HW4 HADI SHARIFI

Q1-a
Check the code bundled with the report

Q1-b
Check the code bundled with the report

Q1-c
Check the code bundled with the report

Q1-d
Check the code bundled with the report.

I used the KMean function from ScikitLearn package. The best accuracy I got was 80%.

	Accuracy	Misclassification for 2	Misclassification for 6
KMean	94%	0.05	0.07
GMM-EM	81%	0.19	0.15

Q2 - a

To derive the gradiant of cost function. The cost function is the loglikelihood

$$\ell(\theta)=\sum_{i=1}^{n}\{-\log(1+e^{-\theta x_i})+(y_i-1)\theta x_i\}$$

$$\frac{\partial \ell(\theta)}{\partial \theta}=\sum_{i=1}^{n}(y_i x_i - x_i - (\frac{e^{-\theta x_i}}{1+e^{-\theta x_i}})x_i)$$

$$=\sum_{i=1}^{n}(y_i-\frac{1}{1+e^{\theta x_i}})x_i = Fdx$$ this is our gradiant decent steps.

for gradiant decent, we need define the steps which is the formula above
And we need the learning rate λ.

the pseudo code will be:

Procedure GD (Fdx , $\theta(0)$)

$\theta \leftarrow \theta(0)$
while not converged do
$\theta \leftarrow \theta + \lambda$ Fdx(θ)

return θ

Fdx is a function that returns a matrix of form $Fdx(\theta)=\begin{bmatrix}\frac{\partial \ell(\theta)}{\partial \theta_1}\\ \frac{\partial \ell(\theta)}{\partial \theta_2}\\ \vdots \\ \frac{\partial \ell(\theta)}{\partial \theta_n}\end{bmatrix}$

$\theta \in \mathbb{R}$

$\theta(0)$ is a Random initialization

the returned θ is a point in which the log likelihood is the optimum. We run the Algorithm for each and points.

Q2-b

Stochastic gradient decent is little different. gradient decent iterates over all points and for lage data that is very time consuming.

In SGD, we shuffle the data; pick Random K data points of them and run the gradient steps. In SGD, we also have the learning rate λ. the gradient step is similar to Q2-a ($Fdx(\theta)$).

Procedure SGD (Fdx, $\theta(0)$)

    θ ← θ(0)
    while not converged do
        for shuff data {1,2,...N} pick K data points
            For k ∈ {1,2,...K} do
                θ_k ← θ_k + λ Fdx(θ_k)
    return θ_k

Q2-C

In order to identify a training Problem as concave or convex, we need to derive the Hessian matrix of $\ell(\theta)$.

For data Points of type R^n the hessian matrix will be:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

for problem Q2, the data Point are in R Therefore

Taking Second deravative would do the work.

the Hessian matrix has only one item.

$$\frac{\partial}{\partial x} \frac{1}{1+e^{ax}} = \frac{-ae^{ax}}{(1+e^{ax})^2}$$

from Q1-a

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left(y_i - \frac{1}{1+e^{\theta x_i}}\right) x_i$$

$$\frac{\partial^2 \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{x_i^T e^{\theta x_i}}{(1+e^{\theta x_i})^2} \times x_i = \sum_{i=1}^{n} \frac{e^{\theta x_i}}{(1+e^{\theta x_i})^2}$$

$$= \sum_{i=1}^{n} \frac{1}{1+e^{\theta x_i}} \cdot \left(\frac{e^{\theta x_i}}{1+e^{\theta x_i}}\right) = \sum_{i=1}^{n} \frac{1}{1+e^{\theta x_i}} \left(\frac{1}{1+e^{\theta x_i}} - 1\right)$$

(*)

$\circledast \Rightarrow \dfrac{1}{1+e^{\theta x_i}}$ is always $0 < \square < 1$ which makes

$$\sum_{i=1}^{n}\left(\frac{1}{1+e^{\theta x_i}} \times \left(\frac{1}{1+e^{\theta x_i}} - 1\right)\right) < 0 \Rightarrow$$ hence the problem is <u>concave</u> and NOT <u>convex</u>