EECS 545 Machine Learning: Homework #2

Due on February 8, 2022 at 12pm

Professor Honglak Lee Section A

Yuang Huang
yahuang@umich.edu

Problem 1

Logistic regression

Solution
Part a: Hessian **H**

$$l(\mathbf{w}) = \sum_{i=1}^{N} y^{(i)} \log h(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \log(1 - h(\mathbf{x^{(i)}})), \tag{1}$$

where $h(\mathbf{x}) = \sigma(\mathbf{w^T x}) = \frac{1}{1+\exp(-\mathbf{w^T x})}$ and we denote that $pred = \mathbf{w^T x}$.
Then we assume that:

$$l_i(\mathbf{w}) = y^{(i)} \log \sigma(pred^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(pred^{(i)})), \tag{2}$$

where we know that $\frac{\partial pred}{\partial \mathbf{w}} = \mathbf{x^T}$ and $\frac{\partial pred}{\partial \mathbf{w^T}} = \mathbf{x}$.
It can be shown that:

$$\begin{aligned}\nabla l_i(\mathbf{w}) &= \frac{y^{(i)} x^{(i)}}{\sigma(pred^{(i)})} - \frac{(1 - y^{(i)}) x^{(i)}}{(1 - \sigma(pred^{(i)}))} \\ &= y^{(i)} x^{(i)} (1 - \sigma(pred^{(i)})) - (1 - y^{(i)}) x^{(i)} \sigma(pred^{(i)}) \\ &= x^{(i)} y^{(i)} - x^{(i)} \sigma(pred^{(i)})\end{aligned} \tag{3}$$

Then we can be write:

$$\begin{aligned}H^{(i)} = \nabla^2 l_i(\mathbf{w}) &= -x^{(i)} x^{(i)^T} \frac{1}{1 + \exp(pred^{(i)})} \frac{\exp(pred^{(i)})}{1 + \exp(pred^{(i)})} \\ &= -x^{(i)} x^{(i)^T} \sigma(pred^{(i)})(1 - \sigma(pred^{(i)}))\end{aligned} \tag{4}$$

so the Hessian H is written by:

$$H = -\mathbf{XRX^T} \tag{5}$$

where R is the diagnal matrix that the diagnal elements are $\sigma(pred^{(i)})(1-\sigma(pred^{(i)}))$. Thus,

$$\begin{aligned}\mathbf{z^T} H \mathbf{z} &= -\mathbf{zXRX^T z^T} \\ &= -||\mathbf{z^T RX}||^2 \leq 0.\end{aligned} \tag{6}$$

So it is shown that Hessian H is negative semi-definite and thus l is concave and has no local maxima other than the global one.

Problem 1 continued on next page. . .

Part b:

illustrates training data(blue), the prediction points using the BGD(green) and the SGD(orange), respectively. It is known from the figure that the fit of both methods is good and close. Fig. illustrates the mean squared error (E_{MS}) curves of the BGD and the SGD. The convergence speed of the two methods is very close (so I draw the curves separately), and they both converge at around $epoch = 50$, and converge to $E_{MS} = 0.2$. In theory, the SGD will converge faster. In problem1, it may be hard to distinguish the convergence speed of the two methods because the training set is too small.

Part c:

llustrates the trend of E_{RMS} changing with degree. It is easy to know from the figure that 0, 1, 2, 3 degree polynomials under-fitting the date and 9 degree polynomial over-fitting the data. I think 5 degree best fits the date because the Root-Mean-Square Error (E_{RMS}) of 5 degree polynomial function is relatively smaller and it needs relatively less calculations.

