Hw1-math

Mathematic Background (0.8%)

(a) A symmetric matrix $M \in \mathbb{R}^n$ is positive semi-definite if $\forall x \in \mathbb{R}^n$

$$x^T M x \geq 0$$

Now, given a matrix $A \in \mathbb{R}^{n \times n}$. Show that AA^T is a positive semi-definite matrix.

<pf>: $x^T(A^TA)x = (Ax)^T(Ax) = ||Ax||_2^2 \geq 0$

$\therefore AA^T$ is a positive semi-definite matrix.

(b) If $f(x_1, x_2) = x_1 \sin(x_2) \exp(-x_1 x_2)$, what is the gradient $\nabla f(x)$ of f?

Recall that $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$

<Ans>: $\dfrac{\partial f}{\partial x_1} = \sin(x_2) exp(-x_1x_2) - x_1x_2 sin(x_2) \exp(-x_1x_2)$

$$= (1 - x_1x_2) sin(x_2) \exp(-x_1x_2)$$

$$\frac{\partial f}{\partial x_1} = x_1 \cos(x_2) exp(-x_1x_2) - x_1x_2 sin(x_2) \exp(-x_1x_2)$$

$$= x_1 \exp(-x_1x_2)[\cos(x_2) - x_2 \sin(x_2)]$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} (1 + x_1) sin(x_2) \exp(-x_1x_2) \\ x_1 \exp(-x_1x_2)(\cos(x_2) - x_2 \sin(x_2)) \end{bmatrix}$$

(c) Given $X_1, \cdots, X_n$ are identically and independent (i.i.d.) Bernoulli distribution with parameter p. Please find the maximum likelihood estimator of p

<Ans>:

.

Closed-Form Linear Regression Solution (0.8%)

Suppose the linear regression model

$$y = X\theta + \epsilon y$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times d}, \theta \in \mathbb{R}^d$ **and** $\epsilon \in \mathbb{R}^n$**. Note that** $X_i \in \mathbb{R}^{1\times d}$**is the** i**-th row of** X **. Write** $\theta = [w_1, \cdots, w_d, b]^T$ **and** $X_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,m}, 1]$**. If the linear model has the bias term** b**, then** $d = m + 1$ **(denote** $x_{i,m+1} = 1$**). On the other hand,** $d = m$**. For simplicity, we assume no bias term.**

(a) Find the general optimal solution θ^* **that minimizes the weighted MSE:**

$$\sum \omega_i (y_i - X_i\theta)^2$$

<Ans>:

$$L(\theta) = (y_i - X_i\theta)^T \Omega (y_i - X_i\theta) = y_i^T \Omega y_i - \theta^T X_i^T \Omega y_i - y_i^T \Omega X_i \theta + \theta^T X_i^T \Omega X_i \theta$$

$$L(\theta + \Delta\theta) - L(\theta)$$

$$= (y - X\theta - X\Delta\theta)^T \Omega (y - X\theta - X\Delta\theta) - L(\theta)$$

$$= -\Delta\theta^T X^T \Omega y + \Delta\theta^T X^T \Omega X\theta - y^T \Omega X \Delta\theta + \theta^T X^T \Omega X \Delta\theta + \Delta\theta^T X^T \Omega X \Delta\theta$$

$$= -\Delta\theta^T (X^T \Omega y - X^T \Omega X\theta + X^T \Omega y - X^T \Omega \theta X) + c \; (\because X^T \Omega X \text{ is invertible})$$

$$v = 2X^T \Omega y - 2X^T \Omega X\theta$$

$$\rightarrow \theta^* = (X^T \Omega X\theta)^{-1} X^T \Omega y$$

(b) To avoid overfitting, we add a $L2$**-regularization term into the original loss function (MSE):**

$$\sum_i (y_i - X_i\theta)^2 + \lambda \sum_j w_j^2$$

Write down the matrix form of the loss function and find the general optimal solution of the θ^*

<Ans>:

$$L(\theta) = \left\|y - x\theta\right\|_2^2 + \lambda||w||^2 = y^T y - \theta^T x^T y - y^T x\theta + \theta^T x^T x\theta + \lambda I\theta\theta^T$$

$L(\theta + \Delta\theta) - L(\theta)$

$= -\Delta\theta^T x^T y - y^T x \Delta\theta + \theta^T x^T x \Delta\theta + \Delta\theta^T x^T x \theta + \Delta\theta^T x^T x \Delta\theta + \lambda I(\theta + \Delta\theta)(\theta + \Delta\theta)^T - \lambda\theta\theta^T$

$= \Delta\theta^T(-x^T y - x^T y + x^T x\theta + x^T x\theta) + c + \lambda I(\Delta\theta\theta^T + \theta\Delta\theta^T + \Delta\theta\Delta\theta^T)$

