Homework 2

1. In the class, we derived the closed-form solution for solving the parameters of the linear regression with one output. In this problem, you will apply the same technique to derive the equations for learning the parameters of linear regression with two outputs $y \in R^2$ and $y = (y_1, y_2)^\top$ jointly. Given training data $D = \{x[m], y[m]\}, m = 1, 2, .., M$, derive the equations to learn the regression parameter matrix $W = [W_1, W_2]$, where $W_1 = [w_1, w_{1,0}]^\top$ and $W_2 = [w_2, w_{2,0}]^\top$ by minimizing the mean squared errors.

Let's assume each sample $x[m]$ has N features:

so $x[m] \in R^{N\times 1}$, $W \in R^{(N+1)\times 2}$, $y \in R^{2\times 1}$

The regression is:

$$\begin{aligned}\hat{y}[m] &= W^\top \begin{bmatrix} x[m] \\ 1 \end{bmatrix} \\ &= W^\top X[m]\end{aligned}$$

The loss function is:

$$\begin{aligned}\mathcal{L}(D; W) &= \frac{1}{2}\sum_{i=1}^{M}(y[m] - W^\top X[m])^\top (y[m] - W^\top X[m]) \\ &= \frac{1}{2}\sum_{i=1}^{M}(y[m])^\top y[m] - (W^\top X[m])^\top y[m] - (y[m])^\top W^\top X[m] + (W^\top X[m]))^\top W^\top X[m] \\ &= \frac{1}{2}\sum_{i=1}^{M}(y[m])^\top y[m] - 2(W^\top X[m])^\top y[m] + (W^\top X[m]))^\top W^\top X[m] \\ &= \frac{1}{2}\sum_{i=1}^{M}(y[m])^\top y[m] - 2X[m]^\top W y[m] + (W^\top X[m]))^\top W^\top X[m]\end{aligned}$$

To minimize the loss function, we want the gradient $\bigtriangledown_W \mathcal{L}(D; W) = 0$

The close-form solution is:

$$\begin{aligned}\bigtriangledown_W \mathcal{L}(D; W) &= \sum_{i=1}^{M} -\frac{\partial 2(W^\top X[m])^\top y[m]}{\partial W} + \frac{\partial (W^\top X[m]))^\top W^\top X[m])}{\partial W} \\ &= \sum_{i=1}^{M} -2X[m]y^\top + 2X[m]X[m]^\top W \\ \Rightarrow \sum_{i=1}^{M} 2X[m]y^\top &= \sum_{i=1}^{M} 2X[m]X[m]^\top W \\ \Rightarrow W &= \sum_{i=1}^{M}(X[m]X[m]^\top)^{-1}X[m]y^\top\end{aligned}$$

2. For binary classification with logistic regression classification, where $y \in \{+1, -1\}$, we introduced the sigmoid function, where $\sigma()$ is the sigmoid function. For binary classification with $y \in \{1, 0\}$ and $x \in R^N$, we may use a new activation function such that $p(y|x) = \phi(y * (w^\top x + w_0))$, where $phi()$ is defined as:

$$\phi(y(w^\top x + w_0)) = \frac{e^{y(w^\top x + w_0)}}{1 + e^{(w^\top x + w_0)}}$$

Given trainig data $D = \{x[m], y[m]\}, m = 1, 2, \ldots, M$, derive the gradient descent solution to parameters Θ, where $\Theta = \begin{bmatrix} w \\ w_0 \end{bmatrix}$ using the negative conditional log likelihood (cross-entropy) as the loss function.

Let's denote $\Theta = \begin{bmatrix} w \\ w_0 \end{bmatrix}, X = \begin{bmatrix} x \\ 1 \end{bmatrix}$, so $w^\top x + w_0 = X^\top \Theta$

and denote scalar $z = X^\top \Theta$

Then the activation function is $p(y|x) : p = \phi(yz) = \frac{e^{yz}}{1+e^z}$

The loss function $\mathcal{L}(D, \Theta) = -\log p(y|x) = -\log \phi(yX^\top \Theta) = -\log \frac{e^{yX^\top \Theta}}{1+e^{X^\top \Theta}}$

The gradient:

$$\begin{aligned}
\bigtriangledown_\Theta \mathcal{L}(D, \Theta) &= \frac{\partial \mathcal{L}}{\partial p} \frac{\partial p}{\partial z} \frac{\partial z}{\partial \Theta} \\
&= -\frac{1}{p}\left(\frac{\partial e^{yz}}{\partial z} \frac{\partial z}{\partial \Theta}(1+e^z)^{-1} + e^{yz} \frac{\partial (1+e^z)^{-1}}{\partial z} \frac{\partial z}{\partial \Theta} \right) \\
&= -\frac{1}{p}\left(y e^{yz} \frac{\partial z}{\partial \Theta}(1+e^z)^{-1} + e^{yz}(-(1+e^z)^{-2} e^z) \frac{\partial z}{\partial \Theta} \right) \\
&= -\frac{1}{e^{yz}(1+e^z)^{-1}} \left(y e^{yz}(1+e^z)^{-1} X - e^{yz}(1+e^z)^{-2} e^z X \right) \\
&= -\frac{1}{e^{yz}(1+e^z)^{-1}} e^{yz}(1+e^z)^{-1} \left(y - (1+e^z)^{-1} e^z \right) X \\
&= -\left(y - (1+e^z)^{-1} e^z \right) X
\end{aligned}$$

