1.

1.

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

assume $\|X^{i}\| \le R$

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

$$= \frac{1}{n}\sum_{i=1}^{n}(\sigma(w^T x_i) - y_i)x_i + \lambda\|w\|, \text{ where } \sigma(w^T x_i) = \frac{\exp(w^T x_i)}{1+\exp(w^T x_i)}$$

Steps: ① Initalize W_i values as random values $(i=0,1,\dots,d)$

where d is the number of features, $w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$

② Compute $\nabla_w f(w)$ above

③ compute $y_{k+1} = w_k - \eta\nabla_w f(w_k)$, until it converges

④ If $\|y_{k+1}\| \le R$, $w_{k+1} = y_{k+1}$, if $\|y_{k+1}\| > R$, $w_{k+1} = \frac{R}{\|y_{k+1}\|} y_{k+1}$

(b) Yes, for every $w_1 > w_2$, To prove

$$f(w_1) \ge f(w_2) + (w_1 - w_2)^T \nabla f(w_2) + \frac{\lambda}{2}\|w_1 - w_2\|^2$$

We need to prove $\frac{1}{n}\sum_{i=1}^{n}\{-y_i w_1^T x_i + \log(1+\exp(w_1^T x_i))\} + \frac{\lambda}{2}\|w_1\|^2 \ge$

$\frac{1}{n}\sum_{i=1}^{n}\{y_i w_2^T x_i + \log(1+\exp(w_2^T x_i))\} + \frac{\lambda}{2}\|w_2\|^2 + (w_1 - w_2)^T\{\frac{1}{n}\sum_{i=1}^{n}(\sigma(w_2^T x_i) - y_i)x_i$

$+ \lambda\|w_2\|\} + \frac{\lambda}{2}\|w_1 - w_2\|^2$.

To prove $f(w)$ is strongly convex, we can prove $f(w) - \frac{\lambda}{2}\|w\|^2$

is convex

$$\nabla^2\left(f(w) - \frac{\lambda}{2}\|w\|^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\exp(w^T x_i)\cdot[(1+\exp(w^T x_i)) + \exp(w^T x_i)]x_i}{(1+\exp(w^T x_i))^2}\right\}x_i \ge 0$$

Since $\exp(w^T x_i) > 0$ and $x_i \ge 0$ $(i = 0, 1, \dots, n)$

The subjective function $f(w)$ is strongly convex.

教师点评

评后反思

(c) Yes. Since for $\beta > \lambda$, since $\log\left(\frac{1+\exp(x_2)}{1+\exp(x_1)}\right) + (x_1 - x_2)^T \frac{\exp(x_2)}{1+\exp(x_2)} + \frac{\beta}{2}\|x_1 - x_2\|^2 \geq 0$ if $\beta = 2(x_1 - x_2)^T$ because $\frac{\exp(x_1)}{1+\exp(x_1)} > 0$ and $\log\left(\frac{1+\exp(x_2)}{\exp(x_1)}\right)$ is growing less than gradient function, and $-yw^T x$ is β-smooth since $-y(w_1^T - w_2^T)x + (w_1 - w_2)^T(-yx) + \frac{\beta}{2}\|w_1 - w_2\|^2 \geq 0$. Therefore for $\beta > \lambda$, there exists $w_1 > w_2$ where $f(w_1) \leq f(w_2) + (w_1 - w_2)^T \nabla f(w_2) + \frac{\beta}{2}\|w_1 - w_2\|^2$.

If $\beta = 2yx$, obviously $\log \frac{\lambda}{2}(\|w_2\|^2 - \|w_1\|^2) + (w_1 + w_2)^T \cdot \lambda\|w_1\| + \frac{\beta}{2}\|w_1 - w_2\|^2 \geq 0$

(d) For β-smooth and α-strongly convex function.

