CSCI 5521 – Introduction to Machine Learning

Homework – 3

Mourya Karan Reddy Baddam – 5564234
Email – badda004@umn.edu

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

$$\frac{\partial f(w)}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n}\left\{+y_i x_i^j - \frac{x_i^j \exp(w^T x_i)}{1+\exp(w^T x_i)}\right\} + \frac{\lambda}{2}\cdot 2w_j$$

$$\frac{\partial f(w)}{\partial w_j} = \underline{\frac{1}{n}\sum_{i=1}^{n}\left\{(h_w(x_i) - y_i)x_i^j\right\} + \lambda w_j}$$

Where $$h_w(x_i) = \frac{1}{1+\exp(-w^T x_i)}$$

$$\Rightarrow \boxed{\nabla f(w) = \frac{1}{n} X^T\left(h_w(X) - y\right) + \lambda w}$$ Gradient

Where X is $n \times d$ matrix,

y is $n \times 1$ vector

$h_w(X)$ is $n \times 1$ vector

w is $d \times 1$ vector.

<u>Gradient descent</u>:- $w_{t+1} = w_t - \eta \nabla f(w_t)$

$$\Rightarrow w_{t+1} = w_t(1-\eta\lambda) - \frac{\eta}{n} X^T\left(h_{w_t}(X) - y\right)$$

Repeat until convergence ($w_{t+1} \to w^*$)

Where, η is the learning rate/ Step size. It could be a constant or varying for each step.

b) From a, we have

$$\frac{\partial f(w)}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(h_w(x_i)-y_i\right)x_i^j\right\} + \lambda w_j$$

$$\Rightarrow \frac{\partial^2 f(w)}{\partial w_j^2} = \lambda + \frac{1}{n}\sum_{i=1}^{n} x_i^j\left(h_w(x_i)\left(1-h_w(\cancel{x_i})\right)x_i^j\right)$$

$$\frac{\partial^2 f(w)}{\partial w_j^2} = \lambda + \frac{1}{n}\sum_{i=1}^{n}(x_i^j)^2\left(h_w(x_i)\left(1-h_w(x_i)\right)\right)$$

Because $h_w(x_i) = \dfrac{1}{1+\exp(-w^T x_i)}$

$$\frac{\partial h_w(x_i)}{\partial w_j} = \frac{x_i^j \exp(-w^T x_i)}{\left(1+\exp(-w^T x_i)\right)^2}$$

$$= x_i^j \cdot \frac{1}{1+\exp(-w^T x_i)} \cdot \frac{\exp(-w^T x_i)}{1+\exp(-w^T x_i)}$$

$$= \underline{x_i^j \; h_w(x_i)\cdot\left(1-h_w(x_i)\right)}$$

$$\frac{\partial^2 f(w)}{\partial w_j^2} = \lambda + \frac{1}{n}\sum_{i=1}^{n}(x_i^j)^2\left(h_w(x_i)\left(1-h_w(x_i)\right)\right)$$

$$\Rightarrow \frac{\partial^2 f(w)}{\partial w_j^2} \geq \lambda \geq 0 \cdot \text{ since } h_w(x_i)\left(1-h_w(x_i)\right) \geq 0$$

because $0 \leq h_w(x_i) \leq 1$

$\therefore$ $\underline{\alpha = \lambda}$. Hence the objective fuction is <u>strongly convex</u>.

c) From b,

$$\frac{\partial^2 f(w)}{\partial w_j^2} = \lambda + \frac{1}{n}\sum_{i=1}^{n}(x_i^j)^2\left(h_w(x_i)\left(1-h_w(x_i)\right)\right)$$

