CMPSC 448: Machine Learning and AI

Homework 1 (Due 02/14/2021 11:59 PM)

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 3.5+ (you may install the Anaconda distribution of Python)
- For theory problems, please show all the detailed steps.
- You need to submit a report including solutions of theory problems (in pdf format), and a `Jupyter` notebook.

Linear Algebra, Calculus, Probability and Statistics

Problem 1 [20 points] In this problem, you are given two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{2\times 2}$ and a vector $\boldsymbol{x} \in \mathbb{R}^2$

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and asked to answer the following questions about them.

(a) What is $\boldsymbol{A} \times \boldsymbol{B}$?

(b) What is $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$?

(c) What is $\boldsymbol{x}^\top \boldsymbol{x}$?

(d) What is $\boldsymbol{x}\boldsymbol{x}^\top$?

(e) What is the projection of $\boldsymbol{x}$ onto the subspace spanned by the columns of $\boldsymbol{A}$?

(f) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function given by $f(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z}$. What is the gradient of f with respect to $\boldsymbol{z}$, i.e. $\nabla_{\boldsymbol{z}} f(\boldsymbol{z})$?

(g) For the function f defined above, what is $\nabla^2_{\boldsymbol{z}} f(\boldsymbol{z})$ (the Hessian of f with respect to the vector $\boldsymbol{z} \in \mathbb{R}^2$)?

(h) What is the maximizer of f among all vectors with unit Euclidean length, $\|\boldsymbol{z}\|_2 = 1$?

Problem 2 [10 points] For this problem, we use the following notation for random variables:

- $X \sim \mathcal{N}(\mu, \sigma^2)$: X is a Gaussian random variable with mean μ and variance σ^2
- $X \sim \texttt{Bern}(p)$: X is a $\{0, 1\}$-valued Bernoulli random variable with expectation p.
- $\mathbb{E}[X]$: the expected value of random variable X

(a) If $X \sim \mathcal{N}(1, 2)$, then what is $\mathbb{E}[X]$? What is $(\mathbb{E}[X])^2 - \mathbb{E}[X^2]$?

(b) If $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim \texttt{Bern}(p), i = 1, 2, \ldots, n$, what is the distribution of $\sum_{i=1}^{n} X_i$?

(c) Let assume the sequence $\{0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1\}$ is independently drawn from $\texttt{Bern}(p)$ (multiple flips of a biased coin with probability of being head as p which is unknown). What is the maximum likelihood estimator (MLE) of p? Please show the detailed steps (and mathematical derivations you employ).

Problem 3 [5 points] What is the rank of the following matrix and why?

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

Problem 4 [5 points] Use either `numpy.linalg` or `scipy.linalg` to find the eigendecomposition of the following matrix:

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

Problem 5 [5 points] For the function $f(x) = \ln\left(1 + e^{-2x}\right)$, what is it derivative $f'(x) = \frac{df(x)}{dx} = ?$.

Problem 6 [10 points] Let $\boldsymbol{x} \in \mathbb{R}^d$ be a vector in d dimensional space and define the vector valued function $f : \mathbb{R}^d \to \mathbb{R}$ by

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

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

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

Problem 7 [5 points]

(c) What is the maximizer of $g : [-4, 4] \to \mathbb{R}$ given by $g(x) = \frac{1}{2}x^3 - \frac{1}{2}x^2 - 6x + \frac{27}{2}$?

(d) What is $\int_0^1 g(x)dx$ for g defined above?

Exploratory Data Analysis with `pandas`

Problem 5 [40 points] The goal of this problem is to do basic data analysis on a simple data set using `pandas` package in Python (no machine learning for now). 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 found at: http://www.census.gov/ftp/pub/DES/www/welcome.html. 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:

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

You are asked to answer following questions about this data set. Please note that you need to use `pandas` functionalities to answer these questions, rather than implementing pure Python code.

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 you code to answer the questions. You are encouraged to install `Anaconda` distribution of Python to run the Jupyter notebook or directly use `JupyterLab` and accomplish this problem.

Deliverable

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