ML-Assignment-2

midigarahallishanm.r

September 2018

1 Question 1

1.1

Loss Function for logistic regression is given as

$$L(w) = \frac{-1}{N} \sum_{i=1}^{N} \log \left(\sigma \left(y_i W^T x_i \right) \right)$$

where $f(x) = \sigma(y_i W^T x_i)$

when $y = 1$ we have

$$L(w) = -\log \left(f(x) \right)$$

when $y = -1$ we have

$$L(w) = -\log \left(1 - f(x) \right)$$

Consider N samples (x_i, y_i) such that

$$x_i \in R^d$$

and

$$y_i \in R$$

The hypothesis function:

$$f(x) = \sigma \left(z_i \right) = \frac{1}{1 + e^{-z_i}}$$

where $z_i = W^T x_i$

Now consider:

$$1 - \sigma\Big(z\Big) = 1 - \frac{1}{1+e^{-z_i}} = \frac{1}{1+e^{z}} = \sigma(-z)$$

consider the loss function:

$$L = -t\log(\sigma(z_i)) - (1-t)\log(1-\sigma(z_i))$$

where $t = (y+1)/2$
Differentiating first term of the above equation we have :

$$\frac{\partial}{\partial w^T} log(\sigma(z_i)) = \frac{1}{\sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial w^T}$$

using chain rule we have:

$$= \frac{1}{\sigma(z_i)} \frac{\partial z_i}{\partial w^T} \frac{\partial \sigma(z_i)}{\partial z_i}$$

$$= \frac{1}{\sigma(z_i)} x_i \frac{\partial \sigma(z_i)}{\partial z_i} \text{ — Equation 1}$$

also we have:

$$\frac{\partial \sigma(z_i)}{\partial z_i} = \frac{\partial}{\partial z}(1+e^{-z})^{-1} = e^{-z}(1+e^{-z})^{-2}$$

$$= e^{-z}\frac{1}{(1+e^{-z})^2} = \frac{1}{(1+e^{-z})}\frac{e^{-z}}{(1+e^{-z})}$$

$$= \sigma(z)(1-\sigma(z)) \text{ — Equation 2}$$

substituting equation 2 in 1 we have :

$$\frac{\partial}{\partial w^T} log(\sigma(z_i)) = (1-\sigma(z_i))x_i - - - - - Equation 3$$

Considering the second term of the loss equation:

$$\frac{\partial log(1-\sigma(z_i))}{\partial w^T} = \frac{1}{1-\sigma(z_i)} \frac{\partial(1-\sigma(z_i))}{\partial w^T}$$

$$= \frac{-1}{1-\sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial w^T}$$

using chain rule we have:

$$= \frac{-1}{1-\sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial z_i} \frac{\partial z_i}{\partial w^T}$$

$$= \frac{-1}{1-\sigma(z_i)}\Big((\sigma(z_i)(1-\sigma(z_i)\Big)x_i$$

$$= -x_i\sigma(z_i) \text{ — equation 4}$$

Substituting equation 3 and 4 in loss function we have:

$$\frac{\partial L}{\partial w^T} = -t_i x_i(1-\sigma(z_i)) + (1-t_i)x_i\sigma(z_i)$$

$$= x_i(\sigma(z_i) - t_i)$$

Now Calculate the Hessian by taking the second derivative:

$$\frac{\partial L}{\partial w^T} = \frac{\partial x_i(\sigma(z_i) - t_i)}{\partial w^T}$$

$$= x_i\frac{\partial \sigma(z_i)}{\partial w^T}$$

using chain rule we have

$$= x_i\frac{\partial \sigma(\partial z_i)}{\partial z_i}\frac{\partial z_i}{\partial w^T}$$

$$= x_i\sigma(z_i)(1-\sigma(z_i))x_i^T$$

For N samples we have:

$$\sum_{i=1}^{N} x_i\sigma(z_i)(1-\sigma(z_i))x_i^T$$

Therefore,

$$H = XDX^T$$

where,

$$D = \sigma(z_i)(1-\sigma(z_i)) \text{ is a diagonal matrix}$$

1.2

The output of this Hessian function will be always positive as sigmoid function will return values between (0, 1)

i.e.

$$D = \sigma(z_i)(1-\sigma(z_i)) >= 0$$

and

$$X >= 0$$

This implies,

$$H >= 0$$

Consider any vector Z such that

$$ZHZ^T = ZXDX^TZ = ZXDX^TZ$$
$$= ||ZXD||^2 >= 0 \text{ (since } X >= 0 \text{ and } ||ZD|| >= 0)$$

This implies the loss function is convex

Therefore,

$$ZHZ^T >= 0$$

2 Question 2

To prove :

$$E_s[E_{out}(f(s)] = E_x[Bias(x) + Var(x)]$$

Given :

$$F[x] = E_s[F_s(x)]$$
$$E_{out}(f_s) = E_x[(f_s(x) - y(x))^2]$$
$$Bias(x) = (F(x) - y(x))^2$$
$$Var(x) = E_s[(f_s(x) - F(x))^2]$$

