Machine Learning Assignment 3

B08611035

November 2020

1 Linear Regression

(1) **b.** 30

As the description shown, the expected in-sample error $E_{in}(\mathbf{w}_{lin}$ with respect to $\mathcal{D}$ is given by:

$$E_D[E_{in}(\mathbf{w}_{lin}] = \sigma^2(1 - \frac{d+1}{N})$$

And For $\sigma = 0.1$ and $d = 11$, what is the smallest number of examples N such that $E_D[E_{in}(\mathbf{w}_{lin}]$ is no less than 0.006?

$$E_D[E_{in}(\mathbf{w}_{lin}] = \sigma^2(1 - \frac{d+1}{N}) > 0.006$$

$$0, 1^2(1 - \frac{11+1}{N} > 0.006 \rightarrow \frac{12}{N} < 0.4 \rightarrow N > 30$$

Thus, from the inference above, we can find out that as N more than 30 so that $E_D[E_{in}(\mathbf{w}_{lin}]$ would no less than 0.006.

(2) **a.**

From the normal equation of linear regression gradient solving to the linear regression weights:

$$\mathbf{x^T x w} = \mathbf{x^T} y \rightarrow \mathbf{w} = (\mathbf{x^T x})^{\mathbf{-1}} \mathbf{x^T} y$$

In the linear algebra, if $\mathbf{x^T x}$ is invertible, then $\mathbf{w}$ has unique one solution. And if $\mathbf{x^T x}$ is singular (not invertible), then $\mathbf{w}$ has many optimal solutions. Thus, there at least one solution for the normal equation.

(3) **c.**

The operation of multiplying each of the n-th row by $1/n$ is equal to the original matrix dot product with a diagonal square matrix as below:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n} \end{pmatrix}_{n \times n}$$

so that

$$\mathbf{X}\cdot\mathbf{X_A} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & & \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n} \end{pmatrix} \cdot \begin{pmatrix} - & x_1 & - \\ & \vdots & \\ - & x_n & - \end{pmatrix} = \begin{pmatrix} - & x_1 & - \\ - & \frac{1}{2}x_2 & - \\ & \vdots & \\ - & \frac{1}{n}x_n & - \end{pmatrix}$$

Due to $\mathbf{X}$ is diagonal square matrix, it's invertible. And given $H = \mathbf{X_A}(\mathbf{X_A^T X_A})^{-1}\mathbf{X_A^T}$ where $\mathbf{X_A^T X_A}$ is invertible. After scaling, does $(\mathbf{X_i X_A}[(\mathbf{X_i X_A})^{\mathbf{T}}(\mathbf{X_i X_A})]^{-1}(\mathbf{X_i X_A})^{\mathbf{T}} = \mathbf{X_A}(\mathbf{X_A^T X_A})\mathbf{X_A^T}$

$$(\mathbf{X_i X_A}[(\mathbf{X_i X_A})^{\mathbf{T}}(\mathbf{X_i X_A})]^{-1}(\mathbf{X_i X_A})^{\mathbf{T}}$$

$$= \mathbf{X_i X_A}[\mathbf{X_A^T X_i^T X_i X_A}]^{-1}\mathbf{X_A^T X_i^T} = \mathbf{X_i X_A}[\mathbf{X_A^T X_i^2 X_A}]^{-1}\mathbf{X_A^T X_i^T}$$

if $\mathbf{X_A}$ is not invertible, then it will get that

$$\mathbf{X_i X_A}[\mathbf{X_A^T X_i^2 X_A}]^{-1}\mathbf{X_A^T X_i^T} \neq \mathbf{X_A}[\mathbf{X_A^T X_A}]^{-1}\mathbf{X_A^T}$$

Thus, due to the reason above, the answer is **c.**: multiplying each of the n-th row of x by $1/n$ (which is equivalent to scaling the n-th example by $1/n$).

2 Likelihood and Maximum Likelihood

(4) **e.** 4

(5) **a.** $(1/\hat{\theta})^N$

$$y_1, y_2, ..., y_N \sim^{i.i.d} \bigcup(0.\theta)$$

so it implies that:

$$f(y_i) = \begin{cases} 1/\theta & 0 \leq y_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

$$L(y|\theta) = \prod_{i=1}^{N} f(y_i|\theta) = \left(\frac{1}{\theta}\right)^N \prod_{i=1}^{N} I(0 \leq y_i \leq \theta)$$

for I defined as:

$$I(0 \leq y_i \leq \theta \begin{cases} 1 & 0 \leq y_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq y_1, y_2, ..., y_N \leq \theta$, and $\max(y_1, y_2, ..., y_N) \leq \theta$, so that for any $\hat{\theta} \geq \max(y_1, y_2, ..., y_n)$, its likelihood is:

$$L(y|\theta) = \left(\frac{1}{\hat{\theta}}\right)^N$$

3 Gradient and Stochastic Gradient Descent

(6) **b.** $err(\mathbf{w},\mathbf{x},y) = \max(0, -y\mathbf{w^Tx})$

$$w_{t+1} = w_t = \frac{\eta}{N}\sum_{n=1}^{N}[[\text{sign}(w_t^T x_n) \neq y_n]]y_n x_n$$

