CS 529: Assignment #1

Z. Berkay Celik
zcelik@purdue.edu

(Due Sunday 9/8/2019 11:59 PM)

Computer Science, Purdue University — August 23, 2019

Instructions

Info: This HW includes both theory and coding problems. Please read course policy before starting your HW.

- Your code must work with Python 3.5+ (you may install the Anaconda distribution of Python).
- You need to submit a report including solutions of theory problems (in pdf format), and a Jupyter notebook that includes your source code.

1 Problem 1: Background [25pt]

1. Define the terms in a couple of sentences with your own words (10pt):
   - Generalization
   - Overfitting and underfitting
   - Regularization
   - No free lunch theorem
   - Occam's razor
   - Independent and identically distributed data points
   - Cross-validation
   - Degrees of freedom
2. Assume that you observe two different coins being tossed as follows (5pt):

   Coin_1=H,H,H,H,T,T,H,H,H,H,T,T,H,H,H,H,T

   Coin_2=H,H,T,T,T,T,H,H,T,T ,T,T,H,H,T,T,T

   Assume the coin tosses are i.i.d. random variables. Each coin will be tossed one more time and you will be given \$100 for each correct guess. What is your guess for Coin_1 and Coin_2's next toss and why?
3. Find the closed form solution $\mathtt{w}^*$ to minimize the error function $\mathtt{E(w)}$ (10pt).

$$E(w)=\frac{1}{2}\sum_{n=1}^{N}\left\{y\left(x_n,w\right)-t_n\right\}^2, \text{ where}$$

$$y(x,w)=w_0+w_1x+w_2x^2+\ldots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

Then use the found $\mathtt{w}^*$ to illustrate the resulting polynomial $\mathtt{y}(\mathtt{x},\mathtt{w}^*)$ (**Hint:** Use the matrix representation which is often simple and clean).

2 Problem 2: Exploratory Data Analysis with Pandas [25pt]

> **Info:** To answer this question, you are provided with a Jupyter notebook named **HW1P2.ipynb** with questions. Please complete the notebook with your code that answers the questions. You are encouraged to install Anaconda distribution of Python to run the Jupyter notebook to accomplish this task.

As it has been emphasized in the lectures, we need to have a good understanding of data before training a machine learning model. In this assignment, you are asked to analyze the UCI Adult data set, **adult.data.csv**. The Adult data set is a standard machine learning data set that contains demographic information about the US residents. This data was extracted from the census bureau database found here. The data set contains 32561 instances and 15 features (please check the notebook for possible values of each feature) with different types (categorical and continuous). The data is provided as a csv file and can be loaded into panda's DataFrame object as `data = pd.read_csv('adult.data.csv')`.

1. How many men and women (sex feature) are represented in this data set?
2. What is the average age (age feature) of women?
3. What is the percentage of German citizens (native-country feature)?
4. What are the mean and standard deviation of age for those who earn more than 50K per year (salary feature) and those who earn less than 50K per year?
5. Is it true that people who earn more than 50K have at least high school education? (education –Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters or Doctorate feature)
6. Display age statistics for each race (race feature) and each gender (sex feature).
7. What is the maximum number of hours a person works per week (hours-per-week feature)? How many people work such a number of hours, and what is the percentage of those who earn a lot (>50K) among them?
8. Count the average time of work (hours-per-week) for those who earn a little and a lot (salary) for each country (native-country). What will these be for Japan?

3 Problem 3: Linear Regression [50pt]

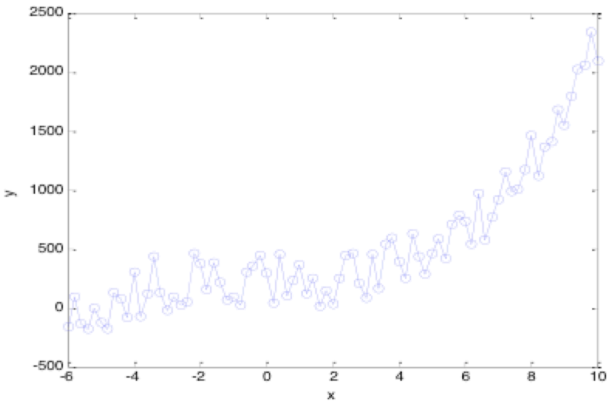
> **Info:** You are given a simple dataset, `data.txt`, which includes an input (`x`) and a continuous output (`y`). We can visualize the dataset below. This problem requires a Jupyter notebook named **HW1P3.ipynb** to be created, that includes the answers for the questions below.

1. Partition all data randomly into 10 folds and produce 10 different training-validation set pairs (3 pts).
2. Normalize your training inputs and outputs by using training sample mean and std deviation (2 pts).
3. For each of the 10 training sets, compute the weights that minimizes the training error:

$$E(w) = \frac{1}{N}\sum_{i=1}^{N}\left(y^t - g\left(x^t, w\right)\right)^2. \tag{1}$$

   , for the following hypotheses classes (15 pts):

   (a) $g(x, w) = w_0 + xw_1$



(b) $g(x, w) = w_0 + xw_1 + x^2w_2 + x^3w_3$

(c) $g(x, w) = w_0 + x^1w_1 + x^2w_2 + x^3w_3 + x^4w_4 + x^5w_5$

(d) $g(x, w) = w_0 + x^1w_1 + x^2w_2 + x^3w_3 + x^4w_4 + x^5w_5 + \ldots + x^{50}w_{50}$

4. For the hypothesis that minimizes the training error (20 pts),

   (a) Plot the errorbar (i.e., the mean and std/sqrt(10)) of training and validation errors (you can use errorbar function, (**Hint:** x axis = 1,. . . , 4 (hypothesis class), y axis = mean error over 10 folds)).

   (b) Plot the training input and outputs and the minimum training error hypothesis outputs for each hypothesis class above (4 plots, 10 hypotheses on each plot).

5. Which hypothesis class would you choose among (a),. . . , (d) and why (10pts)?