CS189 HW4

David Chen

February 2022

Contents

1 Write-up and Honor Code

Collaborated with: N/A.

I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted.

Signed: David Chen

2 Logistic Regression with Newton's Method

2.1 Part 1

First note that the derivative of the logistic function is given by

$$s'(y) = s(y)(1 - s(y)).$$

Therefore, we can use this result to get

$$\begin{aligned}
\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}} - \boldsymbol{y} \cdot \ln \boldsymbol{s} - (\mathbf{1} - \boldsymbol{y}) \cdot \ln(\mathbf{1} - \boldsymbol{y}) \\
&= \nabla_{\boldsymbol{w}} - \sum y_i \ln s_i + (1 - y_i) \ln(1 - s_i) \\
&= -\sum (\nabla_{\boldsymbol{w}} \ln s_i) y_i + (\nabla_{\boldsymbol{w}} \ln(1 - s_i))(1 - y_i) \\
&= -\sum \frac{1}{s_i} (\nabla_{\boldsymbol{w}} s_i) y_i - \frac{1}{1 - s_i} (\nabla_{\boldsymbol{w}} s_i)(1 - y_i) \\
&= -\sum \frac{1}{s_i} (\nabla_{\boldsymbol{w}} s(\boldsymbol{x}_i \cdot \boldsymbol{w})) y_i - \frac{1}{1 - s_i} (\nabla_{\boldsymbol{w}} s(\boldsymbol{x}_i \cdot \boldsymbol{w}))(1 - y_i) \\
&= -\sum \frac{\boldsymbol{x}_i}{s_i} (s_i') y_i - \frac{\boldsymbol{x}_i}{1 - s_i} (s_i')(1 - y_i) \\
&= -\sum \boldsymbol{x}_i \left(\frac{1}{s_i} s_i (1 - s_i) y_i - \frac{1}{1 - s_i} s_i (1 - s_i)(1 - y_i) \right) \\
&= -\sum \boldsymbol{x}_i ((1 - s_i) y_i - s_i (1 - y_i)) \\
&= -\sum \boldsymbol{x}_i (y_i - s_i y_i - s_i + s_i y_i) \\
&= -\sum \boldsymbol{x}_i (y_i - s_i) \\
&= -X^T (\boldsymbol{y} - \boldsymbol{s}).
\end{aligned}$$

2.2 Part 2

$$\begin{aligned}
\nabla^2_{\boldsymbol{w}} J(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}} \left(\nabla_{\boldsymbol{w}} J(\boldsymbol{w})\right) \\
&= \nabla_{\boldsymbol{w}} \left(-X^T (\boldsymbol{y} - \boldsymbol{s})\right) \\
&= \nabla_{\boldsymbol{w}} X^T \boldsymbol{s} \\
&= \sum \nabla_{\boldsymbol{w}} {X_i}^T s_i \\
&= \sum \left(\nabla_{\boldsymbol{w}} s_i\right) {X_i}^T \\
&= \sum s_i (1 - s_i) X_i {X_i}^T \\
&= X^T \Omega X
\end{aligned}$$

where diagonal matrix Ω is given by

$$\Omega = \begin{bmatrix} s_1 (1 - s_1) & 0 & \dots & 0 \\ 0 & s_2 (1 - s_2) & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & s_n (1 - s_n) \end{bmatrix}.$$

2.3 Part 3

Recall that Newton's method updates using vector $\boldsymbol{e}$, which is the solution to the linear system $\left(\nabla^2_{\boldsymbol{w}} J\left(\boldsymbol{w}\right)\right) \boldsymbol{e} = -\nabla_{\boldsymbol{w}} J\left(\boldsymbol{w}\right)$. Therefore, for logistic regression, this becomes

$$\begin{aligned}
\boldsymbol{w} &\leftarrow \boldsymbol{w} + \boldsymbol{e} \\
&\leftarrow \boldsymbol{w} + \left(X^T \Omega X\right)^{-1} X^T \left(\boldsymbol{y} - \boldsymbol{s}\right).
\end{aligned}$$

2.4 Part 4

2.4.1 Section (a)

Design matrix X is given by

$$X = \begin{bmatrix} 0.2 & 3.1 & 1 \\ 1.0 & 3.0 & 1 \\ -0.2 & 1.2 & 1 \\ 1.0 & 1.1 & 1 \end{bmatrix}.$$

