1. Understanding Binary Class Logistic Regression.

CSCC11 Assignment 2.
Ti, Zhang
1000424517

1. First, let's investigate the decision boundary of a binary class logistic regression model. Recall that the decision boundary is the set of points where $P(c_1|\mathbf{x}) = 0.5$ (i.e. when the decision function $\alpha(\mathbf{x}) = 0$). Show that when $P(c_1|\mathbf{x}) = 0.5$, the decision boundary $\alpha(\mathbf{x}) = 0$ is a linear function.

$$P(c_1|\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} = g(\mathbf{w}^T\mathbf{x})$$

Known $\begin{cases} P(c_1|x) = \frac{1}{1+e^{-w^Tx}} \\ P(c_1|x) = 0.5. \end{cases}$ $\therefore$ $\begin{cases} 1+e^{-w^Tx} = 2 \\ e^{-w^Tx} = 1 \\ w^Tx = 0 \end{cases}$

$w^Tx=0$, which is the decision boundary, is appearently a linear function. ▨

2. Knowing that logistic regression has a linear decision boundary, what datasets would this model perform poorly on? In this case, performing poorly means the model is unable to classify all the training data correct. Let's look at a toy example – the XOR (Exclusive OR) dataset. XOR is a binary input Boolean function that outputs **TRUE** when both inputs are the same and outputs **FALSE** otherwise. The dataset is represented as the table below. Our goal is to predict the output $y \in \{0, 1\}$ given the input vector $\mathbf{x} = [x_1, x_2]^T$, where $x_1, x_2 \in \{0, 1\}$.

$\mathbf{x}$			
x_1	x_2	y	
0	0	0	T
0	1	1	F
1	0	1	F
1	1	0	T

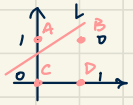
(a) Assume we are using binary class logistic regression on the XOR dataset. What is the maximum classification accuracy we can obtain? Explain.

The classification accuracy can be expressed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

The given example contains all 4 situations we can possibly have.

If we put them in a graph:



All the pink points representing y. (A. B. C. D)

If we put a decision line L, we can get 1 or 3 correct classification out of 4.

So the best I can do is 3 out of 4, A as TN. B as TP. C as TP. D as FN

which means the maximum classification accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

$$= \frac{2+1}{2+1+0+1}$$

$$= 75\%.$$ ▨

CSCC11 Assignment 2.
Ti, Zhang
1000429457

(b) As in basis function regression, we can apply basis functions to create a more sophisticated model. Consider the following feature map (basis functions) on the inputs:

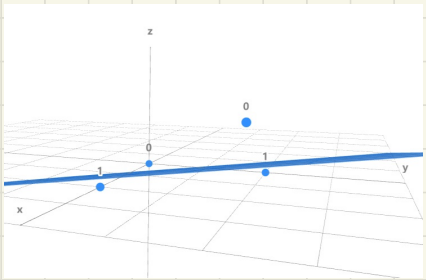
$$\psi(\mathbf{x}) = \begin{pmatrix} \psi_1(\mathbf{x}) \\ \psi_2(\mathbf{x}) \\ \psi_3(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{pmatrix}. \quad (2)$$

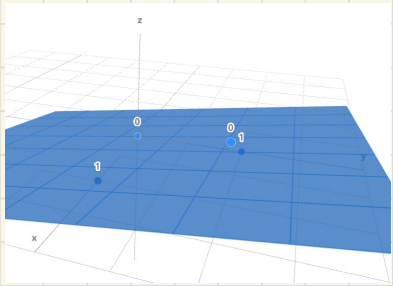
Then, we can express the posterior probability of class c_1 given input $\mathbf{x}$ as $P(c_1|\mathbf{x}) = g(\mathbf{w}^T\psi(\mathbf{x}))$, where $\mathbf{w} = [w_1, w_2, w_3]^T$ is the weight vector of the model. Note that we exclude the bias term in this question. Specify the conditions and provide an example for the weight vector $\mathbf{w}$ such that this model perfectly classifies the XOR dataset.

If we map all possible situation we can get:

	$\psi(x)$			
x_1	0	0	1	1
x_2	0	1	0	1
x_1x_2	0	0	0	1
XOR	0	1	1	0

Express x_1 as x, x_2 as y, x_3 as z and put them in a 3D graph.





It's positive that we can find a surface perfectly classifies the XOR dataset.