To update:

$$\mathrm{w}^{t+1} = \mathrm{w}^t - \eta(-(y - (1+e^z)^{-1} e^z)\mathrm{x}) = \mathrm{w}^t + \eta(y - (1+e^z)^{-1} e^z)\mathrm{x}$$
$$\mathrm{w}_0^{t+1} = \mathrm{w}_0^t - \eta(-(y - (1+e^z)^{-1} e^z)) = \mathrm{w}_0 + \eta(y - (1+e^z)^{-1} e^z)$$

3. In this class, we derive the equation for the parameters for binary discrimination. For multi-class discriminative classification using softmax, where $x \in R^D$, and $y \in \{C1, C2, C3\}$, given the training data $D = \{x[m], y[m]\}, m = 1, 2, 3..., M,$ derive the gradient equations to iteratively learn the parameters $\Theta_1, \Theta_2, \Theta_3$ by minimizing the total negative log conditional likelihood, subject to the L1 norm on the parameters.

The x $\in R^{D\times 1}$, and let $X = \begin{bmatrix} \text{x} \\ 1 \end{bmatrix}$, dim is $[(D+1), 1]$

Total negative log conditional likelihood:

$$\begin{aligned}
\mathcal{L}(D;\Theta) = -\sum_{m=1}^{M} \log p(y[m]|x[m],\theta) &= -\sum_{m=1}^{M} \log \prod_{k=1}^{K} p(y[m]=k|x[m],\Theta_k)^{\mathbb{I}_{y[m]=k}} + \lambda|\Theta|_1 \\
&= -\sum_{m=1}^{M}\sum_{k=1}^{K} \mathbb{I}_{y[m]=k} \log p(y[m]=k|x[m],\Theta_k) + \lambda|\Theta|_1 \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K} \mathbb{I}_{y[m]=k} \log \frac{\exp(X^\top\Theta_k)}{\sum_{k'=1}^{K}\exp(X^\top\Theta_{k'})} + \lambda|\Theta|_1 \\
&= -\sum_{m=1}^{M}\sum_{k=1}^{K} \mathbb{I}_{y[m]=k}[\log\exp(X^\top\Theta_k) - \log\sum_{k'=1}^{K}\exp(X^\top\Theta_{k'})] + \lambda|\Theta|_1 \\
&= -\sum_{m=1}^{M}\sum_{k=1}^{K} \mathbb{I}_{y[m]=k}[X^\top\Theta_k - \log\sum_{k'=1}^{K}\exp(X^\top\Theta_{k'})] + \lambda|\Theta|_1
\end{aligned}$$

The gradient $\bigtriangledown_\Theta \mathcal{L}(D;\Theta) = \left[\frac{\partial\mathcal{L}(D;\Theta)}{\partial\Theta_1}, \ldots, \frac{\partial\mathcal{L}(D;\Theta)}{\partial\Theta_k}\right]$

For each Θ_k, the gradient $\bigtriangledown_{\Theta_k}\mathcal{L}(D;\Theta_k)$:

$$\begin{aligned}
\bigtriangledown_{\Theta_k}\mathcal{L}(D;\Theta_k) &= \sum_{m=1}^{M} -\mathbb{I}_{y[m]=k}X[m] + \frac{\partial \log\sum_{k'=1}^{K}\exp(X^\top\Theta_{k'})}{\partial\Theta_k} + \lambda\frac{\partial|\Theta_k|}{\partial\Theta_k} \\
&= \sum_{m=1}^{M} -\mathbb{I}_{y[m]=k}X[m] + \frac{\exp(X^\top\Theta_k)}{\sum_{k'=1}^{K}\exp(X^\top\Theta_{k'})}X[m] + \lambda\text{sign}(\Theta_k) \\
&= \sum_{m=1}^{M}(\sigma_M[m](k) - \mathbb{I}_{y[m]=k})X[m] + \lambda\text{sign}(\Theta_k)
\end{aligned}$$

The gradient descent:

For $k = 1:K$

$$\Theta_k^{t+1} = \Theta_k^t - \eta(\sum_{m=1}^{M}(\sigma_M[m](k) - \mathbb{I}_{y[m]=k})X[m] + \lambda\text{sign}(\Theta_k))$$

In [ ]: