Math 7243 Homework 2

Author: Jacob Zelko

Collaborators: Christopher Cesares, ChatGPT

Date: January 30th, 2024

Summary: Ridge regression, cost functions, and expected values

Keywords: #covariance #matrix #expected #value #loss #regression

Statement of Academic Integrity: Outside of course office hours, I collaborated with Chris Cesare and utilized ChatGPT on this assignment.

Notes

Set-Up

Importing necessary packages:

```
using DataFrames
using LinearAlgebra
```

Problem 1

Let X be the data matrix and θ be the parameter vector.

Part A In Lecture 2, we showed that the residual sum of square can be written

$$\text{RSS}(\theta) = (Y - X\theta)^T(Y - X\theta)$$

Find a critical point for RSS(θ) by calculating $\frac{\partial}{\partial\theta}$RSS($\theta$) $= 0$.

I first simplified the expression for RSS as follows:

$$(Y^T - \theta^T X^T)(Y - X\theta)$$

Via multiplication operations, I get the expanded form:

$$Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta$$

Which can be simplified even further to:

$$Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta$$

Then, finally, taking the partial derivative of this function, I get the formulation:

$$2X^T X\theta - 2X^T Y = \frac{\partial}{\partial\theta}\text{RSS}(\theta)$$

Then, evaluating when this formulation is equal to 0, I get the form (assuming that X^TX is invertible):

$$\boxed{\theta = (X^TX)^{-1}X^TY}$$

Part B Ridge regression changes the loss function to add in a term penalizing the θ if they get to large: For any positive number λ, the Ridge loss function

$$\text{Ridge}_\lambda(\theta) = (Y - X\theta)^T(Y - X\theta) + \lambda\theta^T\theta$$

Find an expression for the location of the critical point of Ridge$_\lambda(\theta)$.

My approach here was nearly the same as Part A, but I considered the additional term $\lambda\theta^T\theta$. For that reason, I will jump to where I evaluate the partial derivative of this function:

$$2X^TX\theta - 2X^TY + 2\lambda\theta = \frac{\partial}{\partial\theta}\text{Ridge}_\lambda(\theta)$$

And setting this equal to 0, I get the following formulation:

$$X^TX\theta + \lambda\theta^T = X^TY$$

And then from there, I can rearrange the formulation as:

$$\boxed{\theta = (X^TX + \lambda I)^{-1}X^TY}$$

(This assumes that $(X^TX + \lambda I)$ is invertible)

Problem 2

Part A Description: Fit a linear function to this dataset when the loss is RSS. You may use a computer to solve the matrix equation but you should report the best fit function.

```
x = [1.2, 3.2, 5.1, 3.5, 2.6]
y = [7.8, 1.2, 6.4, 2.6, 8.1]
```

```
h(X, theta) = theta[1]*X[:, 1] .+ theta[2]*X[:, 2];
# The ones vector here allows for calculation of the intercept
X = hcat(ones(length(y)), x)
theta = (X' * X)^-1 * X' * y
```

The best fit function is:

$\boxed{h_\theta(x) = 7.3311 - 0.67663x_2}$

And using the RSS loss,

$$\sum_{i=1}^{n}(h_\theta(\vec{x}^{(i)}) - y^{(i)})^2$$

```
rss(X, theta) = sum((h(X, theta) .- y).^2)
```

We then get: $\boxed{35.6925}$ as the RSS loss.

Part B Fit a linear function to this dataset when the loss is the Ridge Loss from Problem 1.b) with $\lambda = 1$ and with $\lambda = 10$. What specifically explains the difference in values between the three fits?

```
x = [1.2, 3.2, 5.1, 3.5, 2.6]
y = [7.8, 1.2, 6.4, 2.6, 8.1]
```

Function definition for ridge regression

```
function ridge_regression(x, y, lambda)
   n = length(x)
   X = hcat(ones(n), x)
   beta = (X'X + lambda * I) \ (X'y)
   return beta
end
```

Function definition for ridge regression loss

```
function ridge_regression_loss(x, y, beta, lambda)
   # Prediction function
   y_pred = beta[1] .+ beta[2] * x
   residuals = y_pred .- y
   # RSS - L1
   loss = sum(residuals.^2) + lambda * sum(beta[2:end].^2)
   return loss
end
```

To evaluate the ridge regression fit and loss, I first calculate the following:

```
beta_1 = ridge_regression(x, y, 1)
loss_1 = ridge_regression_loss(x, y, beta_1, 1)

beta_10 = ridge_regression(x, y, 10)
loss_10 = ridge_regression_loss(x, y, beta_1, 10)
```

With $\lambda = 1$, we get: $\boxed{48.5046}$ as the ridge regression loss and for the fit being $h_\theta(x) = 3.1152 + 0.4749 * x_2$

Then with $\lambda = 10$, we get: $\boxed{50.5344}$ as the ridge regression loss and for the fit being $h_\theta(x) = 0.7334 + 0.9679 * x_2$

The difference between the three fits is that there is a changing penalty term applied to each of the loss functions. The generic RSS reports the "best fit" since there was no penalty followed by the Ridge Loss fits at $\lambda = 1, 10$ which imposed a penalty that depended on the value of what λ was.

Problem 3

Assume the data $\mathcal{D} = (X, \vec{y})$ was drawn from $y = \vec{\theta}_*^T \vec{x} + \epsilon$ and $\epsilon \sim \text{Normal}(0, \sigma^2)$. The Ridge Regression estimate for $\vec{\theta}_*$ is given by

$$\hat{\vec{\theta}} = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

Part 1 Find the Expected value $E_{\mathcal{D}}(\hat{\vec{\theta}})$ of the Ridge Regression $\hat{\vec{\theta}}$ over data $\mathcal{D} = (X, \vec{y})$.

My process for evaluating this was as follows:

$$E_{\mathcal{D}}(\hat{\vec{\theta}}) = E((X^T X + \lambda I)^{-1} X^T \vec{y})$$

Then, doing the substituion:

$$E_{\mathcal{D}}(\hat{\vec{\theta}}) = E((X^T X + \lambda I)^{-1} X^T (X\theta^T + e))$$

Doing some manipulation, this can be rearranged as:

$$E_{\mathcal{D}}(\hat{\vec{\theta}}) = E((X^T X + \lambda I)^{-1} X^T X\theta^T + (X^T X + \lambda I)^{-1} X^T e)$$

Then knowing that the mean of e is 0 (as it is sampled from the given Normal distribution), we can further simplify this expression to get the following expected value:

$$\boxed{(X^T X + \lambda I)^{-1} X^T X\theta^T}$$

Part 2 Is $\hat{\vec{\theta}}$ an unbiased estimator for $\vec{\theta}_*$?

$\boxed{\textbf{unbiased}}$

Part 3 Find the Covariance matrix $\text{Cov}(\hat{\vec{\theta}})$ of the Ridge Regression $\hat{\vec{\theta}}$ over data $\mathcal{D} = (X, \vec{y})$

How I evaluated this was by doing the following:

$$\mathbf{Var}(\hat{\vec{\theta}}) = \mathbf{Var}((X^T X + \lambda I)^{-1} X^T \vec{y})$$

Through some manipulation, we can get this:

$$(X^T X + \lambda I)^{-1} X^T \mathbf{Var}(\vec{y})((X^T X + \lambda I)^{-1} X^T)^T$$

Then, substitution into the variance term and some rearrangement, we can get the final value being:

$$\boxed{\sigma^2 (X^T X + \lambda I)^{-1} X^T ((X^T X + \lambda I)^{-1} X^T)^T}$$