Strong-convexity: $f(w_T) - f(w^*) \leq (w_T - w^*)^T \nabla f(w^*) + \frac{\beta}{2}\|w_T - w^*\|^2 = \frac{\beta}{2}\|w_T - w^*\|^2$

since $\nabla f(w^*) = 0$

$w_{T+1} = w_T - \eta_T \nabla f(w_T)$: $\|w_{T+1} - w^*\|^2 = \|w_T - w^*\|^2 + \eta_T^2\|\nabla f(w_T)\|^2 - 2\eta_T \nabla f(w_T)^T(w_T - w^*)$

$\leq \|w_T - w^*\|^2 + \eta_T^2\|\nabla f(w_T)\|^2 - 2\eta_T\left(\frac{\partial\beta}{\partial+\beta}\|w_T - w^*\|^2 + \frac{1}{\partial+\beta}\|\nabla f(x_T)\|^2\right)$

$= \left(1 - 2\eta_T\frac{\partial\beta}{\partial+\beta}\right)\|w_T - w^*\|^2 + \eta_T\left(\eta_T - \frac{2}{\partial+\beta}\right)\|\nabla f(x_T)\|^2$

Suppose $\eta_T < \frac{2}{\partial+\beta}$, $\|w_{T+1} - w^*\|^2 \leq \left(1 - 2\eta_T\frac{\partial\beta}{\partial+\beta}\right)\|w_T - w^*\|^2$

$\leq \|w_0 - w^*\|^2 \prod_{n=1}^{k}\left(1 - \frac{2\eta_n\partial\beta}{\partial+\beta}\right)$ and $1 - x \leq$

Therefore, $f(w_T) - f(w^*) \leq \frac{\beta}{2}\|w_T - w^*\|^2 \leq \frac{\beta}{2}\prod_{n=1}^{T}\left(1 - \frac{2\eta_n\partial\beta}{\partial+\beta}\right)\|w_0 - w^*\|^2 \leq \frac{\beta}{2}\|w_0 - w^*\|^2$

作业用时

教师点评

评后反思

2.

作业内容

教师点评

评后反思

2.

(a)

E-step: Denote the current parameter values as Φ. Compute W_{ik} for all data points x_i, given the old set of parameters Φ^l and set the label G_i with the largest likelihood $P(G_i|X, \Phi^l) = h_i^t$ to the data point.

M-step: recompute the prior probability $P(G_i)$, mean vector and the covariance matrix with the posterior h_i^t, the data points X^t and the total number of class N, update the new set of parameters to be $\Phi^{(l)}$, maximum likelihood is what to get new parameters.

Keep doing E-step & M-step until h_i^t converges.

(b) $\pi_i^{(l)} = \dfrac{\sum_{t=1}^{N} h_i^t}{N}$, $\mu_i^{(l)} = \dfrac{\sum_{t=1}^{N} h_i^t x^t}{\sum_{t=1}^{N} h_i^t}$

$$\Sigma_i^{(l)} = \frac{\sum_{t=1}^{N} h_i^t (X^t - \mu_i^{(l+1)})(X^t - \mu_i^{(l+1)})^T}{\sum_{t=1}^{N} h_i^t}$$ where l is the step number

(c) $$h_i^t = P(\pi_i | X^t, \mu_i \Sigma_i) = \frac{P(X^t | \pi_i, \mu_i^l \Sigma_i^l) P(\pi_i)}{\sum_{j=1}^{k} P(X^t | \pi_j, \mu_j^l, \Sigma_j^l) P(\pi_j)}$$

$$P(X^t | \pi_i, \mu_i^l, \Sigma_i^l) = \pi_i \frac{1}{(2\pi)^{d/2} |\Sigma_i^l|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X^t - \mu_i^l)^T \Sigma^{-1} (X^t - \mu_i^l)\right)$$

3.

Error rates for MyLogisticReg2 with Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.19801980198019797	0.1188118811881188	0.1089108910891089	0.24752475247524752	0.1188118811881188	0.1584158415841584	0.05494826505153116

Error rates for MyLogisticReg2 with Boston25						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.13861386138613863	0.18811881188118806	0.03960396039603964	0.05940594059405946	0.2178217821782178	0.12871287128712872	0.06972996776894654

Error rates for LogisticRegression with Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.13861386138613863	0.13861386138613863	0.09900990099009899	0.27722772277227725	0.09900990099009899	0.1504950495049505	0.06579504049836 1

Error rates for LogisticRegression with Boston25						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.15841584158415845	0.2178217821782178	0.03960396039603964	0.12871287128712872	0.13861386138613863	0.13663366336633664	0.057528075431025526