CMPSC 448: Machine Learning and AI
Homework 1

Instruction

This HW includes both theory and coding problems. Please note that:

- You cannot look at anyone else's code
- Your code must work with Python 3.7 (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.

Linear Algebra and Calculus

Problem 1. [10 points] What is the rank of the following matrix? Justify your answer.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{bmatrix}$$

Problem 2. [10 points] Use either numpy.linalg or scipy.linalg to find the eigendecomposition of the following matrix:

$$\boldsymbol{X} = \begin{bmatrix} 3 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & -1 & 1 \end{bmatrix}$$

Then, can you reconstruct the original matrix $\boldsymbol{X}$ from obtained eigenvalues and eigenvectors? Do you see any difference between original and reconstructed matrix? Can you Justify?

Problem 3. [20 points] First, compute the derivative $f'(z) = \frac{df(z)}{dz}$ of the scalar function

$$f(z) = \ln(1 + e^{-2z})$$

Then, using the chain rule, what is the gradient $\nabla_{\boldsymbol{w}} g(\boldsymbol{w})$ of the scalar valued function

$$g(\boldsymbol{w}) = f(\boldsymbol{w}^\top \boldsymbol{x}) = \ln(1 + e^{-2\boldsymbol{w}^\top \boldsymbol{x}})$$

Problem 4. [20 points] Let $\boldsymbol{x} \in \mathbb{R}^d$ be a vector in a d dimensional space and define the scalar valued function of d dimensional vectors $f : \mathbb{R}^d \to \mathbb{R}$ by

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a fixed symmetric matrix and $\boldsymbol{b} \in \mathbb{R}^d$ is a fixed vector. Using the definition of gradient show that

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \tag{2}$$

Exploratory Data Analysis with pandas

Problem 5. [40 points] 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. 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. 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 shown below:

```
data = pd.read_csv('adult.data.csv')
```

You are asked to answer following questions about this data set:

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?

To answer these questions, you are provided with a Jupyter notebook 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.

Deliverable

This homework comes with a comma separated csv data file adult.data.csv, and a Jupyter notebook with questions. You are asked to submit a PDF file including the answers for first four questions and the completed notebook for the fifth problem. Make sure your code is running and include enough details about your code.