Let $h_w(x_i) = t$

$\Rightarrow \quad 0 \le t \le 1$

Max value of $t(1-t) \Rightarrow \dfrac{\partial\, t(1-t)}{\partial t} = 0$

$\Rightarrow \quad 2t = 1 \Rightarrow t = 1/2.$

$\therefore$ max value of $h_w(x_i)\left(1-h_w(x_i)\right) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}.$

$$\therefore \frac{\partial^2 f(w)}{\partial w_j^2} \le \lambda + \frac{1}{4}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i^j)^2$$

$$\therefore \quad \nabla^2 f(w) \le \beta I$$

where $\beta = \lambda + \frac{1}{4n}\max_j\left\{\sum_{i=1}^{n}(x_i^j)^2\right\}$

$\therefore$ The objective function is <u>smooth</u>.

d) If learning rate (step-size) $\eta = \frac{2}{\alpha + \beta}$

we have

$$f(w_T) - f(w^*) \leq \frac{\beta}{2} \exp\left(\frac{-4T}{\frac{\beta}{\alpha}+1}\right) \| w_0 - w^* \|^2$$

where $w_T \rightarrow$ iterate after T steps

$w^* \rightarrow$ Global minimizer

$\alpha = \lambda$ & $\frac{\partial^2 f(w)}{\partial w_j^2} \geq \lambda$

$\beta = \lambda + \frac{1}{4n} \max \left\{ \sum_{i=1}^{n} (x_i^j)^2 \right\}$

w_0 = starting point of w

Q2)

(a) In EM algorithm, for mixture of Gaussians model, we first initialize the π_h, μ_h and Σ_h values randomly and then until convergence we iterate the E- (Expectation) and M- (Maximization) steps as below:

- E- Find the posterior probabilities - $p(G_h|x_i)$ for all points x_i, i ∈ [1, n] and for all components G_h, h ∈ [1, k] using Bayes' Rule. We need the values of π_h, μ_h and Σ_h during this step. Based on the values of posterior probabilities - $p(G_h|x_i)$, we assign the component labels that has highest posterior probability to each point x_i.
- M- Using the posterior probabilities found in the E- step, we will find the maximum likelihood estimates of the π_h, μ_h and Σ_h parameters.

  Both the above steps are repeated until max iterations or until they converge using a convergence criterion.

(b) In the M- step, we calculate the component prior π_h, mean μ_h and covariance Σ_h as below:

$\mu_h = \frac{1}{N_h} \sum_{i=1}^{n} p(G_h|x_i) x_i$, for h ∈ [1, k]

$\Sigma_h = \frac{1}{N_h} \sum_{i=1}^{n} p(G_h|x_i)(x_i - \mu_h)(x_i - \mu_h)^T$, for h ∈ [1, k]

$\pi_h = \frac{N_h}{N}$, for h ∈ [1, k]
where : N_h = number of points in component h & N = Total number of points.

(c) In the E- step, we calculate the posterior probabilities $p(G_h|x_i)$ for all points x_i, i ∈ [1, n] and for all components G_h, h ∈ [1, k] using Bayes' Rule as below:

$$p(G_h|x_i) = \frac{p(G_h)p(x_i|G_h)}{p(x_i)}$$

$$p(G_h|x_i) = \frac{\pi_h \mathcal{N}(x_i|\mu_h, \Sigma_h)}{\sum_{j=1}^{k} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

Where, $\mathcal{N}(x_i|\mu_h, \Sigma_h) = \frac{1}{(2\pi)^{d/2}|\Sigma_h|^{1/2}} exp\left(-\frac{1}{2}(x_i - \mu_h)^T \Sigma_h^{-1}(x_i - \mu_h)\right)$

Q3: Summary of Error Rates:

Summary: MyLogisticReg2 with Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
19.61%	18.81%	9.9%	19.8%	22.77%	18.18%	4.35%

Summary: MyLogisticReg2 with Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
12.75%	11.88%	19.8%	13.86%	9.9%	13.64%	3.34%

Summary: LogisticRegression with Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
14.71%	11.88%	9.9%	14.85%	21.78%	14.62%	4.03%

Summary: LogisticRegression with Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
8.82%	8.91%	9.9%	10.89%	9.9%	9.69%	0.76%