitted with a linear model? For which is the linear model not appropriate? Provide a discussion of your results and give reasons for your conclusions.