$P(c_1|x) = g(w^Tx) = g(w_1x_1 + w_2x_2 + w_3x_1x_2) = 0.5$

As proved in Q1, when $g(w^Tx) = 0.5$, $w^Tx = 0$.

$\therefore w_1x_1 + w_2x_2 + w_3x_1x_2 = 0$

Want to find w_1, w_2, w_3 s.t $g(w^Tx)$ satisfy chart A.

Chart A.

	$\psi(x)$			
x_1	0	0	1	1
x_2	0	1	0	1
x_1x_2	0	0	0	1
XOR	0	1	1	0
$g(w^Tx)$	≤ 0	>0	>0	≤ 0

And this last entry is the condition

Chart B

	$\psi(x)$			
x_1	0	0	1	1
x_2	0	1	0	1
x_1x_2	0	0	0	1
XOR	0	1	1	0
$g(w^Tx)$	0	1	1	-1
	≤ 0	>0	>0	≤ 0

Example $w = \begin{bmatrix} 1 \\ 1 \\ -3 \end{bmatrix}$ Satisfy condition.

Example: $w_1 = w_2 = 1$, $w_3 = -3$ in Chart B. ■

2. Multiclass Logistic Regression.

CSCC11 Assignment 2.
Ti, Zhang
1000424517

1. Find the gradient of $\sigma(z_{1:K}, k)$ in Eqn. (7) with respect to z_j:

$$\frac{\partial \sigma(z_{1:K}, k)}{\partial z_j}. \tag{9}$$

Hint: Consider the cases when $k = j$ and $k \neq j$. You may find it helpful to look at the structure of the gradient in the 2-class case in Eqn. (46) in Chapter 9.6 of the online lecture notes.

Known that $\sigma(z_{1:K}, k) = \frac{e^{z_k}}{\sum_{\ell=1}^{K} e^{z_\ell}}$

$$\therefore \frac{\partial \sigma(z_{1:K}, k)}{\partial z_j} = \frac{\partial \left[e^{z_k} / \sum_{i=1}^{k} e^{z_i} \right]}{\partial z_j} \quad \cdots ①$$

Case 1: $k = j$

$$① = \frac{\partial \left[e^{z_j} / \sum_{i=1}^{j} e^{z_i} \right]}{\partial z_j} = \frac{e^{z_j} \cdot \sum_{i=1}^{j} e^{z_i} - e^{z_j} \cdot e^{z_j}}{\left(\sum_{i=1}^{j} e^{z_i} \right)^2}$$

$$= \frac{e^{z_j} \left(\sum_{i=1}^{j} e^{z_i} - e^{z_j} \right)}{\left(\sum_{i=1}^{j} e^{z_i} \right)^2}$$

$$= \frac{e^{z_j}}{\sum e^{z_i}} \cdot \frac{\left(\sum e^{z_i} - e^{z_j} \right)}{\sum e^{z_i}}$$

$$= \frac{e^{z_j}}{\sum e^{z_i}} \cdot \left(1 - \frac{e^{z_j}}{\sum e^{z_i}} \right)$$

$$= \sigma(z_{1:k}, k) \left[1 - \sigma(z_{1:k}, k) \right]$$

Case 2: $k \neq j$

$$① = \frac{\partial \left[e^{z_k} / \sum_{i=1}^{k} e^{z_i} \right]}{\partial z_j} = \frac{0 - e^{z_k} \cdot e^{z_j}}{\left(\sum_{i=1}^{k} e^{z_i} \right)^2}$$

$$= \frac{-e^{z_k}}{\sum e^{z_i}} \cdot \frac{e^{z_j}}{\sum e^{z_i}}$$

$$= -\sigma(z_{1:k}, k) \cdot \sigma(z_{1:k}, j)$$

$$\therefore \frac{\partial \sigma(z_{1:k}, k)}{\partial z_j} = \begin{cases} \sigma(z_{1:k}, k) \left[1 - \sigma(z_{1:k}, k) \right] & k = j \\ -\sigma(z_{1:k}, k) \cdot \sigma(z_{1:k}, j) & k \neq j \end{cases}$$

∎

CSCC11 Assignment 2.