Therefore,

$$\begin{aligned} X\boldsymbol{w}^{(0)} &= \begin{bmatrix} 0.2 & 3.1 & 1 \\ 1.0 & 3.0 & 1 \\ -0.2 & 1.2 & 1 \\ 1.0 & 1.1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 2.9 \\ 2.0 \\ 1.4 \\ 0.1 \end{bmatrix}. \end{aligned}$$

$\boldsymbol{s}^{(0)}$ is equal to the logistic function applied element-wise to $X\boldsymbol{w}^{(0)}$, which gives

$$\begin{aligned} \boldsymbol{s}^{(0)} &= \begin{bmatrix} s\,(2.9) \\ s\,(2.0) \\ s\,(1.4) \\ s\,(0.1) \end{bmatrix} \\ &= \begin{bmatrix} 0.9478 \\ 0.8808 \\ 0.8022 \\ 0.5250 \end{bmatrix}. \end{aligned}$$

2.4.2 Section (b)

We first find $\boldsymbol{e}$ with

$$\begin{aligned} \boldsymbol{e} &= \left(X^T \Omega X\right)^{-1} X^T \left(\boldsymbol{y} - \boldsymbol{s}\right) \\ &= \begin{bmatrix} 0.0953 \\ 0.5623 \\ -1.6783 \end{bmatrix}. \end{aligned}$$

Following the update rule, we get

$$
\begin{aligned}
\boldsymbol{w}^{(1)} &= \boldsymbol{w}^{(0)} + \boldsymbol{e} \\
&= \begin{bmatrix} -0.9047 & 1.5623 & -1.6783 \end{bmatrix}^T.
\end{aligned}
$$

2.4.3 Section (c)

Following the same procedure as above gives

$$
\begin{aligned}
X\boldsymbol{w}^{(1)} &= \begin{bmatrix} 2.9839 \\ 2.1039 \\ 0.3774 \\ -0.8645 \end{bmatrix} \\
\therefore \boldsymbol{s}^{(1)} &= \begin{bmatrix} s\,(2.9839) \\ s\,(2.1039) \\ s\,(0.3774) \\ s\,(-0.8645) \end{bmatrix} \\
&= \begin{bmatrix} 0.9518 \\ 0.8913 \\ 0.5932 \\ 0.2964 \end{bmatrix}.
\end{aligned}
$$

2.4.4 Section (d)

Following the same procedure as above gives

$$
\begin{aligned}
\boldsymbol{e} &= \left(X^T \Omega X\right)^{-1} X^T \left(\boldsymbol{y} - \boldsymbol{s}\right) \\
&= \begin{bmatrix} 0.1614 \\ 0.3733 \\ -1.1480 \end{bmatrix} \\
\therefore \boldsymbol{w}^{(2)} &= \boldsymbol{w}^{(1)} + \boldsymbol{e} \\
&= \begin{bmatrix} -0.7433 & 1.9356 & -2.8263 \end{bmatrix}^T.
\end{aligned}
$$

3 Wine Classification with Logistic Regression

3.1 Part 1

The only difference between logistic regression and logistic regression l_2 regularization is the added $\lambda\|\boldsymbol{w}\|^2$ term to the cost function. This gives

$$\begin{aligned}
J(\boldsymbol{w}) &= -\boldsymbol{y}\cdot\ln\boldsymbol{s} - (\mathbf{1}-\boldsymbol{y})\cdot\ln(\mathbf{1}-\boldsymbol{s}) + \lambda\|\boldsymbol{w}\|^2 \\
\therefore \nabla_{\boldsymbol{w}} J(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}}\left(-\boldsymbol{y}\cdot\ln\boldsymbol{s} - (\mathbf{1}-\boldsymbol{y})\cdot\ln(\mathbf{1}-\boldsymbol{s})\right) + \nabla_{\boldsymbol{w}}\left(\lambda\|\boldsymbol{w}\|^2\right) \\
&= -X^T(\boldsymbol{y}-\boldsymbol{s}) + 2\lambda\boldsymbol{w} \\
\therefore \boldsymbol{w} &\leftarrow \boldsymbol{w} - \varepsilon\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) \\
&\leftarrow \boldsymbol{w} - \varepsilon\left(-X^T(\boldsymbol{y}-\boldsymbol{s}) + 2\lambda\boldsymbol{w}\right) \\
&\leftarrow \boldsymbol{w} + \varepsilon\left(X^T(\boldsymbol{y}-\boldsymbol{s}) - 2\lambda\boldsymbol{w}\right).
\end{aligned}$$

3.2 Part 2

Included in attached Python Notebook.

