HW2

Eric Mariasis

10/10/2021

Exercise 1

If we use ridge regression and both the training error and validation error are high, the model suffers from high bias because the model is likely heavily regularized and underfitting the training data. Three potential ways to mitigate high bias are to decrease the regularization term λ, add more input features, and increase the amount of time training the model.

Exercise 2

If you want to classify pictures as outdoor/indoor and daytime/nighttime, you would implement two logistic regression classifiers because the classes are not exclusive/all combinations of the classes are possible. Softmax regression should only be used when the classes are exclusive. Therefore this is a two-class two-label classification problem.

Exercise 7.2

$\hat{W} = (\phi(x)^T \phi(x))^{-1} \phi^T Y$

```
x = matrix(c(0,0,0,1,1,1), nrow=6, byrow=TRUE)
Y = matrix(c(-1,-1,-1,-2,-2,-1,1,1,1,2,2,1), nrow=6, byrow = TRUE)
phi_x = matrix(c(1,0,1,0,1,0,0,1,0,1,0,1), nrow=6, byrow = TRUE)
# t function takes transpose and solve takes inverse
phi_x_transpose_phi_x_inv = solve(t(phi_x) %*% phi_x)
W = phi_x_transpose_phi_x_inv %*% t(phi_x) %*% Y
print(W)
```

```
##           [,1]      [,2]
## [1,] -1.333333 -1.333333
## [2,]  1.333333  1.333333
```

Exercise 7.5

First, take the derivative of RSS(w, w_0) w.r.t w_0 and set to 0 in order to find the MLE of w_0.

$\frac{d}{dw_0} = \frac{d}{dw_0} of \sum_{i=1}^{N} (y_i - w^T x_i - w_0)^2$

$= -2 \sum_{i=1}^{N} (y_i - w^T x_i - w_0) = 0$

$= \sum_{i=1}^{N} (y_i - w^T x_i) = N w_0$

The $W^T x_i$ is equivalent to $x_i^T w$ because of the design matrix of 1's.

$\sum_{i=1}^{N} (y_i - x_i^T w) = N w_0$

$\hat{w_0} = \frac{1}{N} \sum_{i=1}^{N} (y_i) - \frac{1}{N} \sum_{i=1}^{N} (x_i^T w)$

$\frac{d}{dw} = \frac{d}{dw} of \sum_{i=1}^{N}(y_i - w^T x_i - w_0)^2$

$\frac{d}{dw} = \frac{d}{dw} of \sum_{i=1}^{N}(y_i - w^T x_i - \overline{y}_i + \overline{x}_i^T w)^2$

$= \sum_{i=1}^{N}(y - \overline{y}_i - (x - \overline{x}_i)^T w)^2$

$= \sum_{i=1}^{N}(y_{ic} - (x_{ic})^T w)^2$

$= (y_c - x_c w)^T * (y_c - x_c w)$

The above is in the form needed to derive $\hat{w}_{ols}$ in the book.

This is

$(X_c^T X_c)^{-1} X c^T y_c$

Exercise 7.8

```
# Part A
x = matrix(c(1,1,1,1,1,1,1,1,1,1,1,94,96,94,95,104,106,108,113,115,121,131), nrow = 11)
y = c(0.47, 0.75, 0.83, 0.98, 1.18, 1.29, 1.40, 1.60, 1.75, 1.90, 2.23)
D = matrix(c(x[,2], y), nrow=11)
N = length(x[,2])
w1 = cov(x[,2],y)/cov(x[,2],x[,2])
w0 = mean(y) - w1 %*% mean(x[,2])
w0_orig = matrix(c(0,1))
print(paste("w0 is",w0,"w1 is",w1))
```

```
## [1] "w0 is -3.25642848403279 w1 is 0.0426514131897712"
```

```
y_hat = c(w0) + c(w1) * x[,2]
sig2_hat = (1/(N-2))*sum((y-y_hat)^2)
print(paste("sig2_hat is",sig2_hat))
```

```
## [1] "sig2_hat is 0.0169746237611648"
```

```
# Part C
V0_inv = matrix(c(0,0,0,1), byrow = TRUE, nrow = 2)
Vn = sig2_hat * solve((sig2_hat * V0_inv) + (t(x) %*% x))
Wn = Vn %*% V0_inv %*% w0_orig + (1/sig2_hat) * Vn %*% t(x) %*% y
print(paste("marginal posterior is",Vn[2,2]))
```

```
## [1] "marginal posterior is 1.14229003107198e-05"
```

```
# Part D
print(paste("The 95% credible interval is","(",Wn[2,]-Vn[2,2],",",Wn[2,]+Vn[2,2],")"))
```

```
## [1] "The 95% credible interval is ( 0.0426509259869298 , 0.0426737717875513 )"
```

For the prior, $p(w_0) = 1$ because it is a uniform prior. Therefore, $p(w) = p(w_0)p(w_1)$ is also of Gaussian form because $p(w_1)$ which is Gaussian is just being multiplied by 1.

