Machine Learning
Jacobs University Bremen
Dr. Peter Zaspel

Course: CO22-320372
Date: 2020-02-21
Due: 2020-02-28

ML 2020 Homework 2

1 Instructions

This HW has two regular problems, and two programming problems. You may work in groups of (up to) three on these problems. Include the names of your group members on your submission. Solutions to the two written problems should be in a single PDF file. **For the programming part, you are not allowed to use any Machine Learning/Mathematical libraries unless explicitly mentioned in the question**. Your code must generate all the expected results in a single run. Please include a simple Makefile for C/C++ code. Package your PDF file and code in a zipfile and submit via Moodle. **Please note that we will only accept submission via Moodle.**

2 Problem 1 (4 Points)

In this exercise you have the freedom to choose two data sets of your own preference from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php) and do the following tasks:

- Analyze the training data and state its properties including properties of the input and output variables and number of training samples.
- Formulate a question that can be solved using machine learning on this data set and give the type of machine learning (supervises / unsupervised / regression / classification) that will allow to answer the question
- Describe the expected output.

For this exercise, you should again have a look at the examples 1.1 - 1.4 from the lecture of February 19, 2020. Try to carry out the above tasks following the general idea of these examples.

3 Problem 2 (4 Points)

In this exercise you will get familiar with the concept of mixed joint densities $\rho(x, y)$ is a two-dimensional mixed joint probability density function, where X follows the standard normal distribution, Y follows the Bernoulli distribution, with parameter p. (Note: the probability of getting the value 1 is p).

- Write down the function $\rho(x, y)$
- Find the expectation of $\rho(x, y)$

4 Programming 1 (4 Points)

In this task, you will do some self-study of visualization techniques. Familiarize yourself with one of the following tools / libraries for visualization:

- Matplotlib (Python) `https://matplotlib.org`
- Matplotlib (Wrapper for C++) `https://matplotlib-cpp.readthedocs.io/en/latest/`
- Gnuplot (stand-alone) `http://www.gnuplot.info/`

and carry out the following tasks:

- Try to find a dataset with features and labels. For example, you can use ***The Iris Dataset*** from Scikit-learn. Take first two features as well as target labels and create a 2D scatter plot where individual data points are colored by classes.
- Plot the sine wave in the interval [0, 2π].

5 Programming 2 (4 Points)

In this portion of the assignment, you will implement KNN regression. You should follow the lecture notes, understand the KNN algorithm and finally implement your own KNN regression functions. Test your functions first, using a randomly generated data following the structure described below:

- **Regresssion Data**: To train your regression algorithm, generate a dataset defined as labeled pair (x_i,y_i) with the following target vector (y) and and data vector X.

  $$a = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \end{bmatrix}$$

  $$b = \begin{bmatrix} 0 & 0.3 & 0.75 & 1 & 2 \end{bmatrix}$$

  Next generate evaluation data by sampling 1000 random points between [0,4] and predict the target value for this evaluation set using your KNN regression function.

After you have applied your KNN functions in the respective data sets, visualize your results using the appropriate visualization techniques you have implemented in task 3 of this the assignment.

For the second part of this assignment, you will apply the KNN predictor you implemented above on larger data sets. For this you are required to choose two data sets (one for classification and one for regression) from the following link: Click Here

For each chosen dataset, make sure you do a 70% to 30% split of the dataset to create a the training set and an evaluation set. After you have trained and evaluated your KNN functions, visualize your results using the appropriate visualization techniques you have implemented in task 3 of this the assignment.