

Stat 432 Homework 4

Assigned: Sep 16, 2024; Due: 11:59 PM CT, Sep 26, 2024

- Instruction
- Question 1: Sparsity and Correlation
- Question 2: Shrinkage Methods and Testing Error
- Question 3: Penalized Logistic Regression

Instruction

Please remove this section when submitting your homework.

Students are encouraged to work together on homework and/or utilize advanced AI tools. However, **sharing, copying, or providing any part of a homework solution or code to others** is an infraction of the University's rules on Academic Integrity (<https://studentcode.illinois.edu/article1/part4/1-401/>). Any violation will be punished as severely as possible. Final submissions must be uploaded to Gradescope (<https://www.gradescope.com/courses/570816>). No email or hard copy will be accepted. For **late submission policy and grading rubrics** (<https://teazrq.github.io/stat432/syllabus.html>), please refer to the course website.

- You are required to submit the rendered file `HWx_yourNetID.pdf`. For example, `HW01_rqzhu.pdf`. Please note that this must be a `.pdf` file. `.html` format **cannot** be accepted. Make all of your `R` code chunks visible for grading.
- Include your Name and NetID in the report.
- If you use this file or the example homework `.Rmd` file as a template, be sure to **remove this instruction** section.
- Make sure that you **set seed** properly so that the results can be replicated if needed.
- For some questions, there will be restrictions on what packages/functions you can use. Please read the requirements carefully. As long as the question does not specify such restrictions, you can use anything.
- **When using AI tools**, you are encouraged to document your comment on your experience with AI tools especially when it's difficult for them to grasp the idea of the question.
- **On random seed and reproducibility**: Make sure the version of your `R` is $\geq 4.0.0$. This will ensure your random seed generation is the same as everyone else. Please note that updating the `R` version may require you to reinstall all of your packages.

Question 1: Sparsity and Correlation

During our lecture, we considered a simulation model to analyze the variable selection property of Lasso. Now let's further investigate the prediction error of both Lasso and Ridge, and understand the bias-variance trade-off. Consider the linear model defined as:

$$Y = X^T \beta + \epsilon$$

Where $\beta = (\beta_1, \beta_2, \dots, \beta_{100})^T$ with $\beta_1 = \beta_{11} = \beta_{21} = \beta_{31} = 0.4$ and all other β parameters set to zero. The p -dimensional covariate X follows a multivariate Gaussian distribution:

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_{p \times p}).$$

In Σ , all diagonal elements are 1, and all off-diagonal elements are ρ .

a. [15 points] A single Simulation Run

- Generate 200 training and 500 testing samples independently based on the above model.
- Use $\rho = 0.1$.
- Fit Lasso using `cv.glmnet()` on the training data with 10-fold cross-validation. Use `lambda.1se` to select the optimal λ .

- Report:
 - Prediction error (MSE) on the test data.
 - Report whether the true model was selected (you may refer to HW3 for this property).
- b. [15 points] Higher Correlation and Multiple Simulation Runs
- Write a code to compare the previous simulation with $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$.
 - Perform 100 simulation runs as in part a) and record the prediction error and the status of the variable selection for Lasso.
 - Report the average prediction error and the proportion of runs where the correct model was selected for each value of ρ .
 - Discuss the reasons behind any observed changes.
- c. [15 points] Ridge Regression
- Repeat task b) with the ridge regression. You do not need to record the variable selection status since ridge always select all variables.
 - Report the average prediction error, do you see any difference between ridge and Lasso? Any performance differences within ridge regression as ρ changes?
 - Discuss the reasons behind any observed changes.

Question 2: Shrinkage Methods and Testing Error

In this question, we will predict the number of applications received using the variables in the College dataset that can be found in ISLR2 package. The output variable will be the number of applications (Apps) and the other variables are predictors. If you use Python, consider migrating the data to an excel file and read it in Python.

- a. [10 pts] Use the code below to divide the data set into a training set (600 observations) and a test set (177 observations). Fit a linear model (with all the input variables) using least squares on the training set using `lm()`, and report the test error (i.e., testing MSE).

```
library(ISLR2)
data(College)

# generate the indices for the testing data
set.seed(7)
test_idx = sample(nrow(College), 177)
train = College[-test_idx,]
test = College[test_idx,]
```

- b. [10 pts] Compare Lasso and Ridge regression on this problem. Train the model using cross-validation on the training set. Report the test error for both Lasso and Ridge regression. Use `lambda.min` and `lambda.1se` to select the optimal λ for both methods.
- c. [20 pts] The `glmnet` package implemented a new feature called `relaxed fits` and the associated tuning parameter `gamma`. You can find some brief explanation of this feature at the documentation of this package. See
- CRAN Documentation (<https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>)
 - glmnet Vignette (<https://glmnet.stanford.edu/articles/glmnet.html>)

Read these documentations regarding the `gamma` parameter, and summarize the idea of this feature in terms of the loss function being used. You need to write it specifically in terms of the data vectors \mathbf{y} and matrix \mathbf{X} and define any notations you need. Only consider the Lasso penalty for this question.

After this, implement this feature and utilize the cross-validation to find the optimal λ and γ for the College dataset. Report the test error for the optimal model.

Question 3: Penalized Logistic Regression

In HW3, we used `golub` dataset from the `multtest` package. This dataset contains 3051 genes from 38 tumor mRNA samples from the leukemia microarray study Golub et al. (1999). The outcome `golub.cl` is an indicator for two leukemia types: Acute Lymphoblastic Leukemia (ALL) or Acute Myeloid Leukemia (AML). In genetic analysis, many gene expressions are highly correlated. Hence we could consider the Elastic net model for both sparsity and correlation.

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("multtest")
```

[15 pts] Fit logistic regression to this dataset. Use a grid of α values in $[0, 1]$ and report the best α and λ values using cross-validation.