SUBHODIP SAHA
5485317
HW3

1. $f(w) = \frac{1}{n} \sum_{i=1}^{n} \{ -y_i w^T x_i + \log(1+\exp(w^T x_i)) \} + \frac{\lambda}{2} \|w\|_2^2$

Let's define $\boxed{h_w(x^i) = \frac{1}{1+e^{-w^T x^i}}}$

Let's atfirst prove,

$y^i w^T x^i + \log(1+\exp(w^T x^i)) = -y^i \log h_w(x^i) - (1-y^i)\log(1-h_w(x^i))$

then it would be easier to calculate gradient.

$y^i \log h_w(x^i) + (1-y^i)\log(1-h_w(x^i))$

$= y^i \log h_w(x^i) + \log h_w(x^i) - y^i \log(1-h_w(x^i))$

$= y^i \log h_w(x^i) - \log(1-h_w(x^i))] + \log(1-h_w(x^i))$

$= y^i [\log h_w(x^i) - \log(1-h_w(x^i))] + \log(1-h_w(x^i))$

$= y^i \frac{h_w(x^i)}{1-h_w(x^i)} + \log(1-h_w(x^i))$

$h_w(x^i) = \frac{1}{1+e^{-w^T x_i}} = \frac{e^{w^T x_i}}{e^{w^T x^i}+1}$

$1-h_w(x^i) = \frac{1}{e^{w^T x^i}+1}$

$\frac{h_w(x^i)}{1-h_w(x^i)} = e^{w^T x^i}$

$y^i \frac{h_w(x^i)}{1-h_w(x^i)} + \log(1-h_w(x^i)) = y^i \theta^T x_i - \log(1+e^{w^T x_i})$

So, we have proved,

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} -y_i w^T x_i + \log(1+\exp(w^T x_i)) + \frac{\lambda}{2}\|w\|_2^2 = -\frac{1}{n}\sum_{i=1}^{n} y^i \log h_w(x^i) + (1-y^i)\log(1-h_w(x^i)) + \frac{\lambda}{2}\|w\|_2^2}$$

Now we will differentiate this expression.

$$\frac{\partial f(w)}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n} y^i \frac{\partial}{\partial w_j}\log h_w(x^i) + (1-y^i)\frac{\partial}{\partial w_j}\log(1-h_w(x^i)) + \lambda w_j$$

$$= -\frac{1}{n}\sum_{i=1}^{n} y^i \frac{\frac{\partial h_w(x^i)}{\partial w_j}}{h_w(x^i)} + (1-y^i)\frac{\frac{\partial}{\partial w_j}(1-h_w(x^i))}{1-h_w(x^i)}$$

$$h_w(x^i) = \frac{1}{1+e^{-w^T x^i}} \qquad \frac{\partial h_w(x^i)}{\partial w_j} = +\frac{e^{-w^T x^i}\, x^i_j}{(1+e^{-w^T x^i})^2} = h_w(x^i)(1-h_w(x^i))x^i_j$$

$$-\frac{1}{n}\sum_{i=1}^{n} y^i(1-h_w(x^i))x^i_j - (1-y^i)h_w(x^i)x^i_j + \lambda w_j$$

$$= -\frac{1}{n}\sum_{i=1}^{n} x^i_j\left(y^i - \cancel{y^i h_w(x^i)} - h_w(x^i) + \cancel{y^i h_w(x^i)}\right) + \lambda w_j$$

$$= \frac{1}{n}\sum_{i=1}^{n}(h_w(x^i) - y^i)x^i_j + \lambda w_j$$

$$\boxed{\frac{\partial f(w)}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n}(h_w(x^i)-y^i)x^i_j + \lambda w_j}$$

Expression for gradient.

$$\Delta w_j = -\eta \frac{\partial f}{\partial w_j} \qquad \eta \rightarrow \text{step size}$$ 1'

$$= +\eta\left[\frac{1}{n}\sum_{i=1}^{n}\left(y^i - h_w(x^i)\right)x^i_j - \lambda w_j\right]$$

$$\boxed{w_j^{new} = w_j + \eta\left[\frac{1}{n}\sum_{i=1}^{n}\left(y^i - h_w(x^i)\right)x^i_j - \lambda w_j\right]}$$

→ We should update the ~~...~~ w_j according to that.

b. Let's derive 2nd derivative of $f(w)$.

$$\frac{\partial f(w)}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n}\left(h_w(x^i) - y^i\right)x^i_j + \lambda w_j$$

$$\frac{\partial^2 f(w)}{\partial w_j \partial w_k} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w_k}\left(h_w(x^i) - y^i\right)x^i_j + \lambda \delta_{jk}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial h_w(x^i)}{\partial w_k}x^i_j + \lambda \delta_{jk}$$

We know, $\frac{\partial h_w(x^i)}{\partial w_k} = h_w(x^i)\left(1 - h_w(x^i)\right)x^i_k$

$$\cancel{= \frac{1}{n}\sum_{i=1}^{n} x^i}$$

$$\frac{\partial^2 f(w)}{\partial w_j \partial w_k} = \frac{1}{n}\sum_{i=1}^{n} h_w(x^i)\left(1 - h_w(x^i)\right)x^i_k x^i_j + \lambda \delta_{jk}$$

$$\boxed{\frac{\partial^2 f}{\partial w_j \partial w_k} = \frac{1}{n}\sum_{i=1}^{n} h_w(x^i)\left(1-h_w(x^i)\right)x^i_j x^i_k + \lambda \delta_{jk}}$$

$\rightarrow$ The 2nd derivative of objective fun^c.

for strong convexity we need to show

$\nabla^2 f(w) \geq \alpha \mathbb{1}$ for some $\alpha > 0$.

