Subject 2014_Q2 MLE

Date April 12th

Attendants

$f: \mathbb{R}^d \to \{0, 1\}$

$$f(x) = \begin{cases} 0 & \text{if } g_\theta(x) < 0.5 \\ 1 & \text{otherwise} \end{cases}$$

where $g_\theta : \mathbb{R}^d \to [0, 1]$

(a) Let $x_1, \cdots x_i, \cdots x_n \sim p(x)$ i.i.d

independently and the same dist for each x_i

$$y_i \sim \text{Bern}(g_\theta(x_i))$$

Then $p(x_i, y_i) = p(x_i)\,\text{Bern}(g_\theta(x_i))$

$\perp\!\!\!\perp\ p(x_j)\,\text{Bern}(g_\theta(x_j))$

if $i \neq j$

$\therefore \mathcal{D} = \{(x_1, y_1) \cdots (x_i, y_i) \cdots (x_n, y_n)\}$

is drawn as independent and identically distributed

Subject 20H_Q2 MLE

Date 2

Attendants

(d) Negative log Conditional likelihood

$$\mathcal{L}(\theta) := -\log p(y_1 \dots y_n \mid x_1 \dots x_n, \theta)$$

$$= -\log \prod_{i=1}^{n} p(y_i \mid x_n, \theta)$$ by Q(a)

since $\log AB = \log A + \log B$

then

$$L(\theta) = -\sum_{i=1}^{n} \log \underbrace{p(y_i \mid x_i, \theta)}$$

$$= \mathrm{Bern}(g_\theta(x_i))$$

$$= -\sum_{i=1}^{n} \log \left\{ g_\theta(x_i)^{y_i} (1 - g_\theta(x_i))^{1-y_i} \right\}$$

$$= -\sum_{i=1}^{n} \left[y_i \log g_\theta(x_i) + (1-y_i) \log (1 - g_\theta(x_i)) \right]$$

Subject 20H_Q2 MLE

Date 3

Attendants

(c) This Negative log conditional likelihood is equivalent to a common loss function used for ERM.

→ Cross entropy loss.

$$J = -\sum_{k} q(k) \log(p(k))$$

$$k \in \{y_i, 1-y_i\}.$$

especially Binary Cross Entropy Loss in pytorch implementation.

Subject 20H_Q2 Date 4

Attendants

(d) Special case where g_θ is sigmoid func

$$g_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$$

derive the grad and Hessian of negative log likelihood.

$$\text{gradient } \nabla = \frac{\partial L}{\partial \theta}$$

$$= -\sum_{i=1}^{n} \left\{ y_i \underbrace{\left[\frac{\partial}{\partial \theta} \log g_\theta(x_i)\right]}_{①} + (1-y_i) \underbrace{\frac{\partial}{\partial \theta} \left[\log(1-g_\theta(x_i))\right]}_{②} \right\}$$

$$① = \frac{g_\theta(x_i)'}{g_\theta(x_i)} = \frac{1+e^{-\theta^T x}}{1} \times \frac{-xe^{-\theta^T x}}{(1+e^{-\theta^T x})^2}$$

$$\left(\frac{1}{1+e^{-\theta x}}\right)' = \frac{-xe^{-\theta^T x}}{(1+e^{-\theta x})^2}$$

$$\because (1+e^{-\theta^T x})' = (-x e^{-\theta^T x})$$

$$(e^{-3x})' = (-3x)' e^{-3x}$$

$$= \frac{-x(e^{-\theta^T x})}{(1+e^{-\theta^T x})}$$

$$= \frac{-x(1+e^{-\theta x}-1)}{(1+e^{-\theta^T x})}$$

Subject 2014-Q2 Date 5

Attendants

$$① = \frac{g_\theta(x_i)'}{g_\theta(x_i)} = -x_i \left\{ \frac{1+\exp(-\theta^T x)}{1+\exp(-\theta^T x)} - \frac{1}{1+\exp(-\theta^T x)} \right\}$$

$$= -x_i \left(1 - g_\theta(x_i) \right)$$

$$③ = \frac{-g_\theta(x_i)}{1-g_\theta(x_i)}' = \left\{ \frac{1}{1-\frac{1}{1+\exp(-\theta^T x)}} \right\} \times \left\{ \frac{-1}{1+\exp(-\theta^T x)} \right\}'$$

$$= \left\{ \cdot\cdot \right\} \times \left\{ \frac{+x\, e^{-\theta^T x}}{\left(1+\exp(-\theta^T x)\right)^2} \right\}$$

$$= \left(\frac{1}{\frac{\exp(-\theta^T x)}{1+\exp(-\theta^T x)}} \right) \times \left\{ \cdot\cdot \right\} = \left\{ \frac{1+\exp(-\theta^T x)}{\exp(-\theta^T x)} \right\} \times \left\{ \cdot\cdot \right\}$$

$$= \frac{x \exp(-\theta^T x)}{1+\exp(-\theta^T x)} \times \frac{1}{\exp(-\theta^T x)}$$