3.3 Part 3

Since stochastic gradient descent works with one training example rather than all, it uses a similar update

$$\begin{aligned} w &\leftarrow w - \varepsilon\left(-\boldsymbol{X}_i^T\left(y_i - s\left(\boldsymbol{X}_i \cdot \boldsymbol{w}\right)\right) + 2\lambda\boldsymbol{w}\right) \\ &\leftarrow w + \varepsilon\left(\boldsymbol{X}_i^T\left(y_i - s\left(\boldsymbol{X}_i \cdot \boldsymbol{w}\right)\right) - 2\lambda\boldsymbol{w}\right). \end{aligned}$$

3.4 Part 4

Included in attached Python Notebook.

3.5 Part 5

Included in attached Python Notebook.

3.6 Part 6

Included in attached Python Notebook.

4 A Bayesian Interpretation of Lasso

4.1 Part 1

By Bayes' Theorem,

$$f\left(\boldsymbol{w}|(\boldsymbol{x}_i, y_i)_{i\in[n]}\right) = \frac{f\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}\right) \cdot f\left(\boldsymbol{w}|\boldsymbol{x}_i\right)}{f\left(y_i\right)}.$$

Since $\boldsymbol{w}$ is a random parameter, $f\left(\boldsymbol{w}|\boldsymbol{x}_i\right) = f\left(\boldsymbol{w}\right)$. Therefore,

$$f\left(\boldsymbol{w}|(\boldsymbol{x}_i, y_i)_{i\in[n]}\right) = \frac{f\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}\right) \cdot f\left(\boldsymbol{w}\right)}{f\left(y_i\right)},$$

where $f\left(y_i\right)$ is the maximum likelihood estimate of y_i attained by counting the amount of y_i's and dividing by the total length of y.

4.2 Part 2

$$
\begin{aligned}
\max_{\boldsymbol{w}} l\left(\boldsymbol{w}\right) &= \max_{\boldsymbol{w}} \ln f\left(\boldsymbol{w}|(\boldsymbol{x}_i, y_i)_{i\in[n]}\right)\\
&= \max_{\boldsymbol{w}} \ln \frac{f\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}\right)\cdot f\left(\boldsymbol{w}\right)}{f\left(y_i\right)}\\
&= \max_{\boldsymbol{w}} \ln\left(\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}\right)\right) + \ln\left(f\left(\boldsymbol{w}\right)\right) - \ln\left(f\left(y_i\right)\right)\\
&= \max_{\boldsymbol{w}} \ln\left(\left(y_i|\boldsymbol{x}_i, \boldsymbol{w}\right)\right) + \ln\left(f\left(\boldsymbol{w}\right)\right)\\
&= \max_{\boldsymbol{w}} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\boldsymbol{w}\cdot\boldsymbol{x})^2}{2\sigma^2}}\right) + \ln\left(\frac{1}{2b}e^{-\frac{\|\boldsymbol{w}\|_1}{b}}\right)\\
&= \max_{\boldsymbol{w}} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\left(-\frac{\left(y-\boldsymbol{w}\cdot\boldsymbol{x}\right)^2}{2\sigma^2}\right) + \ln\left(\frac{1}{2b}\right)\left(-\frac{\|\boldsymbol{w}\|_1}{b}\right)\\
&= \max_{\boldsymbol{w}} -\left(y-\boldsymbol{w}\cdot\boldsymbol{x}\right)^2 - \lambda\|\boldsymbol{w}\|_1\\
&= \min_{\boldsymbol{w}} \left(y-\boldsymbol{w}\cdot\boldsymbol{x}\right)^2 + \lambda\|\boldsymbol{w}\|_1,
\end{aligned}
$$

where $\lambda = \frac{\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}{\ln\left(\frac{1}{2b}\right)}\frac{2\sigma^2}{b}$.

5 l_1-regularization, l_2-regularization, and Sparsity

5.1 Part 1

Included in attached Python Notebook.

5.2 Part 2

Proof.