Let's take x is 1 dim, we can generalize to multidimension.

then,

$$\frac{\partial^2 f}{\partial w^2} = \frac{1}{n}\sum_i h_w(x^i)\left(1-h_w(x^i)\right)x^{i^2} + \lambda$$

we know, $\underline{1} \geq h_w(x^i) \geq 0$ so, $\underline{1} \geq 1-h_w(x^i) \geq 0$

$\lambda > 0$.

$x_i^2 \geq 0$

$$\frac{\partial^2 f}{\partial w^2} = \frac{1}{n}\sum_i \underbrace{h_w(x^i)\left(1-h_w(x^i)\right)}_{\geq 0}\underbrace{x^{i^2}}_{\geq 0} + \underbrace{\lambda}_{>0}$$

So, $\frac{\partial^2 f}{\partial w^2} > 0$ ~~$\rightarrow \nabla^2 f(w) \geq \alpha$~~

can be written as $\nabla^2 f(w) \geq \alpha$

If it is true for 1-dim, it should be true for multidim.

So, the fun^c is strongly convex

$$\boxed{\nabla^2 f(w) - \alpha \mathbb{1} \geq 0}$$

b)

I am proving the strong convexity of LR using another method, (for any dim - general case)

$$f(w) = -\sum_{i=1}^{n} \underbrace{+y^i \log h_w(x^i)}\ \underbrace{+(1-y^i)\log(1-h_w(x^i))}\ \underbrace{+\frac{\lambda}{2}\|w\|_2^2}$$

$$= \sum_{i=1}^{n} \underbrace{-y^i \log h_w(x^i)}_{\text{Need to prove it is convex fun}^c.} \underbrace{-(1-y^i)\log(1-h_w(x^i))}_{\text{Need to prove it is convex fun}^c.} + \underbrace{\lambda/2\,\|w\|_2^2}_{\text{Need to prove convex fn}^c.}$$

Let's take,

$h(z) = \frac{1}{1+\exp(z)}$ $\quad f_1(z) = -\log(h(z)) \quad f_2(z) = -\log(1-h(z))$

$\frac{d}{dz} f_1(z) = -\log(1/1+\exp(-z)) = -1 + 1/(1+\exp(-z)) = -1 + h(z)$

$\frac{d^2}{dz^2} f_1(z) = \frac{d}{dz}(h(z)) = \underbrace{h(z)(1-h(z))}_{\geq 0} \rightarrow f_1(z)$ is convex.

$\frac{d}{dz} f_2(z) = \log(1+\exp(-z)) = f_1(z) + z$.

$\frac{d^2}{dz^2} f_2(z) = \frac{d^2}{dz^2} f_1(z) = \underbrace{h(z)(1-h(z))}_{\geq 0} \rightarrow f_2(z)$ is also convex.

$\frac{d^2}{dw^2}\lambda\|w\|_2^2 = \underbrace{\lambda}_{>0} \rightarrow$ is also convex.

$f(w)$ is strongly convex.

Now, we need to prove

$A \in R^{m\times n}$

$b \in R^m$

Let's take, $g(y) = f(Ay+b)$

$\nabla_y g(y) = A^T \nabla_y f(Ay+b) \in R^n$

$\nabla_y^2 g(y) = A^T \nabla_y^2 f(Ay+b) A \in R^{n\times n}$

Since f_1 & f_2 (sigmoid) is convex func,

Proving: $\nabla_x^2 f(x) \geq 0$

$$z^T \nabla_y^2 g(y) z = z^T A^T \nabla_x^2 f(Ay+b) Az$$

$$= (Az)^T \nabla_x^2 f(Ay+b)(Az) \geq 0.$$

So, $\nabla_y^2 g(y)$ is also positive semi definite matrix.

~~If $\nabla_y^2 g(y)$ is +ve semi definite, i.e. $\nabla_y^2 g(y) \geq \alpha$~~

~~$f(w)$ is strongly convex.~~

If $\nabla_w^2 g(w)$ is +ve semi-definite, i.e. $\nabla_w^2 g(w) \geq \alpha$

then $f(w)$ is strongly convex.

