1

Minor 1 Solutions
COL774
Spring 2018-19

Q.1. For logistic regression:-

$$LL(\theta) = \sum_{i=1}^{m} \log P(y^{(i)} | x^{(i)}; \theta). \quad \text{where } P(y^{(i)}=1 | x^{(i)}; \theta) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$$

$$\Rightarrow LL(\theta) = \sum_{i=1}^{m} \left[y^{(i)} \log \frac{1}{[1+e^{-\theta^T x^{(i)}}]} + [1-y^{(i)}] \log\left[1 - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right]\right]$$

We know that $\frac{\partial g(z)}{\partial z} = g(z)(1-g(z))$ where $g(z)$ is the sigmoid function.

$$\Rightarrow \frac{\partial}{\partial \theta_j} LL(\theta) = \sum_{i=1}^{m} \left[\frac{y^{(i)}}{\cancel{g(\theta^T x^{(i)})}} \cancel{g(\theta^T x^{(i)})} (1-g(\theta^T x^{(i)})) x_j^{(i)} \right] + \frac{(1-y^{(i)})}{\cancel{(1-g(\theta^T x^{(i)}))}}(-1) g(\theta^T x^{(i)}) \cancel{[1-g(\theta^T x^{(i)})]} (x_j^{(i)}) \Big]$$

$$= \sum_{i=1}^{m} \left[y^{(i)} - y^{(i)} g(\theta^T x^{(i)}) + (1-y^{(i)})(-1) g(\theta^T x^{(i)}) \right] x_j^{(i)} \Big]$$

$$= \sum_{i=1}^{m} [y^{(i)} - g(\theta^T x^{(i)})] x_j^{(i)}$$

$$\Rightarrow \frac{\partial}{\partial \theta_k \partial \theta_j} LL(\theta) = \sum_{i=1}^{m} (-1) g(\theta^T x^{(i)}) [1-g(\theta^T x^{(i)})] x_j^{(i)} x_k^{(i)}$$

$$\text{Now, } H_{jk} = \frac{\partial}{\partial \theta_k \partial \theta_j} LL(\theta) = \sum_{i=1}^{m} (-1) g(\theta^T x^{(i)}) [1-g(\theta^T x^{(i)})] x_j^{(i)} x_k^{(i)}$$

$$z^T H z = \sum_{j=1}^{n} \sum_{k=1}^{n} z_j H_{jk} z_k$$

$$\Rightarrow z^T H z = \sum_{j,k=1,1}^{n,n} z_j \left[\sum_{i=1}^{m} (-1) g(\theta^T x^{(i)}) [1-g(\theta^T x^{(i)})] x_j^{(i)} x_k^{(i)} \right] z_k$$

$$= \sum_{i=1}^{m} -1 \Big[\underbrace{\sum_j z_j x_j^{(i)} \sum_k z_k x_k^{(i)}}_{\text{identical}} \Big] [g(\theta^T x^{(i)}) (1-g(\theta^T x^{(i)}))]$$

$$= \sum_{i=1}^{m} \cancel{(-1)} \cancel{z_j x} -1 [z^T x^{(i)}]^2 g(\theta^T x^{(i)}) (1-g(\theta^T x^{(i)})) \leq 0$$

since each term inside the sum is ~~+ve~~ (or non-~~+ve~~ positive)

$\Rightarrow z^T H z \cancel{\geq 0} \leq 0$

$\Rightarrow H$ is ~~+ve~~ semi-definite

$\Rightarrow LL(\theta)$ is a concave function θ

(since its ~~se~~ matrix of corresponding second order derivatives is ~~+ve~~ semi-definite).

Q.2. Procedure:-

First Divide the training set Tr into two subsets. Train & validation. Let us say the split be 80/20.

Then, we first "train" the model on 80 examples & test/validate on 20 ~~w~~ for various values of τ.

te_best = ∞; τ_{best} = ~~1~~0;

~~For τ in range($0, \tau_{max}$) {~~

~~for($\tau = 0$; $\tau < \tau_{max}$; $\tau = \tau + \delta$) {~~ ~~increment~~

Let Tv set of examples in validation set.

For τ in range $(0, \tau_{max})$ with increments τ_Δ {

$\hat{T}_r = T_r - T_v$; te = 0;

For $(x^{(i)}, y^{(i)}) \in T_v$ {

Compute error of model on each example

$M = $ learnLWR$(\hat{T}_r, x^{(i)}, \tau)$

$\hat{y}^{(i)} = M(x^{(i)})$;

Compute total error

te = te + Err$(y^{(i)}, \hat{y}^{(i)})$;

}

Update τ-best if required

if (te < te_best) {

$\tau_{best} = \tau$;

te_best = te;

// if total error is best till now, update te-best & τ-best

}

}

Scanned by CamScanner

Next, compute the error on test set.

te = 0;

For $\forall$ $(x^{(i)}, y^{(i)}) \in$ Te {

M = learnLWR(Tr, $x^{(i)}$, τ_{best});

$\hat{y}^{(i)} = M(x^{(i)})$;

te = test Error($\hat{y}^{(i)}, y^{(i)}$);

}

return te; } Return the error on test set.