$$
\begin{aligned}
J_1(\boldsymbol{w}) &= \|X\boldsymbol{w}-\boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|_1 \\
&= (X\boldsymbol{w}-\boldsymbol{y})^T(X\boldsymbol{w}-\boldsymbol{y}) + \lambda\sum_i w_i \\
&= \left(\boldsymbol{w}^T X^T - \boldsymbol{y}^T\right)(X\boldsymbol{w}-\boldsymbol{y}) + \lambda\sum_i w_i \\
&= \|\boldsymbol{y}\|^2 + \|X\boldsymbol{w}\|^2 - 2\boldsymbol{y}^T X\boldsymbol{w} + \lambda\sum_i w_i \\
&= \|\boldsymbol{y}\|^2 + \sum_i \left(\boldsymbol{w}_i{}^2\|\boldsymbol{x}_{*i}\|^2 - 2\boldsymbol{y}^T\boldsymbol{x}_{*i}\boldsymbol{w}_i + \lambda\boldsymbol{w}_i\right) \\
&= \|\boldsymbol{y}\|^2 + \sum_i \left(\boldsymbol{w}_i{}^2\|nI_{*i}\|^2 - 2\boldsymbol{y}^T\boldsymbol{x}_{*i}\boldsymbol{w}_i + \lambda\boldsymbol{w}_i\right) \\
&= \|\boldsymbol{y}\|^2 + \sum_i \left(\boldsymbol{w}_i{}^2 n^2 - 2\boldsymbol{y}^T\boldsymbol{x}_{*i}\boldsymbol{w}_i + \lambda\boldsymbol{w}_i\right) \\
&= \|\boldsymbol{y}\|^2 + \sum_i f(\boldsymbol{x}_{*i}, \boldsymbol{w}_i).
\end{aligned}
$$

$\square$

5.3 Part 3

$$\begin{aligned}\frac{\partial f\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)}{\partial \boldsymbol{w}_i} &= \frac{\partial}{\partial \boldsymbol{w}_i}\left({\boldsymbol{w}_i}^2 n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i}\boldsymbol{w}_i + \lambda \boldsymbol{w}_i\right)\\ &= 2\boldsymbol{w}_i n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i} + \lambda\\ \therefore \boldsymbol{w}_i^* &= \frac{2\boldsymbol{y}^T \boldsymbol{x}_{*i} - \lambda}{2n^2}.\end{aligned}$$

This gives that the sign of $\boldsymbol{w}_i^*$ is equal to the sign of $2\boldsymbol{y}^T \boldsymbol{x}_{*i} - \lambda$, as the denominator is always guaranteed to be positive. In addition, since

$$\begin{aligned}\frac{\partial f^2\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)}{\partial \boldsymbol{w}_i^2} &= \frac{\partial}{\partial \boldsymbol{x}_{*i}}\left(2\boldsymbol{w}_i n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i} + \lambda\right)\\ &= 2n^2\\ &> 0,\end{aligned}$$

$f\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)$ is a convex function and $\boldsymbol{w}_i^*$ is guaranteed to be a minimum.

5.4 Part 4

Changing from $l1$-regularization to $l2$-regularization changes $f\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)$ to

$$f\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right) = {\boldsymbol{w}_i}^2 n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i} \boldsymbol{w}_i + \lambda \boldsymbol{w}_i^2.$$

Therefore, the first derivative becomes

$$\begin{aligned}
\frac{\partial f\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)}{\partial \boldsymbol{w}_i} &= \frac{\partial}{\partial \boldsymbol{w}_i}\left({\boldsymbol{w}_i}^2 n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i} \boldsymbol{w}_i + \lambda \boldsymbol{w}_i^2\right) \\
&= 2\boldsymbol{w}_i n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i} + 2\lambda \boldsymbol{w}_i \\
\therefore \boldsymbol{w}_i^{\#} &= \frac{\boldsymbol{y}^T \boldsymbol{x}_{*i}}{n^2 + \lambda},
\end{aligned}$$

which implies that $\boldsymbol{w}_i^{\#} = 0$ **if** $\boldsymbol{y}^T \boldsymbol{x} = 0$. Furthermore, this is only guaranteed to hold if $f\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)$ is convex, which occurs when

$$\begin{aligned}
\frac{\partial f^2\left(\boldsymbol{x}_{*i}, \boldsymbol{w}_i\right)}{\partial \boldsymbol{w}_i^2} &= \frac{\partial}{\partial \boldsymbol{x}_{*i}}\left(2\boldsymbol{w}_i n^2 - 2\boldsymbol{y}^T \boldsymbol{x}_{*i} + 2\lambda \boldsymbol{w}_i\right) \\
&= 2n^2 + 2\lambda \\
\therefore n^2 &> -\lambda,
\end{aligned}$$

which we add as another necessary and sufficient condition.

5.5 Part 5

$\boldsymbol{w}^*$ is more likely to be sparse, as it is more unlikely for $\boldsymbol{y}^T \boldsymbol{x}_{*i}$ to be equal to a specific value λ rather than 0. The exception to this is when $\lambda = 0$, which causes $\boldsymbol{w}^*$ and $\boldsymbol{w}^\#$ to have the same sparsity.