Ti, Zhang

1000424517

2. Find the gradient of the log likelihood for a single point $(\mathbf{x}, y)$ with respect to $\mathbf{w}_j$:

$$\frac{\partial}{\partial \mathbf{w}_j} \sum_{k=1}^{K} y_k \log \sigma(z_{1:K},\, k) \quad \text{---} \quad (2)$$

$$(2) = \sum_{1}^{K} y_k \cdot \frac{\partial}{\partial w_j} \left[\log \sigma(z_{1:K}, k)\right]$$

$$= \sum_{1}^{K} y_k \cdot \left[\frac{\partial \log \sigma(z_{1:K}, k)}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_j}\right]$$

$$= \sum_{1}^{K} y_k \cdot \left[\frac{\partial \log \sigma(z_{1:K}, k)}{\partial z_j} \cdot x_{1:K}\right]$$

$$= \sum_{1}^{K} y_k \cdot \left[\frac{1}{\sigma(z_{1:K}, k)} \cdot \frac{\partial \sigma(z_{1:K}, k)}{\partial z_j} \cdot x_{1:K}\right]$$

$$= \sum_{k=j}^{K} y_k \cdot \frac{1}{\sigma(z_{1:K}, k)} \sigma(z_{1:K}, k)\left[1 - \sigma(z_{1:K}, k)\right] \cdot x_{1:K}$$

$$+ \sum_{k \neq j}^{K} y_k \cdot \frac{1}{\sigma(z_{1:K}, k)} \left[-\sigma(z_{1:K}, k) \cdot \sigma(z_{1:K}, j)\right] \cdot x_{1:K}$$

$$= \sum_{k=j}^{K} y_k \left[1 - \sigma(z_{1:K}, k)\right] \cdot x - \sum_{k \neq j}^{K} y_k \cdot \sigma(z_{1:K}, j) \cdot x$$

$$= \sum_{k=j}^{K} y_k \cdot x - \sum_{k=j}^{K} y_k \sigma(z_{1:K}, j) x - \sum_{k \neq j}^{K} y_k \sigma(z_{1:K}, j) \cdot x$$

$$= \sum_{k=j}^{K} y_j \cdot x - \sum_{1}^{K} y_k \sigma(z_{1:K}, j) x$$

$$= y_j x - \sigma(z_{1:K}, j) x \cdot \sum_{1}^{K} y_k$$

$$= y_j x - \sigma(z_{1:K}, j) x$$ ∎

Know that $z_{i,k} = \mathbf{w}_k^T \mathbf{x}_i$

$$\frac{\partial}{\partial w_j} = \frac{\partial}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_j}$$

($x_{1:K}$ note as x)

$\because y_{i,k} = \begin{cases} 1 & \text{when } k \text{ is the correct class} \\ 0 & \text{otherwise} \end{cases}$

$\therefore \sum_{1}^{K} y_k = 1$

3. Find the gradient of the loss with respect to $\mathbf{w}_j$:

$$\frac{\partial E(\mathbf{w}_{1:K})}{\partial \mathbf{w}_j}. \tag{11}$$

Hint: Use the results above. And the gradient should have a form similar to Eqn. (48) in Chapter 9.6 of the online lecture notes.

Known that $E(\mathbf{w}_{1:K}) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \sigma(z_{i,1:K},\, k)$

$$\frac{\partial E(w_{1:K})}{\partial w_j} = \frac{\partial}{\partial w_j} \left[-\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \sigma(z_{i,1:K}, k)\right].$$

$$= -\sum_{i=1}^{N} \left\{\frac{\partial}{\partial w_j} \sum_{k=1}^{K} y_{i,k} \log \sigma(z_{i,1:K}, k)\right\}$$ which $\{\}$ part is done in Q2.

$$= -\sum_{i=1}^{N} \left\{y_{ij} x_i - \sigma(z_{i,1:K}, j) x_i\right\}$$ ∎

CSCC11 Assignment 2.
Ti, Zhang
1000424517

4. Now, suppose we have D dimensional inputs and K classes. For each of K classes, let there be a weight vector, $\mathbf{w}_k = (b_k, w_{k,1}, ..., w_{k,d})^T$. If we include regularization, with a (diagonal) Gaussian prior, the negative log-posterior becomes, up to an additive constant,

$$\hat{E}(\mathbf{w}_{1:K}) = -\log\left[p(\mathbf{w}_{1:K})\prod_{i=1}^{N} p(y=y_i \mid \mathbf{x}_i, \mathbf{w}_{1:K})\right], \tag{12}$$

where $p(\mathbf{w}_{1:K})$ is the joint distribution over K indepdendent Gaussian densities with the same diagonal covariance. That is,