Part c(i):

The closed form solution of the ridge regression is:

$$W_{ML} = (\mathbf{\Phi^T \Phi} + \lambda \mathbf{I})^{-1} \mathbf{\Phi^T y} \tag{7}$$

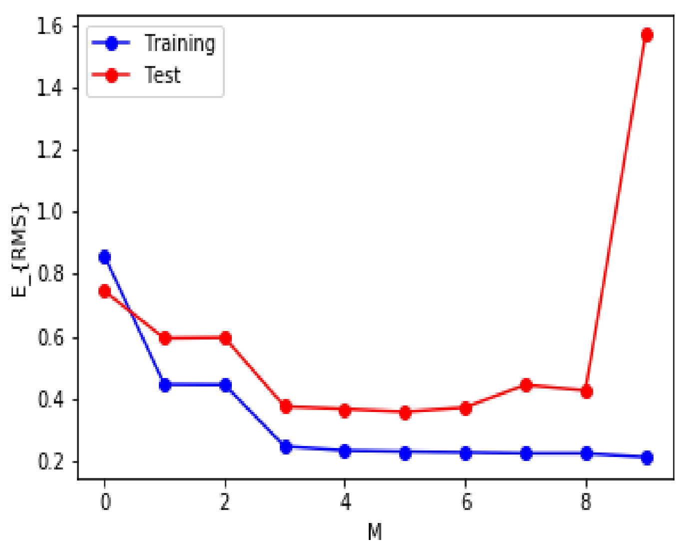

Figure 1: The trend of E_{RMS} changing with regulization factor λ using closed form solution

Part c(ii):

As shown in Fig. 2, the closed form solution reaches the lowest test E_{RMS} at $\lambda = 10^{-4}$, so $\lambda = 10^{-4}$ seemed to work the best.

Problem 2

Softmax Regression via Gradient Ascent

Solution
Part a:

$$\nabla_{\mathbf{w_m}} l(\mathbf{w}) = \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)}) \left[\mathbf{I}(y^{(i)} = m) - \frac{\exp(\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))} \right], \tag{8}$$

and we know :

$$\begin{aligned} p(y = k|\mathbf{x}, \mathbf{w}) &= \frac{\exp(\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))}, \\ l(\mathbf{w}) &= \sum_{i=1}^{N} \sum_{k=1}^{K} \log(\left[p(y^{(i)} = k|\mathbf{x}^{(i)}, \mathbf{w}) \right]^{\mathbf{I}(y^{(i)=k})}). \end{aligned} \tag{9}$$

with (8) and (9), we can get:

$$\begin{aligned}
\nabla_{\mathbf{w_m}} l(\mathbf{w}) &= \nabla_{\mathbf{w}} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})) \right]) && (10) \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})) \right]) && (11) \\
&= \sum_{i=1}^{N} \Big(\nabla_{\mathbf{w}} \sum_{k \neq m}^{K} \mathbf{I}(y^{(i)} = k) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})) \right]) && (12) \\
&+ \nabla_{\mathbf{w}} \mathbf{I}(y^{(i)} = m) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})) \right]) \Big) && (13) \\
&= \sum_{i=1}^{N} \Big(- \nabla_{\mathbf{w}} \sum_{k \neq m}^{K} \mathbf{I}(y^{(i)} = k) \left[\log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})) \right]) && (14) \\
&+ \mathbf{I}(y^{(i)} = m) \left[\phi(\mathbf{x}^{(i)}) - \nabla_{\mathbf{w}} \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})) \right]) \Big) && (15) \\
&= \sum_{i=1}^{N} \Big(\mathbf{I}(y^{(i)} = m) \phi(\mathbf{x}^{(i)}) - \nabla_{\mathbf{w}} \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \Big) && (16) \\
&= \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)}) \Big(\mathbf{I}(y^{(i)} = m) - \nabla_{\mathbf{w}} \frac{\exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))}{(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})))} \Big) && (17) \\
&= \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)}) \Big(\mathbf{I}(y^{(i)} = m) - p(y^{(i)} = m|\mathbf{x}^{(i)}, \mathbf{w}) \Big) && (18)
\end{aligned}$$

Problem 2 continued on next page. . .