And consider that $y_n = \{-1, +1\}$ and what if $\text{sign}(w_t^T y_n) = \{-1, +1\}$. In the four output of the combination is either 0 or $y_n w_n^T x_n$. So that the error function could be:

$$err(\mathbf{w},\mathbf{x},y) = \max(0, -y\mathbf{w^Tx})$$

(7) **a.**

$$err_{exp}(\mathbf{w},\mathbf{x},y) = \exp(-y\mathbf{w^Tx}) = \exp\Big(-y(w_0x_0 + w_1x_1 +,,,+ w_nx_n)\Big)$$

The gradient of error function would be:

$$\nabla err_{exp} = \begin{pmatrix} derr/dw_0 \\ derr/dw_1 \\ \vdots \\ derr/dw_n \end{pmatrix} = \begin{pmatrix} -y\exp(-y\mathbf{w^Tx})x_0 \\ -y\exp(-y\mathbf{w^Tx})x_1 \\ \vdots \\ -y\exp(-y\mathbf{w^Tx})x_n \end{pmatrix}$$

$$= -y\exp(-y\mathbf{w^Tx})\begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix} = -y\exp(-y\mathbf{w^Tx})\mathbf{x}$$

Thus the negative gradient of error function is:

$$-\nabla err_{exp}(\mathbf{w},\mathbf{x},y) = y\mathbf{x}\exp(-y\mathbf{w^Tx})$$

4 Hessian and Newton Method

(8) **b.**

$$w \leftarrow u + v \Rightarrow v = w - u$$

$$\Rightarrow E(w) \approx E(u) + b_E(u)^T \cdot v + \frac{1}{2}v^T \cdot A_E(u)v$$

$$\nabla_v E_{in}(w) = b_E(u)^T + \frac{1}{2}\cdot 2A_E(u)v = b_E(u)^T + vA_E(u) = 0$$

$$\Rightarrow vA_E(u) = -b_E(u)^T \Rightarrow v = -b_E(u)^T \cdot \Big(A_E(u)\Big)^{-1}$$

Due to $A_E(u)$ is symmetric, its inverse also symmetric. Thus:

$$v = -b_E(u)^T (A_E(u))^{-1} = -(A_E(u))^{-1} b_E(u)$$

(9) **b.** $\mathbf{2x^Tx}/N$

$$E = E_{in} = \frac{1}{N}\sum_{i=1}^{N}(w^T x_n - y_n)^2$$

$$\Rightarrow E_{in} = \frac{1}{N}\left[(w_1x_{11} + w_2x_{12} + ... - y_1)^2 + ... + (w_1x_{N1} + w_2x_{N2} + ... - y_n)^2\right]$$

The second derivative for k from w_1 to w_d would be:

$$\frac{d^2E_{in}}{dw_k{}^2} = \frac{2}{N}[x_{1k}^2 + x_{2l}^2 + ... + x_{Nk}^2]$$

$$\frac{d^2E_{in}}{dw_kw_j} = \frac{2}{N}[x_{1j}x_{1k} + x_{2j}x_{2k} + ... + x_{Nj}x_{Nk}]$$

Thus,

$$A_E(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_1 \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_d}(\mathbf{w}) \\ \frac{\partial^2 E}{\partial w_2 \partial w_1}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_2^2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_2 \partial w_d}(\mathbf{w}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial w_d \partial w_1}(\mathbf{w}) & \frac{\partial^2 E}{\partial w_d \partial w_2}(\mathbf{w}) & \cdots & \frac{\partial^2 E}{\partial w_d^2}(\mathbf{w}) \end{bmatrix}$$

$$= \frac{2}{N}\begin{bmatrix} x_{11}^2 + x_{21}^2 +,,, + x_{N1}^2 & \cdots & x_{1d}x_{11} + x_{2d}x_{21}... \\ \vdots & \ddots & \vdots \\ x_{1d}x_{11} + x_{1d}x_{12} +,,, + x_{Nd}x_{N1} & \cdots & x_{1d}^2 + x_{2d}^2 ... + x_{Nd}^2 \end{bmatrix}$$

$$= \frac{2}{N}\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{Nd} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{12} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix} = \frac{2}{N}\mathbf{x^Tx}$$

5 Multinomial Logistic Regression

(10) **b.** $(h_k(\mathbf{x}) - [[y=k]])x_i$

minimizing the negative log likelihood is equivalent to minimizing an $E_{in}(\mathbf{w})$ that is composed of the following error function:

$$err(\mathbf{w}, \mathbf{x}, y) = -\ln h_y(\mathbf{x}) = -\sum_{j=1}^{k} [[y=j]] \ln h_j(\mathbf{x})$$