$$p(\mathbf{w}_{1:K}) = \prod_{k=1}^{K}\left(\frac{1}{((2\pi)^{(D+1)}\beta\alpha^D)^{1/2}}\exp\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)\right). \tag{13}$$

where the covariance matrix is given by $C = \text{diag}(\beta, \alpha, \alpha, \ldots, \alpha) \in \mathbb{R}^{(D+1)\times(D+1)}$. Here, α denotes the variance of the prior on each of the weights, and β is the prior variance on the bias term. Usually we don't want a strong prior on the bias term so $\beta \gg \alpha$. Derive $\frac{\partial \hat{E}}{\partial \mathbf{w}_k}$ for this regularized objective function.

Hint: Your negative log-posterior should have form $\hat{E}(\mathbf{w}_{1:K}) = E(\mathbf{w}_{1:K}) + E_2(\mathbf{w}_{1:K})$, where $E_2(\mathbf{w}_{1:K})$ is the regularization term.

$$\hat{E}(w_{1:K}) = -\log\left[p(w_{1:K})\prod_{i=1}^{N} p(y=y_i \mid x_i, w_{1:K})\right]$$

$$= -\log\left[p(w_{1:K})\right] - \log\left[\prod_{i=1}^{N} p(y=y_i \mid x_i, w_{1:K})\right]$$

$$\therefore \frac{\partial}{\partial w_k}\hat{E}(w_{1:K}) = \frac{\partial}{\partial w_k}\left\{-\log\left[p(w_{1:K})\right] - \log\left[\prod_{i=1}^{N} p(y=y_i \mid x_i, w_{1:K})\right]\right\}$$

$$= \frac{\partial}{\partial w_k}\left\{-\log\left[p(w_{1:K})\right]\right\} - \frac{\partial}{\partial w_k}\left\{\log\left[\prod_{i=1}^{N} p(y=y_i \mid x_i, w_{1:K})\right]\right\} \quad (3)$$

Known that $E(\mathbf{w}_{1:K}) = -\log\prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}_{1:K})$, $\therefore$ (3) $= E(w_{1:K})$

$$\frac{\partial}{\partial w_k}\hat{E}(w_{1:K}) = -\frac{\partial}{\partial k}\left\{\log\left[p(w_{1:K})\right]\right\} - \frac{\partial}{\partial w_k}E(w_{1:K}) \quad \text{----} \quad (4)$$

$$\log\left[p(w_{1:K})\right] = \log\left[\prod_{k=1}^{K}\left(\frac{1}{((2\pi)^{(D+1)}\beta\alpha^D)^{1/2}}\exp\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)\right)\right]$$

$$= \sum_{k=1}^{K}\log\left(\frac{1}{((2\pi)^{(D+1)}\beta\alpha^D)^{1/2}}\exp\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)\right)$$

$$= K\log\left(\frac{1}{((2\pi)^{(D+1)}\beta\alpha^D)^{1/2}}\right) + \sum_{k=1}^{K}\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)$$

$$\frac{\partial}{\partial w_k}\left\{\log\left[p(w_{1:K})\right]\right\} = \frac{\partial}{\partial w_k}\left\{K\log\left(\frac{1}{((2\pi)^{(D+1)}\beta\alpha^D)^{1/2}}\right) + \sum_{k=1}^{K}\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)\right\}$$

$$= \frac{\partial}{\partial w_k}\left\{\sum_{k=1}^{K}\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)\right\} = \frac{\partial}{\partial w_k}\left\{\left(-\frac{\mathbf{w}_k^T C^{-1}\mathbf{w}_k}{2}\right)\right\}$$

$$= -\frac{1}{2}\left\{C^{-1}w_k + (C^{-1})^T w_k\right\}$$

$$= -\frac{1}{2}\cdot 2C^{-1}w_k = -C^{-1}w_k$$

$$\therefore \frac{\partial}{\partial w_k}\hat{E}(w_{1:K}) = (4) = -\frac{\partial}{\partial w_k}\left\{\log\left[p(w_{1:K})\right]\right\} - \frac{\partial}{\partial w_k}E(w_{1:K}) \quad (4)$$

(4) is calculated in Q3.

$$= C^{-1}w_k + \sum_{i=1}^{N}\left\{y_{ij}x - \sigma(x_{i,1:K,j})x\right\}$$