Subject 2014_Q2 Date 6

Attendants

(3) $= \dfrac{x_i}{1+\exp(-\theta^T x)} = x_i\, g_\theta(x_i)$

Then

$$G = \sum_{i=1}^{n} \left\{ y_i x_i (g_\theta(x_i) - 1) + (1-y_i) x_i g_\theta(x_i) \right\}$$

$$= \sum_{i=1}^{n} x_i \left\{ \underline{y_i g_\theta(x_i)} - y_i + g_\theta(x_i) - \underline{y_i g_\theta(x_i)} \right\}$$

$$= \sum_{i=1}^{n} x_i (g_\theta(x_i) - y_i)$$

$$H = \frac{\partial G}{\partial \theta} = \sum_{i=1}^{n} x_i \times \frac{\partial}{\partial \theta} g_\theta(x_i)$$

$$= \sum_{i=1}^{n} x_i \times \frac{-x_i \exp(-\theta^T x_i)}{(1+\exp(-\theta^T x_i))^2}$$

$$= -\sum_{i=1}^{n} x_i^T x_i \times \left(\underline{1} - \frac{1}{1+\exp(-\theta^T x_i)} \right) \times g_\theta(x_i)$$

Subject 20H_Q2 MLE 1(d) Date 7

Attendants

$$H = \sum_{i=1}^{n} x_i^T x_i \left(1 - g_\theta(x_i)\right) g_\theta(x_i)$$

$$x_i \in \mathbb{R}^d$$

Let $X = \begin{pmatrix} x_1 \\ x_2 \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times d}$ n row (行) d column (列))

$$\underset{d \times n}{H = X^T} \underset{n \times n}{\mathrm{diag}\left(g_\theta(x_i)\left(1 - g_\theta(x_i)\right)\right)} \underset{n \times d}{X}$$

Let $D \in \mathbb{R}^{n \times n} = \mathrm{diag}\left(g_\theta(x_i)(1 - g_\theta(x_i)\right)$

$$= \begin{pmatrix} g_\theta(x_1)(1 - g_\theta(x_1)) & \cdots & 0 \\ & \ddots & \\ 0 & & \underbrace{g_\theta(x_n)(1 - g_\theta(x_n))}_{\geq 0} \end{pmatrix}$$

≥ 0 because this is probability

Subject 20H_Q2 MLE

Date 8 April 12th

Attendants

(e) show NLCL is convex

$$H = X^T \underbrace{\mathrm{diag}\left(g_\theta(x_i)(1-g_\theta(x_i))\right)}_{=D} X$$

where D is positive semidefinite

$$H = X^T D X \in \mathbb{R}^{d \times d}$$

let $u \in \mathbb{R}^d$

$$u^T H u = u^T X^T D X u$$

$$= (uX)^T D X u$$

$$= \| D^{\frac{1}{2}} (uX) \|^2 \geq 0$$

then H is also positive definite, NLCL is convex

Subject 20H_Q2 MLE 2 Date 9

Attendants

S is noisy in the sense that for each example (x_i, y_i), the label y_i is flipped w/ proba b_i (b_i's are known constants)

(a) examples in S independent? identically distributed?

$$y_i = \begin{cases} 1 & g_\theta(x_i)(1-b_i) \\ 0 & g_\theta(x_i)\, b_i \\ 1 & (1-g_\theta(x_i))\, b_i \\ 0 & (1-g_\theta(x_i))(1-b_i) \end{cases}$$

independent but not identical.

$$y_i = \begin{cases} 1 & \text{w/ prob} \quad g_\theta(x_i)(1-b_i) + (1-g_\theta(x_i))b_i \\ 0 & \text{w/ prob} \quad g_\theta(x_i)b_i + (1-g_\theta(x_i))(1-b_i) \end{cases}$$

(b) What is the distribution of y_i given x_i?

Subject 20H_Q2 MLE Date 10

Attendants 2.(b)

$$p(y_i|x_i) = \begin{cases} 1 & \text{with probability } g_\theta(x_i)(1-b_i) + (1-g_\theta(x_i))b_i \\ 0 & \text{with probability } g_\theta(x_i)b_i + (1-g_\theta(x_i))(1-b_i) \end{cases}$$

(c) Write down the negative log conditional likelihood in this case.
& Comment on the relation between this objective and the one drived in question 1(b)

$$\mathcal{L}(\theta) = -\log p(y_1 \cdots y_n | x_1 \cdots x_n, \theta)$$

$$= -\log \prod_{i=1}^{n} p(y_i|x_i,\theta) = -\sum_{i=1}^{n} \log p(y_i|x_i,\theta)$$

