Machine Learning

Problem Set 1

Hesam Montazeri
Fereshteh Fallah
Mozhgan Mozaffari Legha

Bahman 15, 1398
(Feb 4, 2020)

Problem 1: Review part
Write your reviews for the whiteboard notes and the slides of the lectures of this week. Write down all formulas and explain in detail each step of the derivations, if applicable. In addition,

(a) Explain blind source separation problem

(b) (Optional) Write a half-page summary of [PGM] A.5.1

Problem 2: Conceptual questions
[ISL] chapter 2: questions 2, 4, 7a-c.

Problem 3: Linear Algebra

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

$$A = \begin{pmatrix} 1 & 2 \\ 5 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 5 \\ 7 & 3 \end{pmatrix} \qquad v = \begin{pmatrix} -5 \\ 3 \end{pmatrix}$$

(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) Find the rank of the following matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 6 & 9 \\ 1 & 5 & 11 \end{pmatrix}$$

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

(e) Prove the following matrix is positive definite

$$\begin{pmatrix} 7 & 2 \\ 2 & 1 \end{pmatrix}$$

Problem 4: Programming: prediction of acute aquatic toxicity
The aim of this exercise is to predict acute aquatic toxicity for the Pimephales promelas. The input data for this exercise is a subset of UCI QSAR fish toxicity data set ([1]). The dataset includes 908 chemical components. Your task is to study the prediction of LD50, the concentration that cause death in 50% of fish over a test duration, given MLOGP (a molecular property). In particular, you need to

(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) implement KNN and Nadaraya-Watson kernel regression without using existing packages and apply it to the above dataset.

(c) randomly split the data into training and test sets and compare the performance of the above models.

We encourage discussing the problems with other students, however, similarity between solutions is not allowed. (**Important**) Studying any online or previous solutions, no matter to what extent, is strictly forbidden and is considered as a violation of the academic honor code. Submit your solutions by Bahman 19, 1398.

References

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