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 $\mathbf{H}$

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

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:

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

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

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

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.

2

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:

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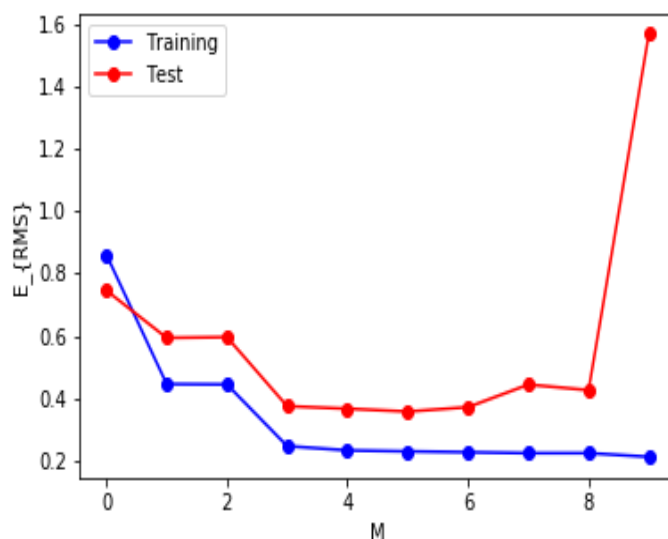

Figure 1: The trend of E_{RMS} changing with regulization factor λ using closed form solution

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 :

$$p(y = k | \mathbf{x}, \mathbf{w}) = \frac{\exp(\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))},$$

$$l(\mathbf{w}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \log\left(\left[p(y^{(i)} = k | \mathbf{x}^{(i)}, \mathbf{w}) \right]^{\mathbf{I}(y^{(i)}=k)} \right). \tag{9}$$

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

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

$$= \sum_{i=1}^{N} \nabla_{\mathbf{w}} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \right])$$

$$= \sum_{i=1}^{N} \left(\nabla_{\mathbf{w}} \sum_{k \neq m}^{K} \mathbf{I}(y^{(i)} = k) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \right] \right)$$

$$+ \nabla_{\mathbf{w}} \mathbf{I}(y^{(i)} = m) \left[\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}) - \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \right])$$

$$= \sum_{i=1}^{N} \left(-\nabla_{\mathbf{w}} \sum_{k \neq m}^{K} \mathbf{I}(y^{(i)} = k) \left[\log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \right] \right) \tag{10}$$

$$+ \mathbf{I}(y^{(i)} = m) \left[\phi(\mathbf{x}^{(i)}) - \nabla_{\mathbf{w}} \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \right])$$

$$= \sum_{i=1}^{N} \left(\mathbf{I}(y^{(i)} = m) \phi(\mathbf{x}^{(i)}) - \nabla_{\mathbf{w}} \log(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))) \right)$$

$$= \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)}) \left(\mathbf{I}(y^{(i)} = m) - \nabla_{\mathbf{w}} \frac{\exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)}))}{(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi((\mathbf{x})^{(i)})))} \right)$$

$$= \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)}) \left(\mathbf{I}(y^{(i)} = m) - p(y^{(i)} = m | \mathbf{x}^{(i)}, \mathbf{w}) \right)$$
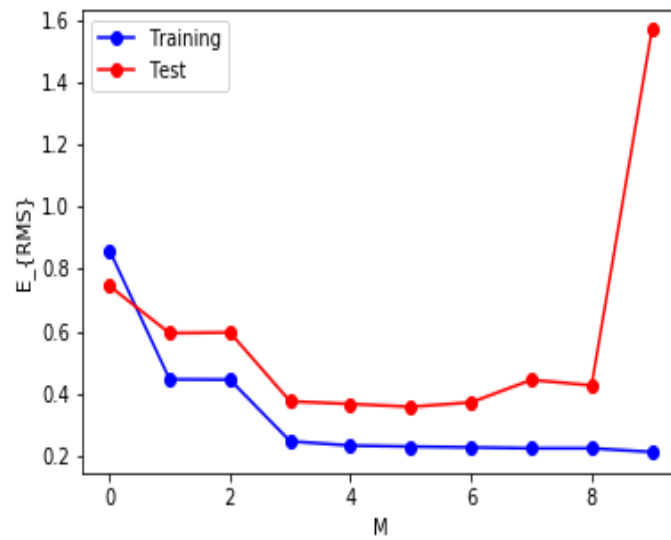
4

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:

$$
p(y = 1|\mathbf{x}^{(i)}) = \frac{p(\mathbf{x}^{(i)}|y = 1)p(y = 1)}{p(\mathbf{x}^{(i)})}
$$

$$
= \frac{p(\mathbf{x}^{(i)}|y = 1)p(y = 1)}{p(\mathbf{x}^{(i)}|y = 1)p(y = 1) + p(\mathbf{x}^{(i)}|y = 0)p(y = 0)}
$$

$$
= \frac{\frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_1))\phi}{\frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_1))\phi + \frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0))(1 - \phi)}
$$

$$
= \frac{\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_1))\phi}{\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_1))\phi + \exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0))(1 - \phi)}
$$

$$
= \frac{1}{1 + \frac{\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_1))\phi}{\exp(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0))(1 - \phi)}}
\tag{11}
$$

Then, we can assum that:

