Assignment 02

Machine Learning, Summer term 2018
Norman Hendrich, Marc Bestmann, Philipp Ruppel
April 16, 2018

Solutions due by April 22

Assignment 02.1 (Numpy array operations, 0 points)

This will be done in-class. Run the exercises in `https://www.machinelearningplus.com/python/101-numpy-exercises-python/`

- Create a 1D-array with the numbers from 0 to 9.
- Create a 3×3 array of Boolean values (all True's).
- Extract all odd numbers from an array.
- Replace all even numbers in a numpy array with the value -1.
- Generate a diagonal matrix of size $n \times n$.
- ...

Assignment 02.2 (Linear Algebra, Hilbert Matrix, 10 points)

In this exercise, you will look how to solve a system of linear equations using numpy. This is usually a fast and efficient operation, but there are special ill-conditioned matrices that make the task more interesting.

a. Write your own Python function to generate the Hilbert matrices H_k of dimension k. Using Python-style (zero-based) addressing, the matrix element in row i and column j has the value $1/(i+j+1)$. Try to use numpy array operations for best performance.

b. Calculate the rank and condition numbers for the Hilbert matrices with different k. Print those numbers for $k \in [1, 30]$.

c. Use *numpy.linalg.slove()* to solve the linear equations $H_k \cdot x = b$, where $b = (1, \ldots 1)$ is a vector of all ones. Print the solutions x_k for $k \in [1, 2, 3, 5, 10, 15, 20, 30, 50, 100]$. Check your solutions by again calculating $||H_k \cdot x_k - b||$ using numpy.

d. Do you trust the solutions?

e. In your own words, what is special about these matrices?

Assignment 02.3 (Housing Data Set, 10 points)

Download the *housing.csv* dataset from the ML webpage. The dataset contains information about houses in California as of 1990: geographical location (longitude, latitude), age of the building, number of rooms, number of bedrooms, population (number of people), number of households, medium income per household, median house value, and an annotation of ocean proximity.

a. Load the dataset into Python (e.g. using csv reader, numpy, or pandas).

b. For all data columns, find and print the minimum and maximum values, find and print the indices of those values, calculate the mean value.

c. For all data columns, calculate and plot a histogram of the data. Is the data from a normal distribution?

d. Generate a geographical map of the houses using a scatter plot based on the longitude and latitude data. Play with transparency (e.g., use alpha 0.1) for less clutter. Next, use a color scale (e.g. blue to red) to color-encode the house value in the scatter plot.

e. Split the dataset randomly into a training set (80% of the data points) and a test set (20% of the data points). Use a fixed random seed, so that your split can be reproduced later. For both sets, repeat the calculation of the minimum, maximum, and mean values. Does your test-set match the distribution in the training-set?

Assignment 02.3 (Housing Data Set, kNN, 10 points)

Try to predict housing value using the k-nearest-neighbor algorithm.

a. Write a Python function that calculates the L_2 loss for the housing value.

b. Design and implement a meaningful distance function for the housing dataset.

c. Implement the kNN algorithm, using the loss and distance functions above, and using the training- and test-sets from the previous exercise.

d. Predict the housing values for different values of k. What are your training and test errors?