Homework 4

Stats 102B Lec 1 and 2

Spring 2024

General Guidelines

Show all your work, including any and all relevant code and output. Any and all collaboration must adhere to **Level 1** collaboration described in the Stats 102B Collaboration Policy.

Please submit your homework as a single file, in PDF format only. Name your assignment file with the convention `123456789_stats102b_hw0.pdf`, where `123456789` is replaced with your UID and `hw0` is updated to the actual homework number. Include your first and last name and UID in your assignment as well.

All R code is expected to follow the Tidyverse Style Guide: https://style.tidyverse.org/. If you scan or take a picture of any written work, please make sure the resolution is high enough that your work is clear and legible. Submissions with severe style or formatting issues may receive a penalty. Any submission that cannot be properly read will not be graded.

Please read in the `volleyball.csv` dataset from Bruin Learn.

Question 1

(a)

In volleyball, players spike or attack the ball in an effort to get the ball to their opponent's side quickly. They also try to block on defense to prevent the opponent's spike. A volleyball player has a maximum spike touch and maximum block touch, measured in centimeters. Fit a simple least squares regression model predicting block from spike and output the parameters. Store the coefficients for later.

(b)

Recall from lecture that least squares regression optimizes the MSE loss function. Instead, we will find parameters that optimize the MAE and Huber loss functions. However, both of these loss functions are not differentiable, so we will have to resort to a zero-order method: coordinate descent!

Write a function `coord_desc(g, w0, d, K)` that runs (zero-order) coordinate descent with the following parameters:

- `g` is a function $g : \mathbb{R}^n \to \mathbb{R}$
- `w0` is the initial point
- `d` is the step size in each direction
- `K` is the number of iterations

The output should be a list object with the following components:

- `$index` should represent w^*, the local minimum
- `$val` should represent $g(w^*)$, the value of the function at the local minimum

(c)

The MAE loss is defined below.

$$g(w) = \frac{1}{n}||y - f(x|w)||_1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i|w)|$$

First, write a function that computes the MAE loss with y corresponding to block and x corresponding to spike. Then use your coordinate descent with w_0 as the simple linear regression coefficients, $d = 0.01$, and $K = 100$ to estimate w. Show the `$index`.

(d)

The Huber loss is defined below.

$$g(w) = \frac{1}{n}\sum_{i=1}^{n}L_\delta(y_i, w)$$

where

$$L_\delta(y_i, w) = \begin{cases} \frac{1}{2}[y_i - f(x_i|w)]^2 & \text{if } |y_i - f(x_i|w)| \leq \delta, \\ \delta|y_i - f(x_i|w)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

First, write a function that computes the Huber loss with y corresponding to block and x corresponding to spike. Then use your coordinate descent with w^0 as the simple linear regression coefficients, $d = 0.01$, $K = 100$, and $\delta = 18$ to estimate w. Show the `$index`.

(e)

Create a scatterplot of block vs spike with the least squares regression line, MAE loss line, and Huber loss line. Give each line a unique color and create an appropriate legend. Comment on your plot.

(f)

Vary the δ parameter in your Huber loss and fit the parameters. Comment on your findings.

Question 2

(a)

Now we will investigate which model can best predict a volleyball player's maximum spike height from their weight.

Consider the following models:

- Linear: $y_i = w_0 + w_1 x_i + \varepsilon_i$.
- Quadratic: $y_i = w_0 + w_1 x_i + w_2 x_i^2 + \varepsilon_i$
- Cubic: $y_i = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + \varepsilon_i$
- Degree 14 Polynomial: $y_i = w_0 + w_1 x_i + \cdots + w_{14} x_i^{14} + \varepsilon_i$
- Exponential: $\log(y_i) = w_0 + w_1 x_i + \varepsilon_i$, equivalent to $y_i = \exp(w_0 + w_1 x_i + \varepsilon_i)$

Create a scatterplot of spike vs weight and plot the models above directly on the graph. Comment on your plot.