}

Q.3. In gradient ascent,

$$g_j = \frac{d}{d\theta_j} LL(\theta)$$

$$LL(\theta) = \sum_{i=1}^{m} \log(P(y^{(i)}; x^{(i)}; \theta))$$

$$\Rightarrow g_j = \frac{d}{d\theta_j} \sum_{i=1}^{m} \log[P(y^{(i)}; x^{(i)}; \theta)] = \sum_{i=1}^{m} \frac{d}{d\theta_j} [\log P(y^{(i)}; x^{(i)}; \theta)]$$

Now, consider SGD.

Define $LL(\theta)^{(i)} = \log(P(y^{(i)}; x^{(i)}; \theta))$

$$g_j^{(i)} = \frac{d}{d\theta_j} \log P(y^{(i)}, x^{(i)}; \theta)$$

~~Now $E[g_j^{(i)}] = \sum$ P(x^{(i)}, y^{(i)}) ... Prob ...~~

~~$\frac{d}{d\theta_j}\log[P(y^{(i)}, x^{(i)}; \theta)] = \frac{1}{m} \times \sum_{i=1}^{m} \frac{d}{d\theta_j}\log[P(y^{(i)}; x^{(i)}; \theta)]$~~

Scanned by CamScanner

$$E[g_j^{(i)}] = \sum_{i=1}^{m} \frac{1}{m} \frac{\partial}{\partial \theta_j}[\log p(y^{(i)}, x^{(i)}; \theta)]$$

$\frac{1}{m}$: Prob. of seeing ith example (sample uniformly at random from training set)

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} \log [p(y^{(i)}, x^{(i)}; \theta)]$$

$$= \frac{1}{m} g_j$$

$$\Rightarrow g_j = m * E[g_j^{(i)}] \quad \forall j \quad (\text{~~since we~~}$$

Thus holds $\forall j$.

Hence, proved.

~~Q4~~. Assumptions:- Yes, we did make i.i.d. assumption over the training set. Because of this assumption we can write:-

$$LL(\theta) = \log [p(y^{(1)}, \ldots y^{(m)}, x^{(1)} \ldots x^{(m)}; \theta)]$$

$\Downarrow$ since examples are independent

$$= \log \prod_{i=1}^{m} [p(y^{(i)}, x^{(i)}; \theta)]$$

$$= \sum_{i=1}^{m} \log [p(y^{(i)}, x^{(i)}; \theta)]$$

& the expression for $g_j = \frac{\partial}{\partial \theta_j} LL(\theta) = \sum_{i=1}^{m} \frac{\partial (\log p(y^{(i)}, x^{(i)}; \theta))}{\partial \theta_j}$ follows. This won't hold if $(x^{(i)}, y^{(i)})$ weren't not i.i.d.

Q.4. Normal Distribution :-

$$p(x^{(i)} | y^{(i)} = 1; \mu_1, \Sigma_1) = \frac{1}{(2\pi)^{n/2} |\Sigma_1|^{1/2}} e^{-\frac{(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}{2}}$$

Decision boundary :-

$$p(y^{(i)} = 1 | x^{(i)}; \Theta) = 0.5$$

$$= \frac{p(x^{(i)} | y^{(i)} = 1) \; p(y^{(i)} = 1)}{p(x^{(i)})}$$

$\hookrightarrow p(x^{(i)} | y^{(i)} = 1) \, p(y^{(i)} = 1) + p(x^{(i)} | y^{(i)} = 0) \, p(y^{(i)} = 0)$

$$\Rightarrow \frac{1}{\frac{p(x^{(i)}|y^{(i)}=0)p(y^{(i)}=0) + \cancel{p(x^{(i)}|y^{(i)}=1)p(y^{(i)}=1)}}{\cancel{p(x^{(i)}|y^{(i)}=1)} \, p(y^{(i)}=1)}} = 1/2$$

$$\Rightarrow \frac{1}{1 + \frac{p(x^{(i)}|y^{(i)}=0) \, p(y^{(i)}=0)}{p(x^{(i)}|y^{(i)}=1) \, p(y^{(i)}=1)}} = 1/2$$

$$\Rightarrow \frac{2}{1} = 1 + \frac{p(x^{(i)}|y^{(i)}=0) \, p(y^{(i)}=0)}{p(x^{(i)}|y^{(i)}=1) \, p(y^{(i)}=1)}$$

Taking log

$$0 = \log [p(x^{(i)} | y^{(i)} = 0) \, p(y^{(i)} = 0)] - \log [p(x^{(i)} | y^{(i)} = 1) \, p(y^{(i)} = 1)]$$

$\Rightarrow \log p(x^{(i)} | y^{(i)}=0) + \log p(y^{(i)}=0)$

$= \log p(x^{(i)} | y^{(i)}=1) + \log p(y^{(i)}=1)$

$\Rightarrow \log \frac{1}{(2\pi)^{n/2} |\Sigma_0|^{1/2}} + (-1)\frac{(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)}{2} + \log \phi$

