In-Class Assignment 01

Machine Learning, Summer term 2018

Norman Hendrich, Marc Bestmann, Philipp Ruppel

April 9, 2018

Assignment 01.1 (Getting Started, 0 points)

To get started: Please get Python working on your system, and step through the tutorials for Numpy and Matplotlib as needed.

Assignment 01.2 (Webserver Statistics, 0 points)

The file *traffic_per_hour.csv* (linked on the webpage) contains (faked) access statistics (hits per hour) for some company webserver.

a. Try to read and parse the file with Pyhton. There are multiple options, for example the CSV reader module, the Numpy method *genfromtext*, or methods in Pandas. Note that the columns in the file are separated by tab characters, not commas.

b. The file includes some invalid (nan, not a number) entries. Use Numpy indexing magic or Pandas to remove these from the data.

c. Plot the raw data using matplotlib.scatter() and add title and axis labels to your plot.

d. For understanding the inductive bias: Use Numpy *polyfit* to fit linear, quadratic and polynoms of higher order to the data, and plot these, too. The companion method *poly1d* directly works with the polynom coefficients returned from polyfit.

e. What happens with the polynom approximation when you try polynoms of higher order? Which polynom seems to approximate the given data best?

   Also try reducing the number of datapoints and observe at which point severe overfitting occurs.

f. The company wants to buy a new webserver when traffic hits 10000 accesses per hour: when will that be? Use either Numpy indexing operations to find the nearest data point in your polyfig approximation, or (better) use the Numpy root-finding functions to calculate the value exactly.