$\underline{c}$ The 2nd derivative of objective fun^c $\underline{1}^2$

$$\frac{\partial^2 f}{\partial w_j \partial w_k} = \frac{1}{n}\sum_{i=1}^{n} h_w(x^i)(1-h_w(x^i))\, x^i_j x^i_k + \lambda \delta_{jK}$$

To establish smoothness we need to show,

$\nabla^2 f(w) \leq \beta \mathbb{1}$ for some β $\beta < \infty$

As. $0 \leq h_w(x^i) \leq 1$

Then. $h_w(x^i)(1-h_w(x^i)) \leq 0.25$ (for $h_w(x^i)=0.5$)

Let's prove for 1 dim.

$$\frac{\partial^2 f}{\partial w^2} = \frac{1}{n}\sum_i h_w(x^i)(1-h_w(x^i))\, x^{i^2} + \lambda.$$

$\max_w h_w(x^i)(1-h_w(x^i)) = 0.25$

$x_i^2 < \infty$ (as ~~[illegible]~~ x_i is finite)

$\lambda < \infty$ (λ is also finite)

$$\frac{\partial^2 f}{\partial w^2} = \frac{1}{n}\sum_i \underbrace{h_w(x^i)(1-h_w(x^i))}_{\max = 0.25}\ \underbrace{x^{i^2}}_{<\infty} + \underbrace{\lambda}_{<\infty}$$

~~So, $\frac{\partial^2 f}{\partial w^2}$ we can always~~

So, $\frac{\partial^2 f}{\partial w^2}$ is bounded & we can always write $\frac{\partial^2 f}{\partial w^2} \leq \beta$. As true for 1 dim. it must be true for multidim.

So the fun^c is smooth.

$$\boxed{\beta \mathbb{1} - \nabla^2 f(w) \geq 0}$$

d for smooth, strongly convex fun^c.

α smooth fun^c ~~strongly convex fun~~ satisfy,

$$f(w) \geq f(v) + (w-v)^T \nabla f(v) + \frac{\alpha}{2} \|w-v\|_2^2$$

β strongly convex fun^c satisfy,

$$f(w) \leq f(v) + (w-v)^T \nabla f(v) + \frac{\beta}{2} \|w-v\|_2^2$$

Initial point x_0,

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

η → step size.

for smooth, strongly convex fun^c $\boxed{\eta = \frac{2}{\alpha+\beta}}$

$$\underbrace{\boxed{f(x_T) - f(x^*) \leq \frac{\beta}{2} \exp\left(-\frac{4T}{\frac{\beta}{\alpha}+1}\right) \|x_0 - x^*\|^2}}_{}$$

This is the bound on difference.

fixed step size η, optimization is substantially faster, exponential.

"Q 2"

2. EM algorithm for learning Gaussian mixture model: EM is a type of clustering algorithm similar to k-means. Rather than having hard assignment into clusters like k-means we have soft assignment. As a result each Gaussian distribution has some responsibility for generating particular data point.

EM algo in high level.

$$\ln p(X|\theta) = \ln\left\{\sum_{z} p(X, Z|\theta)\right\}$$

Our goal is to maximize MLE of X given parameters θ. (X is observed, Z is hidden)

E step: Estimate posterior distribution of responsibilities of each gaussian $p(G_a|x_i)$ depending on weight (π), mean(μ) & covariance (Σ).
i.e. estimate $p(G_a|x_i) = f(\pi, \mu, \Sigma)$

M step: Use $p(G_a|x_i)$ to maximize likelihood w.r.t the parameters θ
i.e. maximize $\ln p(X|\mu, \Sigma, \pi)$

Repeat EM step until converge.

b/ M step (calculate mean, covariance, prior)

~~$\mu_k^{new} = \frac{1}{N}$~~ We define, $n_k = \sum_{i=1}^{n} p(G_k|x_i)$

$$\mu_k^{new} = \frac{1}{n_k} \sum_{i=1}^{n} p(G_k|x_i)\, x_i$$

$$\Sigma_k^{new} = \frac{1}{n_k} \sum_{i=1}^{n} p(G_k|x_i)(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T$$

$$\pi_k^{new} = \frac{n_k}{n}$$ for all component, $k = 1, \dots k$

where, $n_k = \sum_{i=1}^{n} p(G_k|x_i)$ (posterior prob.)

c/ E step

Gaussian with μ_k, mean covar Σ_k.

$$p(G_k|x_i) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

$$= \frac{\pi_k |\Sigma_k|^{-1/2} \exp\left[-\frac{1}{2}(x^i - \mu_k)^T \Sigma_k^{-1} (x^i - \mu_k)\right]}{\sum_{j=1}^{K} \pi_j |\Sigma_j|^{-1/2} \exp\left[-\frac{1}{2}(x^i - \mu_j)^T \Sigma_j^{-1} (x^i - \mu_j)\right]}$$

Problem 3

1. Error rates for myLogisticReg2 with Boston 50

F1	F2	F3	F4	F5	Mean	SD
0.17	0.22	0.24	0.22	0.16	0.20	0.03

2. Error rates for myLogisticReg2 with Boston 75

F1	F2	F3	F4	F5	Mean	SD
0.18	0.27	0.24	0.22	0.24	0.23	0.02

3. Error rates for LogisticRegression with Boston 50

F1	F2	F3	F4	F5	Mean	SD
0.12	0.20	0.25	0.25	0.17	0.20	0.04

4. Error rates for LogisticRegression with Boston 75

F1	F2	F3	F4	F5	Mean	SD
0.11	0.11	0.09	0.08	0.09	0.14	0.01