University of TEHRAN

Machine Learning

Problem Set 1

Hesam Montazeri
Fahimeh Palizban

Mehr 2, 1398
(September 24, 2019)

Problem 1: Review Questions

Write a summary of the lectures of this week. Write down all formulas we discussed in the lectures and explain in detail each step of the derivation. As a guideline, you may consider the following topics:

(a) Supervised learning; regression; classification

(b) Definition of unsupervised learning

(c) Read about the cancer signatures in cancer. What is the input data? What do cancer signatures mean?

(d) Reinforcement learning

(e) Simple linear regression: model; loss functions

(f) Gradient descent algorithm for simple linear regression. Discuss both stochastic and batch versions.

(g) Analytical solution to simple linear regression

(h) Mathematical formulation of multiple linear regression in a matrix form

(i) (Optional) A half-page summary of [PGM] A.5.1

Problem 2: Conceptual questions

[ISL] chapter 2: question 2.

Problem 3: Linear Algebra

(a) Find B^T, B^{-1}, $A \times B$, $v^T B v$, $|B|$, $tr(B)$, $\|v\|_1$ and $\|v\|_2$ where sqaure matrices M and N and the vector v are defined as

$$A = \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 5 \\ 0 & 3 \end{pmatrix} \qquad v = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

(b) Find k such that the following matrices are singular

(i) $\begin{pmatrix} k & 6 \\ 4 & 3 \end{pmatrix}$

(ii) $\begin{pmatrix} 1 & 2 & -1 \\ -3 & 4 & k \\ -4 & 2 & 6 \end{pmatrix}$

(c) Let A be matrix

$$A = \begin{pmatrix} 1 & 3 & 0 \\ 2 & 6 & 0 \\ 1 & 3 & 3 \end{pmatrix}$$

Find the rank of the matrix A.

(d) Given a vector $\beta \in R^p$ and a matrix $C \in R^{p \times p}$, compute gradient and Hessian matrix of $f(\beta) = \beta^T C \beta$ with respect to β.

Problem 4: Programming: Breast Cancer Diagnosis and Prognosis
The input data is a subset of the UCI Breast Cancer Wisconsin (Prognostic) data set ([1]). Each record of the data represents follow-up data for one breast cancer case. The goal is to predict recurrence time of cancer (denoted as T in the data) based on radius of each cell nucleus which computed from a digitized image of a fine needle aspirate of a breast mass (denoted as R) using simple linear regression.

(a) Implement the stochastic and batch gradient descent algorithms as well as the analytical solution. Compare the obtained estimates with those of the built-in function *lm* in R (or the programming language of your choice).

(b) Analyze the data using simple linear regression! (which means plotting and exploring the data, visualization of the estimated model, assessment of error, interpretation of results, etc).

We encourage discussing the problems with other students, however, similarity between solutions is not allowed. Please write in the first page of your submission whom you have brainstormed the questions. Submit your solutions (using Easyclass) by Mehr 6, 1398.

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.