University of TEHRAN

Machine Learning

Problem Set 2

Hesam Montazeri
Fahimeh Palizban
Zohreh Toghraee

Mehr 9, 1398
(October 1, 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 derivations. As a guideline, you may consider the following topics:

(a) Linear algebra review: PD, SPD matrices; linearly independent vectors; linear combination; rank of a matrix

(b) Analytical solution to linear regression in matrix form; its geometric interpretation

(c) Non-linear response from a linear model; overfitting

(d) Ridge regression; LASSO

(e) Generalization error; cross validation

(f) **(Important)** Read carefully [ESL] 7.10.2 and summarize its main points in a paragraph.

Problem 2: Conceptual questions

[ISL] chapter 3: questions 3, 4; chapter 6: questions 3a-b, 4a-b, 5.

Problem 3: Orthogonal projection

Show that

(a) the hat matrix, $H = X(X^TX)^{-1}X^T$, of the multiple linear regression is an orthogonal projection.

(b) (Optional) the ridge hat matrix is not a projection matrix.

Problem 4: Weighted linear regression

Derive the optimal solution $\hat{\beta}$ for the *weighted* loss function:

$$\frac{1}{2}\sum_{i=1}^{n} w[i] \times \left(y[i] - \beta^T x[i]\right)^2$$

where $w[i]$ is the associated weight for ith data point $(x[i],\ y[i])$.

Problem 5: Programming: HIV Drug Resistance

The input data of this exercise is taken from the HIV Drug Resistance database [1]. Your task is to predict phenotypic resistance to the non-nucleoside reverse transcriptase inhibitor, Efavirenz, based on presence of genetic events at selected positions of HIV reverse transcriptase genotype and a phenotypic result from another drug.

(a) Analyze the data using the multiple linear regression, the LASSO, and ridge regression. Compare the performance of the models in terms of the coefficient of determination, R^2.

(b) Repeat part a by using only the first 50 training examples.

(c) Use all training examples again and repeat part a by adding all pairwise interaction terms to the regression models.

Helpful functions in R: cv.glmnet, glmnet, lm, model.matrix.

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 13, 1398.

References

[1] Soo-Yon Rhee, Matthew J Gonzales, Rami Kantor, Bradley J Betts, Jaideep Ravela, and Robert W Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research*, 31(1):298–303, 2003.