Machine Learning

Problem Set 2

Hesam Montazeri
Fereshteh Fallah
Mozhgan Mozaffari Legha

Bahman 22, 1398
(Feb 11, 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.

Problem 2: Conceptual questions
[ISL] chapter 2: questions 1, 3, 5, 6, 7d. Chapter 3: question 4. Chapter 6: question 4. In addition

(a) Repeat chapter 6-question 4 for the parameter K of KNN model.

(b) Repeat chapter 6-question 4 for the parameter λ of NW kernel regression.

(c) Run all R commands in section 2.3. Include all results and plots in a report. Explain the results in your own words, whenever applicable.

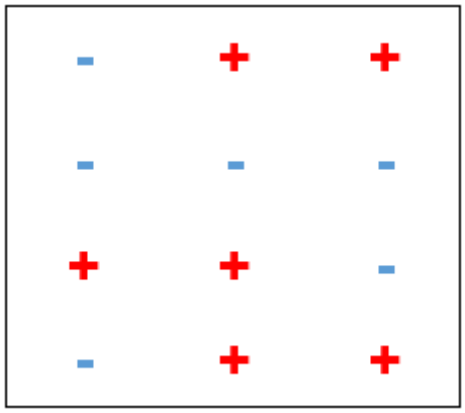
Problem 3: Review- column space
Does the vector w belong to the column space of A? Explain

$$w = \begin{pmatrix} 5 \\ 7 \\ 3 \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 1 & 4 \\ 2 & 3 & 9 \\ 2 & 1 & 7 \end{pmatrix}$$

Problem 4: Feature selection and cross validation
Read carefully [ESL] 7.10.2 and summarize its main points in a paragraph.

Problem 5: Orthogonal projection
Show that the hat matrix, $H = X(X^T X)^{-1} X^T$, of the multiple linear regression is an orthogonal projection.

Problem 6: K-nearest neighbor
On the following dataset, draw the decision boundaries learned by the 1-NN algorithm. Indicate regions with positive or negative labels (assume ties are broken arbitrarily).



Problem 7: 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 UCI QSAR fish toxicity data set [1]. Your task is to study the prediction of LD50, the concentration that cause death in 50% of fish over a test duration, based on six predictors namely CIC0, SM1_Dz(Z), GATS1i, NdsCH, NdssC, MLOGP. In particular, you need to

(a) explore the data.

(b) implement the analytical solution to multiple linear regression and compare the obtained estimates with those of the built-in function *lm* in R.

(c) compare cross validation errors of KNN and multiple linear regression on the above dataset (don't use existing packages for cross validation).

write a short report for this exercise.

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 26, 1398.

References

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