Part b:

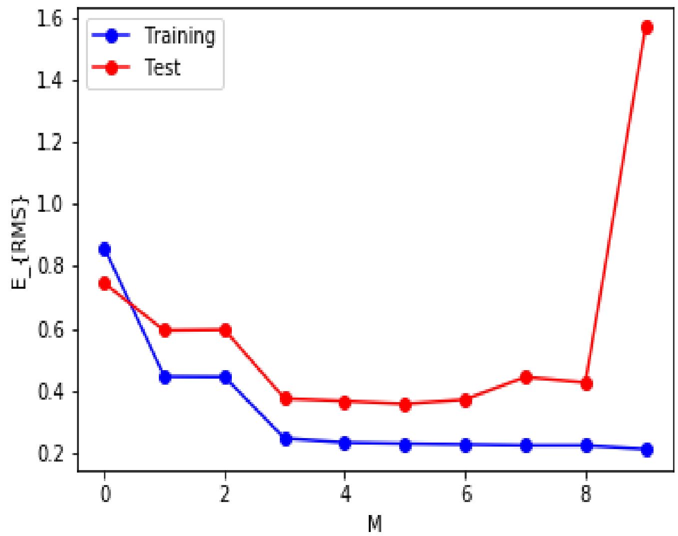

Figure 2: The trend of E_{RMS} changing with regulization factor λ using closed form solution

Problem 3

Gaussian Discriminate Analysis

Solution
Part a:

$$\begin{aligned}
p(y=1|\mathbf{x}^{(i)}) &= \frac{p(\mathbf{x}^{(i)}|y=1)p(y=1)}{p(\mathbf{x}^{(i)})} \\
&= \frac{p(\mathbf{x}^{(i)}|y=1)p(y=1)}{p(\mathbf{x}^{(i)}|y=1)p(y=1)+p(\mathbf{x}^{(i)}|y=0)p(y=0)} \\
&= \frac{\frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_1))\phi}{\frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_1))\phi+\frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_0))(1-\phi)} \\
&= \frac{\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_1))\phi}{\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_1))\phi+\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_0))(1-\phi)}
\end{aligned} \tag{19}$$

Then, we can assum that:

$$\begin{aligned}
\log\frac{p(y=1|\mathbf{x}^{(i)})}{p(y=0|\mathbf{x}^{(i)})} &= \log\frac{\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_1))}{\exp(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_0))}+\log\frac{p(y=1)}{p(y=0)} \\
&= (-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_1))-(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)}-\mu_0))
\end{aligned} \tag{20}$$

The sum squared error:

$$L=\frac{1}{2}\sum_{i=1}^{N}(y^{(i)}-h(x^{(i)}))^2 \tag{21}$$

Problem 3 continued on next page. . .

From eq(18) and eq(21), we can get the partial derivation of L:

$$
\begin{aligned}
\frac{\partial L}{\partial \omega_0} &= \frac{\partial \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \omega_1 x^{(i)} - \omega_0)^2}{\partial \omega_0} \\
&= -\sum_{i=1}^{N}(y^{(i)} - \omega_1 x^{(i)} - \omega_0) \\
&= N\omega_0 + \omega_1 \sum_{i=1}^{N} x^{(i)} - \sum_{i=1}^{N} y^{(i)} \\
&= N(\omega_0 + \omega_1 \bar{X} - \bar{Y}) \\
\frac{\partial L}{\partial \omega_1} &= \frac{\partial \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \omega_1 x^{(i)} - \omega_0)^2}{\partial \omega_1} \\
&= \sum_{i=1}^{N} x^{(i)}(y^{(i)} - \omega_1 x^{(i)} - \omega_0) \\
&= \sum_{i=1}^{N}(x^{(i)}y^{(i)} - \omega_1 {x^{(i)}}^2 - \omega_0 x^{(i)}) \\
&= \sum_{i=1}^{N}(x^{(i)}y^{(i)} - \omega_1 {x^{(i)}}^2) - N\omega_0 \bar{X}
\end{aligned} \tag{22}
$$