Now consider :

$$E_s[E_{out}(f(s)] = E_s[E_x[(f_s(x) - y(x))^2]]$$

$$= E_x[E_s[(f_s(x) - y(x))^2]] \text{ — Equation 1}$$

consider:

$$E_s[(f_s(x) - y(x))^2]$$

Adding and subtracting F(x), we have

$$= E_s[((f_s(x) - F(x)) + (F(x) - y(x)))^2]$$

Expanding using :

$$(a+b)^2 = a^2 + b^2 + 2ab$$

$= E_s[(f_s(x) - F(x))^2] + (F(x) - y(x))^2 + 2(E_s[(f_s(x) - F(x))](F(x) - y(x))$ – Equation 3

consider

$$E_s[(f_s(x) - E_s(f_s(x)))] = E_s[f_s(x)] - E_s[E_s(f_s(x))]$$

also we know that,

$$E_s[f_s(x)] = f_s(x)$$

$$E_s[E_s(f_s(x))] = f_s(x)$$

using $E(E(z)) = z$

Therefore the third term in the equation 3 is equal to 0 as

$$E_s[(f_s(x) - E_s(f_s(x)))] = f_s(x) - f_s(x) = 0$$

Now we have :

$$E_s[(f_s(x) - y(x))^2] = E_s[(f_s(x) - F(x))^2] + (F(x) - y(x))^2$$

By using the given equations in the above equation we have:

$$E_s[(f_s(x) - y(x))^2] = Bias(x) + Var(x) \text{ — Equation 2}$$

Substituting Equation 2 in 1, we have

$$E_s[E_{out}(f(s)] = E_x[Bias(x) + Var(x)]$$

Hence Proved!

3 Question 3

3.1

In general : According to the notion of VC-Dimension, the VC-Dimension of a hypothesis set H is the most data points H can shatter.

The largest data-set that is linearly-separable or that can be shattered when no more than (n+1) data points in the set are collinear is given by $d = (n+1)$

i.e. for dimension $d = 2$, $VC - Dimension = 3$ (for Example: XOR can not be shattered) for dimension $d = 3, VC - Dimension = 4$

Generalizing:
for n dimension, VC-Dimension = $(n+1)$

From this we get to know that the smallest data-set that is not linearly-separable when no more than $n+1$ data points are collinear(in case of 2D) or coplanar(in case of nD for $n > 2$) will be one more than VC-Dimension = $(n+1)+1$ = $(n+2)$

i.e. For $(n+2)$ there will be at-least one arrangement of data points that can be shattered.

Therefore the smallest data set that is not linearly separable in case of 2D will be $2+2$ = 4 and for 3D, it will be $3+2 = 5$.

3.2

Perceptron Learning algorithm will not converge in case the data-points are not linearly separable. Convergence in case of Perceptron Learning algorithm is when there are no points in the data-set that are misclassified. i.e. There is no combination of weights and bias that form a line(in case of 2D or a hyper-plane(in case of n-Dimension, where n $\geq$ 3) that can correctly classify the given data points.

4 Question 4

The probability density function of Gaussian distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For one-dimensional Guassian for each feature feature class is given by:

$$P(x_i|y) = \Pi_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$= (\frac{1}{\sigma\sqrt{2\pi}})^n \, e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Taking log on both the sides:

$$\log(P(x_i|y)) = \log((\frac{1}{\sigma\sqrt{2\pi})})^n e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2})$$

$$= n\log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= -\frac{n}{2}\log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Or

$$L(\mu, \sigma|x_i) = \log(P(x_i|y))$$

Differentiating L w.r.t mean value of the Gaussian distribution we have:

$$\frac{\partial}{\partial\mu}L = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(2x_i - 2\mu)$$

Equating this to zero

$$\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(2x_i - 2\mu) = 0$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}, \text{ Maximum Likelihood estimator for } \mu$$

Now, Differentiating L w.r.t σ of the Gaussian distribution we have:

$$\frac{\partial}{\partial\sigma}L = \frac{-n}{\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{n}{(\sigma^2)^2}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Equating this to zero we have:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

5 Question 5

Hinge Loss in terms of w is given as follows:

$$f(z) = max(0, 1 - yz)$$

Where $z = X.w$

Using Chain rule we have:

$$\frac{\partial}{\partial w_i} f(z) = \frac{\partial f(z)}{\partial z} \frac{\partial z}{\partial w_i})$$

First derivative of $z = X.w$ results in :

$$-y \text{ when } X.W < 1$$

$$0 \text{ when } X.W >= 1$$

Second derivative term is x_i
This can represented as :

$$\nabla_w L_{hinge} = \begin{cases} -yx_i & \text{if y X . w } < 1 \\ 0 & \text{if y X . w } >= 1 \end{cases}$$

Log Loss Equation is given as:

$$\frac{\partial(L_{loss})}{\partial w} = \frac{\partial}{\partial w} \log(1 + e^{-y_i f(x_i)})$$

$$= \frac{1}{1 + e^{-y_i w^T x_i - y_i b_i}} . \frac{\partial(1 + e^{-y_i w^T x_i - y_i b_i})}{\partial w}$$

$$= \frac{e^{-y_i w^T x_i - y_i b_i}}{1 + e^{-y_i w^T x_i - y_i b_i}} . \frac{\partial(e^{-y_i w^T x_i - y_i b_i})}{\partial w}$$

$$= \frac{e^{-y_i w^T x_i - y_i b_i}}{1 + e^{-y_i w^T x_i - y_i b_i}} (-x_i y_i)$$

$$\nabla_w L_{log} = \frac{-x_i y_i}{e^{y_i W^T x_i} + 1} = \frac{XY}{e^{YW^T X} + 1}$$

Given :

$$w_0 = 0, w_1 = 1 w_2 = 0$$

$y = [1, 1, -1]$

Given Bias for all the data points is Equal to 0
By using the above derived equations we have:

S(Given)	$\nabla_w L_{hinge}$	$\nabla_w L_{log}$
(1/2, 3)	(-1/2, -3)	$\frac{-1}{e^{0.5}+1}[\frac{1}{2} \quad 3]$
(2, -2)	0	$\frac{-1}{e^{2}+1}[2 \quad -2]$
(3, 1)	0	$\frac{1}{e^{3}+1}[3 \quad -1]$