Exercise 8.3 Part A

$\frac{d}{da}\frac{1}{1+e^{-a}}$

$= -\frac{\frac{d}{da}[e^{-a}+1]}{(e^{-a}+1)^2}$

$= -\frac{e^{-a}\frac{d}{da}[-a]}{(e^{-a}+1)^2}$

$= \frac{e^{-a}}{(e^{-a}+1)^2}$

$\sigma(a)(1-\sigma(a))$

$= \frac{1}{1+e^{-a}}(1-\frac{1}{1+e^{-a}})$

$= \frac{1}{1+e^{-a}}(\frac{1+e^{-a}}{1+e^{-a}}-\frac{1}{1+e^{-a}})$

$= \frac{1}{1+e^{-a}}(\frac{e^{-a}}{1+e^{-a}})$

$= \frac{e^{-a}}{(e^{-a}+1)^2}$

Exercise 8.3 Part B

$\frac{d}{dw}(\sigma(w^T x_i))$

Using the chain rule, first use $\sigma(a)(1-\sigma(a))$ from the prior part. This yields $\sigma(w^T x_i)(1-\sigma(w^T x_i))$. Then multiply that by x_i which is the derivative of the inner part. The sigma expressions are equivalent to μ so we get:

$\mu_i(1-\mu_i)(x_i)$

Then we take the derivative of $log(\mu_i)$ which is $\frac{1}{\mu_i} * d(\mu_i) = \frac{1}{\mu_i}\mu_i(1-\mu_i)(x_i)$. The μ_is cancel out to get $(1-\mu_i)(x_i)$

For the derivative of $log(1-\mu_i)$ it is similar except the inner $1-\mu_i$ will be in the denominator and we need to multiply by -1 for the $-\mu_i$. This equals $\frac{-\mu_i(1-\mu_i)(x_i)}{1-\mu_i}$

The $1-\mu_i$ terms cancel out to yield $-\mu_i x_i$.

Next we plug these log results back into the original negative log likelihood expression.

$-\sum_{i=1}^{N}(y_i[(1-\mu_i)(x_i)] - (\mu_i x_i(1-y_i)))$

$= -\sum_{i=1}^{N}(y_i x_i - y_i\mu_i x_i - \mu_i x_i + y_i\mu_i xi)$

The $y_i\mu_i x_i$ terms cancel leaving

$= -\sum_{i=1}^{N}(y_i x_i - \mu_i x_i)$

$= -\sum_{i=1}^{N}(x_i(y_i - \mu_i))$

$= \sum_{i=1}^{N}(x_i(\mu_i - y_i))$ which is the gradient of the log likelihood,

Exercise 8.7

Sketches that the answer references are shown in attached file HW2_8_7.png.

For part a, the decision boundary line to minimize the loss can essentially be drawn in any fashion that successfully separates the +'s from the O's. The answer is not unique because the slope of that line can be adjusted slightly so there are multiple possible ways to draw that boundary. And there are no classification errors since the datasets are successfully separated.

For part b, the line drawn needs to pass through the intersection of x1 and x2 in other words the origin. This line can do pretty well with minimizing loss but the lower + sign will be misclassified even in the best case so therefore that line will make 1 error.

For part c, the line drawn needs to be horizontal since the regularization of w1 to 0 effectively pushes all the data points toward the x2 axis and therefore the decision boundary needs to be a horizontal line. The minimized loss for this yields two classification errors, the top most o datapoint and the bottom most + datapoint.

For part d, the line drawn needs to be vertical since the regularization of w2 to 0 effectively pushes all the data points down toward the x1 axis and therefore the decision boundary needs to be a vertical line. The minimized loss for this yields no classification errors since a vertical line can successfully separate the +'s from the o's.

Exercise 10.5 Part A

The only vertex has a directed path to S is G, so the only scenarios that need to be taken into account are when G = 0 and G = 1. The $|V = 1$ can be disregarded since V has no path to G.

$P(S = 1) = P(G = 0) * P(S = 1|G = 0) + P(G = 1) * P(S = 1|G = 1) = \alpha * (1 - \gamma) + (1 - \alpha) * (1 - \beta)$

Exercise 10.5 Part B

As described in the prior part, V does not have a directed path to S so in $P(S = 1|V = 0)$, the V term can be disregarded and the expression is the same which is $\alpha * (1 - \gamma) + (1 - \alpha) * (1 - \beta)$.

Exercise 10.5 Part C

$\alpha = P(G = 0) = \frac{1}{3}$

$\beta = P(S = 0|G = 1) = 0$ (In this chart, S never equals 0 when G = 1)

$\gamma = P(S = 0|G = 0) = 1$