

CMPSC 448: Machine Learning and Algorithmic 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.

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

The rank of the matrix is 2.

$$\text{rank}(\mathbf{X}) = \begin{bmatrix} -2 & 0 & -6 \\ 0 & -2 & 2 \\ 0 & 0 & 0 \end{bmatrix} = 2$$

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

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

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

After using `numpy.linalg` and `scipy.linalg`, we know that eigenvalues are $\lambda_1 = 4$, $\lambda_2 = 2$, $\lambda_3 = 2$ and eigenvectors are $v_1 = [-1, -2, 1]^T$, $v_2 = [-1, 0, 1]^T$, $v_3 = [-1, 1, 0]^T$. Therefore, the decomposition of the matrix is as follow:

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

Also, we can construct the original matrix X from obtained eigenvalues and eigenvectors:

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

Obviously, two matrices don't have any difference.

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

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

$f'(z) = \frac{1}{e^{-2z}+1} \frac{d}{dz}[e^{-2z} + 1] = \frac{\frac{d}{dz}[e^{-2z}] + \frac{d}{dz}[1]}{e^{-2z}+1} = \frac{e^{-2z} \frac{d}{dz}[-2z] + 0}{e^{-2z}+1} = \frac{(-2 \frac{d}{dz}[z])e^{-2z}}{e^{-2z}+1} = -\frac{2e^{-2z}}{e^{-2z}+1} = -\frac{2}{e^{2z}+1}$ Then, using the chain rule, what is the gradient $\nabla_{\mathbf{w}} g(\mathbf{w})$ of the scalar valued function

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

$$\nabla_{\mathbf{w}} g(\mathbf{w}) = \frac{\partial g}{\partial w_i} \text{ where } \frac{\partial g}{\partial w_i} = \frac{1}{1+e^{-2\mathbf{w}^\top \mathbf{x}}} e^{-2\mathbf{w}^\top \mathbf{x}} (-2x_i) = \frac{-2x_i}{1+e^{2\mathbf{w}^\top \mathbf{x}}} = \frac{-2}{1+e^{2\mathbf{w}^\top \mathbf{x}}} x$$

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

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

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

$$\nabla f(\mathbf{x}) = \mathbf{A} \mathbf{x} + \mathbf{b} \quad (2)$$

$$\begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{21} & \dots & \dots & \dots \\ a_{d1} & \dots & \dots & a_{dd} \end{bmatrix} \quad (3)$$

$$= \begin{bmatrix} \sum_{i=1}^d x_i a_{i1} & \sum_{i=1}^d x_i a_{i2} & \dots & \sum_{i=1}^d x_i a_{id} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix} \quad (4)$$

$$= x_1 \sum_{i=1}^d x_i a_{i1} + x_2 \sum_{i=1}^d x_i a_{i2} + \dots + x_d \sum_{i=1}^d x_i a_{id} \quad (5)$$

$$\nabla(\mathbf{x}^\top \mathbf{A} \mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} (x_1 \sum_{i=1}^d x_i a_{i1} + x_2 \sum_{i=1}^d x_i a_{i2} + \dots + x_d \sum_{i=1}^d x_i a_{id}) \\ \frac{\partial}{\partial x_2} (x_1 \sum_{i=1}^d x_i a_{i1} + x_2 \sum_{i=1}^d x_i a_{i2} + \dots + x_d \sum_{i=1}^d x_i a_{id}) \\ \dots \\ \frac{\partial}{\partial x_d} (x_1 \sum_{i=1}^d x_i a_{i1} + x_2 \sum_{i=1}^d x_i a_{i2} + \dots + x_d \sum_{i=1}^d x_i a_{id}) \end{bmatrix} \quad (6)$$

$$= \begin{bmatrix} (2x_1 a_{11} + \sum_{i \neq 1} x_i a_{i1}) + x_2 a_{12} + \dots + x_d a_{1d} \\ x_1 a_{21} + (2x_2 a_{22} + \sum_{i \neq 2} x_i a_{i2}) + \dots + x_d a_{2d} \\ \dots \\ x_1 a_{d1} + x_2 a_{d2} + \dots + (2x_d a_{dd} + \sum_{i \neq d} x_i a_{id}) \end{bmatrix} \quad (7)$$

$$= \begin{bmatrix} x_1 a_{11} + \sum_{i \neq 1}^d x_i a_{i1} + \sum_{i=1}^d x_i a_{1i} \\ x_2 a_{22} + \sum_{i \neq 2}^d x_i a_{i2} + \sum_{i=1}^d x_i a_{2i} \\ \dots \\ x_d a_{dd} + \sum_{i \neq d}^d x_i a_{id} + \sum_{i=1}^d x_i a_{di} \end{bmatrix} \quad (8)$$

$$= \begin{bmatrix} \sum_{i=1}^d x_i a_{i1} + \sum_{i=1}^d x_i a_{1i} \\ \sum_{i=1}^d x_i a_{i2} + \sum_{i=1}^d x_i a_{2i} \\ \dots \\ \sum_{i=1}^d x_i a_{id} + \sum_{i=1}^d x_i a_{di} \end{bmatrix} \quad (9)$$

Because A is a fixed symmetric matrix,

$$= \begin{bmatrix} 2 \sum_{i=1}^d x_i a_{1i} \\ 2 \sum_{i=1}^d x_i a_{2i} \\ \dots \\ 2 \sum_{i=1}^d x_i a_{di} \end{bmatrix} = 2Ax \quad (10)$$

$$\begin{bmatrix} b_1 & b_2 & \dots & b_d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{bmatrix} = b_1 x_1 + b_2 x_2 + \dots + b_d x_d \quad (11)$$

$$\nabla(b^\top x) = \begin{bmatrix} \frac{\partial}{\partial x_1} b_1 x_1 + b_2 x_2 + \dots + b_d x_d \\ \frac{\partial}{\partial x_2} b_1 x_1 + b_2 x_2 + \dots + b_d x_d \\ \dots \\ \frac{\partial}{\partial x_d} b_1 x_1 + b_2 x_2 + \dots + b_d x_d \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_d \end{bmatrix} = b \quad (12)$$

Therefore, $\nabla f(x) = Ax + b$

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 pandas'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.