$= \Delta\theta^T(-x^T y - x^T y + x^T x\theta + x^T x\theta + 2\lambda I\theta) + c$

$= \Delta\theta^T(-2x^T y + (2x^T x + 2\lambda)\theta I) + c$

$v = -2x^T y + (2x^T x + 2\lambda I)\theta$

$\rightarrow \theta^* = (x^T x + \lambda I)^{-1} x^T y$

Logistic Sigmoid Function and Hyperbolic Tangent Function (0.8%)

Consider the logistic sigmoid function defined by,

$$\sigma(a) = \frac{1}{1 + exp(-a)}$$

and the hyperbolic tangent function defined by,

$$tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$$

(a) Show that these two functions are related by,

$$tanh(a) = 2\sigma(2a) - 1$$

<pf>:

$$2\sigma(2a) - 1 = \frac{2}{1 + e^{-2a}} - \frac{1 + e^{-2a}}{1 + e^{-2a}}$$
$$= \frac{1 - e^{-2a}}{1 + e^{-2a}} \times \frac{e^a}{e^a}$$
$$= \frac{e^a - e^{-2a}}{e^a + e^{-a}} = tanh(a)$$

(b) Show that a linear combination of logistic sigmoid functions of the form

$$y(x,w) = w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x-\mu_j}{s})$$

is equivalent to a linear combination of 'tanh' functions of the form

$$y(x,u) = u_0 + \sum_{j=1}^{M} u_j tanh(\frac{x-\mu_j}{2s})$$

and find the expressions to relate the new parameters $u_1, \cdots, u_M$ to the original parameters $w_1, \cdots, w_M$

<pf>:

$$\begin{aligned} y(x,u) &= u_0 + \sum_{j=1}^{M} u_j tanh(\frac{x-\mu_j}{2s}) \\ &= u_0 + \sum_{j=1}^{M} u_j [2\sigma(2 \times \frac{x-\mu_j}{2s}) - 1] \\ &= u_0 + \sum_{j=1}^{M} 2u_j \sigma(\frac{x-\mu_j}{s}) - u_j \\ &= u_0 - \sum_{j=1}^{M} u_j + \sum_{j=1}^{M} 2u_j \sigma(\frac{x-\mu_j}{s}) \end{aligned}$$

$$w_J = 2u_j \text{ for } u_1, \cdots, u_M \text{ and } w_1, \cdots, w_M$$

Noise and Regulation (0.8%)

<pf>:

$$
\begin{aligned}
L_{ss}(w,b) &= \mathbb{E}[\frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i+\eta_i)-y_i)^2] \\
&= \mathbb{E}[\frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i+\eta_i)^2-2f_{w,b}(x_i+\eta_i)y_i+y_i^2)] \\
&=
\end{aligned}
$$

Logistic Regression (0.8%)

(a) Suppose $w=[-1,2,-1,5]^T,\ x=[7,0,3,10]^T,\ \text{and } b=3$ **Please calculate the logistic regression prediction for the above particular example.**

<Ans>:

$$p(C_1|x)=\sigma(w^T+b)=\sigma(-7+0-3+50+3)=\sigma(43)=\frac{1}{1+e^{-43}}=1$$

$$p(C_2|x)=1-p(C_1|x)=0$$

$$\rightarrow C_1$$

(b) Given training data set $\{x_i,y_i\}_i^N=1$**, where** $y_i\in\{0,1\}$**. Suppose** N **observations are generated independent. Please write down the likelihood function of** $p(y\mid x)$ **in terms of** y_i**,** $f_{w,b}(x_i)$**, where** $y=[y_1,\cdots,y_N]^T$**.**

Moreover, write down the loss function $L(w,b)$ **defined as the negative of the log likelihood.**

<Ans>:

$$\text{for N indipendent. } p(y|x)=\textstyle\prod_i p(y_i|x_i)$$

$$p(y_1|x_i)=f_{w,b}(x_i)=\sigma(w^Tx_i+b)$$

$$L(w,b)=f_{w,b}(x_1)f_{w,b}(x_2)\cdots *$$

$$w^*,b^*=arg\min_{w,b}-\ln L(w,b)$$

$$
\begin{aligned}
\ln L(w,b) &= -\ln f_{w,b}(x_1)\ln f_{w,b}(x_2)\cdots \\
&= -[\hat{y}_N\ln f(x_N)+(1-\hat{y}_N)\ln(1-f(x_N))]
\end{aligned}
$$

(c) Derive the formula that describes the update rule of parameters in logistic regression with learning rate η (e.g. $w(t+1) \leftarrow w(t) - \cdots$). Note that the answer in terms of $w(t+1), w(t), f_w, b(x_i), x_i, y_i, \eta$