University of TEHRAN

Machine Learning

Problem Set 5

Hesam Montazeri
Fereshteh Fallah
Mozhgan Mozaffari Legha

Farvardin 27, 1399
(April 15, 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 4: questions 2, 3, 5

Problem 3: Stochastic gradient ascent rule for Poisson regression
Poisson regression is a GLM where the response variable is assumed to have Poisson distribution:

(a) Write the Poisson distribution in an exponential family form.

(b) Derive the stochastic gradient ascent rule for Poisson regression.

(c) Implement your algorithm and compare the results to the built-in function in R.

Problem 4: Generalized linear models
Give some examples of bioinformatics applications where GLM framework with the following distributions of response variable can be used:

(a) Gaussian

(b) Bernoulli

(c) Negative Binomial

(d) Poisson

(e) Beta

(f) Weibull

In addition, use negative binomial and beta regression models for examples of your choice in R (you may use existing packages in R).

Problem 5: Programming: high-dimensional classification
The goal of this exercise is to perform subject classification to Normal (Y=2), Reflux Esophagitis (Y=3), Barrett's Esophagus (Y=0), Esophageal Adenocarcinoma (Y=1) using abundances of microbial taxa. The dataset was obtained from the laboratory of Zhiheng Pei at New York University (NYU) Langone Medical Center ([1]). Perform the following classifiers:

(a) Naïve Bayes classifier where features assumed to have Poisson distributions

(b) Naïve Bayes classifier where features assumed to have negative binomial distributions

(c) Linear discriminant analysis with a feature selection or dimension reduction method of your choice

(d) LASSO and ridge regression

Use your own implementations for the Naïve Bayes models and briefly explain the mathematics behind the parameter estimation for each part. Compare all models in terms of the generalization error.

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 Ordibehesht 4, 1399.

References

[1] Alexander Statnikov, Mikael Henaff, Varun Narendra, Kranti Konganti, Zhiguo Li, Liying Yang, Zhiheng Pei, Martin J Blaser, Constantin F Aliferis, and Alexander V Alekseyenko. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, 1(1):11, 2013.