$= \log \frac{1}{(2\pi)^{n/2} |\Sigma_1|^{1/2}} - \frac{(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2}$

$+ \log (1-\phi)$

$\Rightarrow$

$\frac{1}{2}\left[(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) - (x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0) \right]$

$= \log \frac{(2\pi)^{n/2} |\Sigma_0|^{1/2}}{(2\pi)^{n/2} |\Sigma_1|^{1/2}} + \log \left[\frac{(1-\phi)}{\phi} \right]$

$\Rightarrow$ $\cancel{x^T [\Sigma_1^{-1} - \Sigma_0^{-1}] x} \quad \cancel{+ x^T \Sigma_1^{-1} \mu_1 - \mu_1^T \Sigma_1^{-1} x}$

$x^T \Sigma_1^{-1} x + x^T \Sigma_1^{-1} \mu_1 - \mu_1^T \Sigma_1^{-1} x + \mu_1^T \Sigma_1^{-1} \mu_1$

$- [x^T \Sigma_0^{-1} x + x^T \Sigma_0^{-1} \mu_0 - \mu_0^T \Sigma_0^{-1} x + \mu_0^T \Sigma_0^{-1} \mu_0]$

Note $\Sigma_1^{-1} = (\Sigma_1^{-1})^T$ & $\Sigma_0^{-1} = (\Sigma_0^{-1})^T$ $\quad = \log \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} * \left[\frac{1-\phi}{\phi}\right]$

$\Rightarrow x^T [\Sigma_1^{-1} - \Sigma_0^{-1}] x + 2 [x^T \Sigma_1^{-1} \mu_1 - x^T \Sigma_0^{-1} \mu_0]$

$+ \mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0$

$= \log \frac{|\Sigma_0|^{1/2} (1-\phi)}{|\Sigma_1|^{1/2} \phi}$

$\Rightarrow x^T (\Sigma_1^{-1} - \Sigma_0^{-1}) x + 2x^T (\Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0) = \log \left[\frac{|\Sigma_0|^{1/2} (1-\phi)}{|\Sigma_1|^{1/2} \phi}\right] + \cdots$

(4)

Looking for quadratic terms:-

coefficients of $x_j x_k = 0 \; \forall \; j \neq k$

Since $\hat{\Sigma}_1$ & $\hat{\Sigma}_0$ both are diagonal

Coefficients of x_j^2 is given as:-

$$\Rightarrow (\hat{\Sigma}_1^{-1})_{jj} - (\hat{\Sigma}_0^{-1})_{jj}$$

$$= \left[\frac{1}{(\hat{\Sigma}_1)_{jj}} - \frac{1}{(\hat{\Sigma}_0)_{jj}} \right]$$

(b). When $\hat{\Sigma}_0 = \hat{\Sigma}_1$. Quadratic terms vanish.

So, Decision boundary is simply a linear function of $\underline{x}$. (Hyperplane).

(c) → when $x \in \underline{R^2}$.

(a) → Decision boundary is of form:

$$a x_1^2 + a' x_1 + b x_2^2 + b' x_2 + c = 0$$

⇒ Equation of an ellipse or hyperbola whose axes are aligned with principal x-y axis.

Q.5

$$P(y|x;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!} \quad y=k \quad \lambda = \theta^T x$$

for $y=k$

$$\Rightarrow \log p(y^{(i)}|x^{(i)};\theta) = \cancel{\log e^{-\lambda} + k\log \lambda - \log k!}$$

$$= \log e^{-\lambda} + y^{(i)}\log\lambda - \log y!$$

$$= -\lambda + y^{(i)}\log\lambda - \log y!$$

substituting $\lambda = \theta^T x^{(i)}$

$$= -e^{\theta^T x^{(i)}} + y^{(i)}\log[e^{\theta^T x^{(i)}}] - \log y!$$

$$\Rightarrow LL(\theta) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta)$$

$$\Rightarrow = \sum_{i=1}^{m} -e^{\theta^T x^{(i)}} + y^{(i)}\log[e^{\theta^T x^{(i)}}] - \log y!$$

$$\frac{\partial}{\partial\theta_j} LL(\theta) = \frac{\partial}{\partial\theta_j}\sum_{i=1}^{m} -e^{+\theta^T x^{(i)}} + y^{(i)}\theta^T x^{(i)} - \log y!$$

$$= \sum_{i=1}^{m}[y^{(i)} - e^{+\theta^T x^{(i)}}]x_j^{(i)}$$

$$\Rightarrow \frac{\partial}{\partial\theta_j} LL(\theta) = \sum_{i=1}^{m}[y^{(i)} - h_\theta(x^{(i)})]x_j^{(i)}$$

where $\boxed{h_\theta(x^{(i)}) = e^{\theta^T x^{(i)}} = e^{\lambda}}$

Hence, $E[y^{(i)}|x^{(i)};\theta] = \lambda = e^{\theta^T x^{(i)}} = h_\theta(x^{(i)})$ mean of distribution $p(y|x;\lambda)$

Hence, proved