(b)

Implement K-fold cross validation with $K = 10$ on each of the models using the MAE. Show the validation errors and comment on your results. (If the degree 14 polynomial performs the best, that's okay.)

Question 3

From lecture, we saw that high model complexity can lead to high variance in the fitted model, and we used out-of-sample prediction error on validation data to reduce overfitting. Another way to prevent a model from becoming too complex is to control the variance through imposing a penalty for model complexity in the loss function.

Note: Since variance is sensitive to the units of measurement of the input variables, we assume all input variables are first *standardized*, i.e., each input variable has mean 0 and standard deviation 1.

A linear model $y = Xw + \varepsilon$ will have higher variance if the magnitude of the parameter vector w is larger, so we can use the squared L_2 norm $||w||_2^2 = w^T w$ as a measure of model complexity. Rather than minimizing MSE, we introduce a **penalty term** $\lambda w^T w$ and minimize the **ridge regression** loss function

$$g(w) = \underbrace{\frac{1}{n}||y - Xw||_2^2}_{\text{MSE}} + \underbrace{\lambda||w||_2^2}_{\text{penalty term}} = \frac{1}{n}(y - Xw)^T(y - Xw) + \lambda w^T w.$$

The penalty parameter λ controls the trade-off between not fitting the data well (MSE) and penalizing overly complex models ($w^T w$). The global minimum $w^* = \underset{w \in \mathbb{R}^{p+1}}{\arg\min}\, g(w)$ of the ridge regression loss function is called the **ridge regression estimator**.

(a)

We will now fit a ridge regression model using three different techniques. However, before we do that, it is important to standardize our predictors and response variable. We will try to predict spike height from player height, weight, and block height. Create the appropriate design matrix X where each column is standardized. The first row is shown below to check your work:

```
##            height   weight    block
## 1.000000 1.456365 1.530021 1.147310
```

(b)

Consider the ridge loss function

$$g(w) = \frac{1}{n}(y - Xw)^T(y - Xw) + \lambda w^T w.$$

Using matrix calculus, derive a formula for the global minimum w^* and justify why it is a global minimum. Then, for $\lambda = 1$, compute the value of w^* using your design matrix above and y vector being the spike height.

(c)

Use Newton's Method with $w_0 = (0, 0, 0, 0)^T$ and $K = 10$ to find w^*.

Hint: You can use your function from Homework 3 Question 1(a).

(d)

Using Newton's Method with the ridge loss is equivalent to using another method from lecture with the MSE. Identify which method it is and explain.

(e)

The following code fits a ridge regression model. Run it to check your work (Josh is looking out for you ♡).

```r
library(glmnet)
ridgefit <- glmnet(X[,-1], y, alpha = 0)
as.matrix(t(coef(ridgefit, s = 1)))
```

Question 4

The Libero in volleyball is a strictly back row position that specializes in defense. They wear an off-colored jersey, cannot attack, and substitute freely for the team's middle blockers.

Consider the cross-entropy loss for logistic regression, defined by

$$g(w) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log\left[\sigma(w^T\boldsymbol{x}_i)\right] + (1-y_i)\log\left[1-\sigma(w^T\boldsymbol{x}_i)\right]\right).$$

The gradient can be computed to be

$$\nabla g(w) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sigma(w^T\boldsymbol{x}_i)\right)\boldsymbol{x}_i.$$

Suppose we are trying to predict whether a player is a libero or not from their height and spike height.

(a)

Write functions for the cost and gradient. Output below has been provided for you to check your work.

Hint: Set y to be `is_Libero` and create an appropriate design matrix (be sure to include an intercept column).

(b)

Use normalized gradient descent with $w^0 = (25, 0, 0)^T$, $\alpha = 0.001$, $K = 10000$ to estimate the logistic regression parameters (it will take a while).

(c)

Run the following code that fits a logistic regression model. Do you parameter estimates match?

```r
summary(glm(y ~ X[, -1], family = "binomial"))
```