And, the matrix represents a hypothesis:

$$h_y(\mathbf{x}) = \frac{\exp(w_y^T x)}{\sum_{i=1}^{k} \exp(w_i^T x)}$$

Thus, the error function can be expressed like:

$$err(\mathbf{w}, \mathbf{x}, y) = -\sum_{j=1}^{k} [[y=j]] \ln h_j(\mathbf{x}) = \sum_{j=1}^{k} [[y=j]] \ln \frac{\exp(w_y^T x)}{\sum_{i=1}^{k} \exp(w_i^T x)}$$

$$\frac{\partial err(\mathbf{w}, \mathbf{x}, y)}{\partial w_{ik}} = -[[y=1]] \cdot \frac{1}{h_1(\mathbf{x}} \cdot \frac{-\exp(w_1^T x) \cdot \exp(w_k^T x) x_i}{(\sum_{i=1}^{k} \exp(w_i^T x))^2} - ...$$

$$... - [[y=k]] \cdot \frac{1}{h_k(\mathbf{x}} \cdot \frac{-\exp(w_k^T x) \cdot \exp(w_k^T x) x_i}{(\sum_{i=1}^{k} \exp(w_i^T x))^2}$$

$$= [[y=1]] \frac{1}{h_1(\mathbf{x}} \cdot h_1(\mathbf{x}) h_k(\mathbf{x}) x_i + ... + [[y=k]] \frac{1}{h_k(\mathbf{x}} \cdot h_k(\mathbf{x}) h_k(\mathbf{x}) x_i$$

$$= \sum_{j=1}^{k} [[y=j]] h_k(\mathbf{x}) x_i - [[y=k]] x_i = \left[\sum_{j=1}^{k} [[y=j]] h_k(\mathbf{x}) - [[y=k]] \right] x_i$$

$$= \left[h_k(\mathbf{x}) - [[y=k]] \right] x_i$$

(11) **e.** $w_2^* - w_1^*$

When $k=2$, error in Multinomial Logistic Regression (MLR) is equivalent to error in logistic regression:

$$err(\mathbf{w}, \mathbf{x}, y)|_{MLR} = -\ln h_y(\mathbf{x}) = err(\mathbf{w}, \mathbf{x}, y)|_{logistic} = -\ln \theta(y \mathbf{w}^{\mathbf{T}} \mathbf{x})$$

$$\Rightarrow h_y(\mathbf{x}) = \theta(y \mathbf{x}^{\mathbf{T}} \mathbf{x} = \frac{1}{1 + \exp(-y \mathbf{w}^{\mathbf{T}} \mathbf{x})}$$

In the case of $k = y_n = 1$ that $y'_n = 1$:

$$h_1(x_n) = \frac{1}{1 + \exp(w_2^{*T} - w_1^{*T}) x_n} = \theta(y'_n w^T x_n) = \frac{1}{1 + \exp(\mathbf{w}^{\mathbf{T}} \mathbf{x})}$$

$$w^T = w_2^{*T} - w_1^{*T} \Rightarrow w^T = (w_2^* - w_1^*)^T = w = w_2^* - w_1^*$$

In the case of $k = 2 = y_n$, so that $y_n' = 1$:

$$h_2(x_n) = \frac{1}{1 + \exp(w_1^* - w_2^*)^T x_n} = \theta(y_n w^T x_n) = \frac{1}{1 + \exp(-\mathbf{w^T} x_n)}$$

$$-w^T = (w_1^* - w_2^*)^T = w = w_2^* - w_1^*$$

6 Nonlinear Transformation

(12) **e.** $[-7, 0, 0, 2, -2, 3]$

Plot the five curve on graph, and mark these points will obtain **e.** obviously.

(13) **b.**$2((\log_2 d) + 1)$

$$(x_1, x_2, ..., x_d) \in \mathbb{R}^d \rightarrow_{\Phi_{(h)}} (1, x_k) \in \mathbb{R}^2$$

Each of $\Phi_{(k)}$ is a decision stump that:

$$\Phi_{(d)} : (1, x_d) \Rightarrow w_0 + w_d x_d \Rightarrow m_{\mathcal{H}_\lceil} = 2n$$

from the observation above. so that the VC dimension of that $d_{vc}(\bigcup_{k=1}^{d} H_k)$ will be the largest N for which $m_{\mathcal{H}}(N) = 2^N = N$.

$$2^N \leq m_{\mathcal{H}(N)=2Nd} \Rightarrow 2^{N-1} \leq Nd$$

$$N - 1 \leq \log_2 d + \log_2 N \leq l \log_2 d + \frac{N}{2}$$

$$\Rightarrow N \leq 2(\log_2 d + 1)$$

7 Programming*

(14) **d.** 0.60

(15) **c.** 1800

(16) **c.** 0.56

(17) **a.** 0.44

(18) **a.** 0.32

(19) **b.** 0.36

(20) **d.** 0.44