($g_\theta(x_i) \to g_i$ as shorter notation)

$$= -\sum_{i=1}^{n} \log \left\{ (g_i - 2g_ib_i + b_i)^{(1-y_i)} + (1-b_i-g_i+2g_ib_i)^{y_i} \right\}$$

Subject 2014 – Q2 MLE Date / /

Attendants April 12th

2(c) continue...

$$L(\theta) = -\sum_{i=1}^{n}\left(y_i \log\left(g_i - 2b_i g_i + b_i\right) + (1-y_i)\log\left(1-b_i-g_i+2b_i g_i\right)\right)$$

i-th data (label) point contribute by choosing which term will be ignored.

when $b_i = 0$

$$L_{b_i=0}(\theta) = -\sum_{i=1}^{n}\left(y_i \log(g_\theta(x_i)) + (1-y_i)\log(1-g_\theta(x_i))\right)$$

is equivalent to Q1.(b)

when $b_i = \frac{1}{2}$

$$L_{b_i=\frac{1}{2}}(\theta) = -\sum_{i=1}^{n}\left(y_i \log\left(\frac{1}{2}\right) + (1-y_i)\log\left(\frac{1}{2}\right)\right)$$

$$= \sum_{i=1}^{n}\left(y_i \log 2 + (1-y_i)\log 2\right)$$

$$= \sum_{i=1}^{n} \log 2 = n\log 2 \quad (\text{constant})$$

⇒ We cannot minimize this objective

Subject 20 H_Q2 Date 12

Attendants April 12th

2 (d)

$b = b_1 = b_2 = \cdots b_n$

Propose a learning algorithm to jointly estimate the model params θ and flipping probability b.

$$\frac{\partial L(\theta)}{\partial b} = \frac{d}{db}\left\{-\sum_{i=1}^{n}\left(y_i \log\left(g_i - 2b_i g_i + b_i\right) + (1-y_i)\log\left(1-b_i-g_i+2b_i g_i\right)\right)\right\}$$

$$= -\sum_{i=1}^{n} y_i \frac{(-2g_i+1)}{(g_i-2b_i g_i+b_i)} + (1-y_i)\frac{(2g_i-1)}{(1-b_i-g_i+2b_i g_i)}$$

$$= -\sum_{i=1}^{n}(2g_i-1)\left\{\frac{-y_i}{(g_i-2b_i g_i+b_i)} + \frac{1-y_i}{1-b_i-g_i+2b_i g_i}\right\}$$

---

not necessary?

Subject 20H_Q2

Date 13

Attendants

April 24th

$$\frac{\partial L(\theta)}{\partial b} = 0$$

$\Leftrightarrow$

$$0 = -\sum_{i=1}^{n} (2g_i - 1) \left\{ \frac{-y_i}{(g_i - 2b_i g_i + b_i)} + \frac{1 - y_i}{(1 + b_i - g_i + 2b_i g_i} \right\}$$

$\Leftrightarrow$ $\times (g_i - 2b_i g_i + b_i)(1 + b_i - g_i + 2b_i g_i)$

$$0 = \sum_{i=1}^{n} (2y_i - 1) \left\{ g_i - 2b_i g_i + b_i - \cancel{g_i y_i} + \cancel{2b_i g_i y_i} - b_i y_i - y_i - b_i y_i + \cancel{g_i y_i} - \cancel{2b_i g_i y_i} \right\}$$

$\Leftrightarrow$

$$0 = \sum_{i=1}^{n} (2y_i - 1) \left\{ g_i - y_i + b_i - 2b_i y_i - 2b_i g_i \right\}$$

$\Leftrightarrow$

$$0 = \sum_{i=1}^{n} (2y_i - 1) \left\{ (g_i - y_i) + b_i (1 - 2y_i - 2g_i) \right\}$$

... ?

← not necessary?

Subject 2014_Q2 Date 14 April 12

Attendants

$$\frac{\partial}{\partial\theta}L(\theta) = \frac{\partial}{\partial\theta}\left\{-\sum_{i=1}^{n}\left(y_i\log\left(g_i-2b_ig_i+b_i\right)+(1-y_i)\log\left((1-b_i-g_i+2b_ig_i)\right)\right)\right\}$$

$$= -\sum_{i=1}^{n} y_i\left(\frac{(g_i)'-2b_i(g_i)'}{g_i-2b_ig_i+b_i}\right)+(1-y_i)\left(\frac{-(g_i)'+2b_i(g_i)'}{1-b_i-g_i+2b_ig_i}\right)$$

Algorithm

step 0 : random initialization of $b\in[0,1]$

step 1 : gradient ascent wrt $\frac{\partial L}{\partial\theta}$ and update θ

step 2 : update $b = \underset{b}{\operatorname{argmax}} \frac{\partial L(\theta)}{\partial b}$

Then loop step 1 and step 2.

↑ Answer.

Subject | Date

Attendants