Spring 2023, 958:588 Financial Data Mining Homework 2.

DUE: Midnight, Friday, March 10th

INSTRUCTIONS: Submit all files on Canvas. For the free response parts, type your solution into an electronic document or take a picture of your solution and make sure that all hand-writing is legible. For the code parts, submit the completed files on Canvas. **Email and physical submissions will NOT be accepted.**

For code that imports csv files, make sure that all the imported files are in the same directory as the code.

NOTE: This homework is available in both R and Python. You may choose the programming language most familiar to you.

NOTE: Your code must run without error. Points may be deducted if there is any error that causes the code to crash.

NOTE: You may work in groups with other students but you must write down the names of all the members of your group.

Problem 1: Optimization for Logistic Lasso with Proximal Gradient Descent

Download the `hw2_problem1` R or Python notebook file and complete the code by filling in all parts labeled "FILL IN". You may find it helpful to review Sections 2.3, 3.3, and 4.3 of the lecture notes.

After completing the code, run the code to compare the solution computed by your code against the R package `glmnet` or the Python Scikit-learn function `LogisticRegression`. Report the estimation error as well as the deviation of your solution from the software package solution.

Problem 2: Predicting voting patterns with Kernel SVM

Download the `hw2_problem2` R or Python file and `votes.csv`. Our dataset `votes` consists of over 3000 counties in the United States along with their socio-economic and demographic information and voting records in the 2016 US election; each row corresponds to a single county. Our response variable will be `prefer_trump`, which is 0 or 1 indicating whether the percentage of people who

voted for Trump in that county is greater than that who voted for Clinton in the 2016 US presidential election.

HINT: you may find ways to simplify your code by first working through Problem 4 of this homework.

Part a (code)

The function `kernelSVM` is an incomplete implementation of kernel SVM.

Complete the code in `hw2_problem2` R or Python file by filling in all the parts with the label `FILL IN`. Most of the "FILL IN"'s involve making prediction with kernel SVM.

Part b (free response)

We now use kernel SVM to predict the `prefer_trump` variable from a variety of socio-economic and demographic features.

Run with the code with degree set to 1, then 2, 3, 4, 5, and 6. Report the predictive error as well as the number of support vectors in each of the cases.

How does the predictive error change with the degree of the polynomial kernel? Explain why.

How does the number of support vectors change with the degree of the polynomial kernel? Explain why.

Problem 3: Predicting Shaquille O'Neal's free throw outcomes

Part a (code)

Download the `hw2_problem3` R or Python file and `shaq.csv`. Our dataset `shaq` consists of the 1632 free throws that Shaquille O'Neal attempted in the regular NBA seasons from 2006 to 2011. Each row correspond to one free throw attempt along with some other information such as:

- `shot_made`: 0/1 indicating whether the free throw was successful.
- `home_game`: 0/1 indicating whether the game is an away game or a home game.
- `first_shot`: 0/1 indicating whether the free throw is the first of the two throws.
- `cur_score:` score of Shaq's team at the time of the free throw.
- `opp_score:` score of the opposing team at the time of the free throw.
- `score_ratio:` cur_score/(opp_score + 1).
- `cur_time:` number of seconds into the game at the time of the free throw.
- `period:` the quarter in which the free throw is made.

In the part of the code labeled `Part (a)`, perform some basic exploratory analysis following the instructions listed. Use the R command `chisq.test` or the Python command `chi2_contingency` for the χ^2 association test.

Part b (code)

We will predict `shot_made` using the other features with logistic ridge regression. For R, we will use the package `glmnet` for this problem. Use the documentation
`https://cran.r-project.org/web/packages/glmnet/glmnet.pdf`.

For Python, we will use the scikit-learn function `LogisticRegressionCV` for this problem with `Cs=50, cv=5, penalty='l2', solver='lbfgs', max_iter=1000`. Use the documentation
`https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html`.

Complete the code by filling in all parts labeled "FILL IN". Report the ridge and baseline errors.

Part c (free response)

Choose the statement below which you agree with the most. Justify your answer in a sentence or two.

1. The features are strongly predictive of Shaquille O'Neal's free throws.
2. The features are weakly predictive of Shaquille O'Neal's free throws.
3. There is no evidence that the features are at all predictive of Shaquille O'Neal's free throws.

Part d (free response)

Give a two or three sentences answer to each of the following questions.

1. Each row of the original dataset also contains the final score of the game in which the free throw was made. Why is it that we cannot use the final score of the game as a feature in making the prediction?
2. Why is it that, ideally, when holding out samples for the test set, we should choose a set of games and hold out ALL samples from those games?

Problem 4: Matrix Algebra for SVM

Part a (free response)

Let $X \in \mathbb{R}^{n \times p}$ matrix whose i-th row correspond to the i-th data point X_i. We wish to compute an $n \times n$ matrix M such that $M_{i\ell} = X_i^\top X_\ell$ is the inner product between the i-th data point and the ℓ-th data point.

Take the setting where $n = 3$ and $p = 2$; show that we have $M = XX^\top$.

Part b (free response)

Let $Y \in \mathbb{R}^n$ be a vector. We wish to compute an $n \times n$ matrix $\tilde{M}$ such that $\tilde{M}_{i\ell} = Y_i Y_\ell X_i^\top X_\ell$.

Take the setting where $n = 3$ and $p = 2$; show that we have $M = D_Y X X^\top D_Y$ where D_Y is a diagonal matrix such that $(D_Y)_{ii} = Y_i$.