where $\bar{X}$ is the mean of $\left\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\right\}$ and $\bar{Y}$ is the mean of $\left\{y^{(1)}, y^{(2)}, \cdots, y^{(N)}\right\}$. Let the partial derivation of L be equal to zero, respectively, the solution for ω_0 and ω_1 for this 1D case of linear regression is derived as follow:

$$
\begin{aligned}
&\frac{\partial L}{\partial \omega_0} = N(\omega_0 + \omega_1 \bar{X} - \bar{Y}) = 0 \\
&\Rightarrow \omega_0 = \bar{Y} - \omega_1 \bar{X}
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \omega_1} &= \sum_{i=1}^{N}(x^{(i)}y^{(i)} - \omega_1 {x^{(i)}}^2) - N\omega_0 \bar{X} \\
&= \sum_{i=1}^{N}(x^{(i)}y^{(i)} - \omega_1 {x^{(i)}}^2) - N(\bar{Y} - \omega_1 \bar{X})\bar{X} \\
&= \sum_{i=1}^{N}(x^{(i)}y^{(i)} - \omega_1 {x^{(i)}}^2) - N\bar{Y}\bar{X} - N\omega_1 \bar{X}^2 = 0 \\
\Rightarrow \omega_1 &= \frac{\sum_{i=1}^{N} x^{(i)}y^{(i)} - N\bar{Y}\bar{X}}{\sum_{i=1}^{N} {x^{(i)}}^2 - N\bar{X}^2} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}y^{(i)} - \bar{Y}\bar{X}}{\frac{1}{N}\sum_{i=1}^{N} {x^{(i)}}^2 - \bar{X}^2}
\end{aligned} \tag{24}
$$

Part b(i):

Problem 3 continued on next page...

$$\begin{aligned}\mathbf{A} &= \mathbf{U\Lambda U^T} \\ &= \sum_{i=1}^{d} \lambda_i u_i u_i^T\end{aligned} \tag{25}$$

$$\begin{aligned}\mathbf{z^T A z} &= \mathbf{z^T U \Lambda U^T z} \\ &= \sum_{i=1}^{d} \lambda_i z^T u_i u_i^T z \\ &= \sum_{i=1}^{d} \lambda_i \|u_i^T z\|_2^2, (z \neq 0)\end{aligned} \tag{26}$$

It is obviously that $\sum_{i=1}^{d} \lambda_i \|u_i^T z\|_2^2 > 0$ iff $\lambda_i > 0$ for each i. So, $\mathbf{A}$ is PD iff $\lambda_i > 0$ for each i.

Part b(ii):

The matrix $\mathbf{\Phi^T \Phi}$ is real and symmetric, so it can be expressed by $\mathbf{\Phi^T \Phi} = \mathbf{U \Lambda U^T}$, so

$$\begin{aligned}\mathbf{z^T \Phi^T \Phi z} &= \mathbf{z^T U \Lambda U^T z} \\ &= \sum_{i=1}^{d} \lambda_i^2 z^T u_i u_i^T z \\ &= \sum_{i=1}^{d} \lambda_i^2 \|u_i^T z\|_2^2 \geq 0, (z \neq 0)\end{aligned} \tag{27}$$

$$\begin{aligned}\mathbf{z^T (\Phi^T \Phi} + \beta \mathbf{I) z} &= \mathbf{z^T U (\Lambda} + \beta) \mathbf{U^T z} \\ &= \sum_{i=1}^{d} (\lambda_i^2 + \beta) z^T u_i u_i^T z \\ &= \sum_{i=1}^{d} (\lambda_i^2 + \beta) \|u_i^T z\|_2^2 > 0, (z \neq 0),\end{aligned} \tag{28}$$

because $(\lambda_i^2 + \beta)$ always larger than zero. Hence, for any $\beta > 0$, ridge regression makes the matrix $\mathbf{\Phi^T \Phi} + \beta \mathbf{I}$ PD.