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

$$
= (-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_1)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_1)) - (-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T\Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0)) + \log\frac{p(y = 1)}{p(y = 0)}
$$

$$
= (\mu_1 - \mu_0)^T\Sigma^{-1}\mathbf{x}^{(i)} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log\frac{\phi}{1 - \phi}
\tag{12}
$$

where $(\mu_1 - \mu_0)^T\Sigma^{-1}\mathbf{x}^{(i)}$ is W_T and $-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log\frac{\phi}{1-\phi}$ is W_0. Thus, with (11) and (12) we will get:

$$
p(y = 1|\mathbf{x}^{(i)}) = \frac{1}{1 + \exp(-\log\frac{p(y=1|\mathbf{x}^{(i)})}{p(y=0|\mathbf{x}^{(i)})})}
$$

$$
= \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}^{(i)})}
\tag{13}
$$

so the posterior distribution of the label (y) at $\mathbf{x}$ takes the form of a logistic function, and can be written as:

$$
p(y = 1|\mathbf{x}; \phi, \Sigma, \mu_0.\mu_1) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}
\tag{14}
$$

6

Part b:

Considering that b is a special case of c, we will directly consider M both b and c here. The log-likelihood of the data is:

$$
\begin{aligned}
l(\phi, \mu, \Sigma) &= \log(\Pi_{i=1}^{N} p(\mathbf{x}^{(i)}, y^{(i)}; \phi, \mu, \Sigma)) \\
&= \log(\Pi_{i=1}^{N} p(\mathbf{x}^{(i)} | y^{(i)}; \phi, \mu, \Sigma) p(y^{(i)}; \phi)) \\
&= \log(\Pi_{i=1}^{N} p(\mathbf{x}^{(i)} | y^{(i)}; \phi, \mu, \Sigma)) + \log \Pi_{i=1}^{N} p(y^{(i)}; \phi) \\
&= \sum_{i=1}^{N} (\log(\frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}}) - \frac{1}{2}(\mathbf{x}^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1}(\mathbf{x}^{(i)} - \mu_{y^{(i)}}) \\
&\quad + y^{(i)} \log \phi + (1 - y^{(i)}) \log(1 - \phi))
\end{aligned}
\tag{15}
$$

Then take the partial derivative of ϕ to function l:

$$
\begin{aligned}
\frac{\partial l}{\partial \phi} &= \sum_{i=1}^{N} [\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi}] \\
&= \frac{\sum_{i=1}^{N} y^{(i)}}{\phi} - \frac{N - \sum_{i=1}^{N} y^{(i)}}{1 - \phi},
\end{aligned}
\tag{16}
$$

let the partial derivative equal to zero:

$$
\begin{aligned}
\frac{\sum_{i=1}^{N} y^{(i)}}{\phi} - \frac{N - \sum_{i=1}^{N} y^{(i)}}{1 - \phi} &= 0 \\
\Rightarrow \frac{\sum_{i=1}^{N} y^{(i)}}{\phi} &= \frac{N - \sum_{i=1}^{N} y^{(i)}}{1 - \phi} \\
\Rightarrow \phi &= \frac{1}{N} \sum_{i=1}^{N} 1\left\{y^{(i)} = 1\right\}
\end{aligned}
\tag{17}
$$

Then take the partial derivative of μ_0 to function l:

$$
\begin{aligned}
\nabla_{\mu_0} l &= \nabla_{\mu_0} [\sum_{i: y^{(i)} = 0} -\frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T \Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0)] \\
&= \sum_{i: y^{(i)} = 0} [\Sigma^{-1} \mathbf{x}^{(i)} - \Sigma^{-1} \mu_0],
\end{aligned}
\tag{18}
$$

let the partial derivative equal to zero:

$$
\mu_0 = \frac{\sum_{i=1}^{n} 1\left\{y^{(i)} = 0\right\} \mathbf{x}^{(i)}}{\sum_{i=1}^{N} 1\left\{y^{(i)} = 0\right\}}.
\tag{19}
$$

Similarly, μ_1 can be written as:

$$
\mu_1 = \frac{\sum_{i=1}^{n} 1\left\{y^{(i)} = 1\right\} \mathbf{x}^{(i)}}{\sum_{i=1}^{N} 1\left\{y^{(i)} = 1\right\}}.
\tag{20}
$$

Then take the partial derivative of Σ to the function l, and for this situation, we only consider M equals to 1:

$$
\begin{aligned}
\nabla_\Sigma l &= \nabla_\Sigma [\sum_{i=1}^N \log(\frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}}) - \frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T \Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0)] \\
&= \nabla_\Sigma [\sum_{i=1}^N \log(\frac{1}{(2\pi)^{\frac{M}{2}}}) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)^T \Sigma^{-1}(\mathbf{x}^{(i)} - \mu_0)] \qquad (21) \\
&= \sum_{i=1}^N [-\frac{1}{2\Sigma} + \frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)(\mathbf{x}^{(i)} - \mu_0)^T \frac{1}{\Sigma^2}],
\end{aligned}
$$

let the partial derivative equal to zero:

$$
\begin{aligned}
&\sum_{i=1}^N [-\frac{1}{2\Sigma} + \frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)(\mathbf{x}^{(i)} - \mu_0)^T \frac{1}{\Sigma^2}] = 0 \\
&\Rightarrow \sum_{i=1}^N [\frac{1}{2}(\mathbf{x}^{(i)} - \mu_0)(\mathbf{x}^{(i)} - \mu_0)^T \frac{1}{\Sigma^2}] = \sum_{i=1}^N \frac{1}{2\Sigma} = \frac{N}{2\Sigma} \qquad (22) \\
&\Rightarrow \Sigma = \frac{1}{N}\sum_{i=1}^N (\mathbf{x}^{(i)} - \mu_0)(\mathbf{x}^{(i)} - \mu_0)^T.
\end{aligned}
$$

Part c:
For ϕ, it is the same as b. For μ, it is also similar in b,

$$
\mu_t = \frac{\sum_{i=1}^n 1\{y^{(i)} = t\}\mathbf{x}^{(i)}}{\sum_{i=1}^N 1\{y^{(i)} = t\}}. \qquad (23)
$$

Problem 4

Naive Bayes for classifying SPAM

Solution
Part a: