Subject:
Year. Month. Date. ( )

«Home Work 3»

Kimia Noorbakhsh
[redacted]

1)

first statement: True. we know that $bias(\hat{y}(x)) = E(\hat{y}(x)) - y(x)$ and therefore, increasing the training set size would not affect this term. we should chang our model.

counterexample: assume you have some data points $x_1, ..., x_N$ and you want to estimate ther mean. $\hat{y}(x) = 6$ is an estimator which has high bias, but nothing will happen if you increase N.

second statement: False, it does not happen alway.

For example, if our model is overfitting, then we will have a great training accuracy, but our test error might even increase.

third statement: True, when we increase complexity, we are more likely to overfit which means we would get great accuracy on training data, and poor accuracy on our test data. such model is good at remmembering but not generalizable.

Forth statement: False, if 6th degree polynomial fits our data then it means our data distribution is not linear. may be 3 or 4th order polynomial would be better in this case.

Subject.
Year. Month. Date. ( )

2)

2.1) $\tanh(a) \stackrel{?}{=} 2\sigma(2a) - 1$

From the definition, we know that : $\tanh(a) = \frac{\sinh(a)}{\cosh(a)} = \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}}$

$$= \frac{e^{2a} - 1}{e^{2a} + 1} \quad ①$$

also we have: $\sigma(2a) = \frac{1}{1 + e^{-2a}}$

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

we need to prove ① $\stackrel{?}{=}$ ② $\Leftrightarrow \frac{e^{2a} - 1}{e^{2a} + 1} \stackrel{?}{=} \frac{1 - e^{-2a}}{1 + e^{-2a}}$

$$\Leftrightarrow (e^{2a} - 1)(1 + e^{-2a}) \stackrel{?}{=} (e^{2a} + 1)(1 - e^{-2a})$$

$$\Leftrightarrow e^{2a} - 1 + 1 - e^{-2a} \stackrel{?}{=} e^{2a} + 1 - 1 - e^{-2a}$$

$\Rightarrow$ therefore, the desired equality holds. ✓

2.2) let $\frac{(x - \mu_j)}{2s} = \alpha_j$

$$\Rightarrow y(x, w) = w_0 + \sum_{j=1}^{M} w_j \sigma(2\alpha_j)$$

$$= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} (2\sigma(2\alpha_j) - 1 + 1)$$

(from 2.1) $= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} (\tanh(\alpha_j) + 1)$.

$$= u_0 + \sum_{j=1}^{M} u_j \tanh(\alpha_j)$$

✓

such that : $u_0 = w_0 + \sum_{j=1}^{M} \frac{w_j}{2}$, and $u_j = \frac{w_j}{2} \quad 1 \leqslant j \leqslant M$

PAPCO

Subject,
Year. Month. Date. ( )

3) $$\arg\max_{w} p(w|y) = \arg\max_{w} \left(\ln p(w) + \ln p(y|w) \right)$$

$$p(w) = \frac{1}{\sqrt{2\pi}\,\alpha^{-1/2}} \exp\left(-\frac{1}{2} w^T \alpha w\right)$$

$$\Rightarrow p(w) = \frac{\alpha^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\alpha}{2} w^T w\right) \Rightarrow \ln p(w) = \frac{1}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\|w\|^2$$

$$p(y|w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\beta^{-1/2}} \exp\left(-\frac{\beta}{2}(y_i - w^T x_i)^2\right)$$

$$\Rightarrow \ln p(y|w) = \frac{n}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

$$\Rightarrow \arg\max_{w} p(w|y) = \arg\max_{w}\left(\overbrace{\frac{1}{2}\ln\left(\frac{\alpha}{2\pi}\right) + \frac{n}{2}\ln\left(\frac{\beta}{2\pi}\right)}^{\text{(independent to } w\text{)}} - \frac{\alpha}{2}\|w\|^2 - \frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 \right)$$

$$= \arg\max_{w} -\frac{\alpha}{2}\|w\|^2 - \frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

$$= \arg\min_{w} +\frac{\alpha}{2}\|w\|^2 + \frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

$$= \arg\min_{w} \underbrace{\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2}_{L(w)} + \underbrace{\frac{\alpha}{\beta}}_{\lambda} \times \underbrace{\frac{1}{2} w^T w}_{R(w)} \Rightarrow \checkmark$$

Subject: Year. Month. Date. ( )

4)

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

4.1) $y | x_1, x_2 \sim N(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2, \sigma^2)$

$$\Rightarrow p(y|x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - w_0 - w_1 x_1 - w_2 x_2 - w_3 x_1^2 - w_4 x_2^2)^2}{2\sigma^2} \right)$$

✓

4.2) The conditional log likelihood is:

$$l(w_0, w_1, w_2, w_3, w_4) = \log \prod_{i=1}^{n} p(y^{(i)} | x_1^{(i)}, x_2^{(i)}) = \sum_{i=1}^{n} \log p(y^{(i)} | x_1^{(i)}, x_2^{(i)})$$

from 4.1 $= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{n} \log \exp\left(\frac{-(y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_1^{(i)2} - w_4 x_2^{(i)2})^2}{2\sigma^2} \right)$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_1^{(i)2} - w_4 x_2^{(i)2})^2$$

drop consts $\cong - \sum_{i=1}^{n} (y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_1^{(i)2} - w_4 x_2^{(i)2})^2$ ✓

4.3) we need to maximize the conditional log likelihood and it is equivalent to minimizing the negative conditional log likelihood, therfore we have:

$$f(w_0, w_1, w_2, w_3, w_4) = -l(w_0, w_1, w_2, w_3, w_4)$$

from 4.2 $= \sum_{i=1}^{n} (y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_1^{(i)2} - w_4 x_2^{(i)2})^2$

✓

← continued

4 - continued:

4.4) the gradient is: $\nabla_w f(w) = \left[\frac{\partial f}{\partial w_0} \ \frac{\partial f}{\partial w_1} \ \frac{\partial f}{\partial w_2} \ \frac{\partial f}{\partial w_3} \ \frac{\partial f}{\partial w_4} \right]^T$ ⊛

now we have: $\frac{\partial f}{\partial w_0} = -2 \sum_{i=1}^{n} (y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_3^{(i)2} - w_4 x_4^{(i)2})^2$

$$\frac{\partial f}{\partial w_1} = -2 \sum_{i=1}^{n} x_1^{(i)} (y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_3^{(i)2} - w_4 x_4^{(i)2})^2$$

$$\frac{\partial f}{\partial w_2} = -2 \sum_{i=1}^{n} x_2^{(i)} (y^{(i)} - w_0 - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_3^{(i)2} - w_4 x_4^{(i)2})^2$$

$$\frac{\partial f}{\partial w_3} = -2 \sum_{i=1}^{n} x_1^{(i)2} (y^{(i)} - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_3^{(i)2} - w_4 x_4^{(i)2})^2$$

$$\frac{\partial f}{\partial w_4} = -2 \sum_{i=1}^{n} x_2^{(i)2} (y^{(i)} - w_1 x_1^{(i)} - w_2 x_2^{(i)} - w_3 x_3^{(i)2} - w_4 x_4^{(i)2})^2$$

now we can combine the above equations in ⊛ ✓
to get $\nabla_w f(w)$

4.5) The update rule is $\boxed{w := w - \alpha \nabla_w f(w)}$
which α is the learning rate or the step size ✓

---

5) 5.1) Assume that we have d features, then we have:

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^{d+1}, \quad x^{(i)} = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\Rightarrow X := \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}_{(d+1) \times (d+1)} = \begin{bmatrix} 1 & x^{(1)T} \\ 1 & x^{(2)T} \\ \vdots & \\ 1 & x^{(n)T} \end{bmatrix}$$

continued →

5.1) continued

now we have $J(w) = \sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)})^2$

for optimizing, we should have $\nabla_w J = 0$ and therefore:

$$\Rightarrow \sum_{i=1}^{n} \nabla_w (\underbrace{y^{(i)2}}_{\text{(independent on } w)} - 2w^T x^{(i)} y^{(i)} + \underbrace{w^T x^{(i)} x^{(i)T} w}_{=(w^T x^{(i)})^2}) = 0$$

$$\Rightarrow -\sum_{i=1}^{n} 2x^{(i)} y^{(i)} + \left(\sum_{i=1}^{n} 2x^{(i)} x^{(i)T}\right) w = 0 \quad \circledast$$

[we know that $(A + A^T)P = \nabla P^T A P$
$\Rightarrow \nabla w^T x^{(i)} x^{(i)T} w = (2x^{(i)} x^{(i)T}) w$]

$$\circledast \Rightarrow w_{OPT} = \left(\sum_{i=1}^{n} x^{(i)} x^{(i)T}\right)^{-1} \left(\sum_{i=1}^{n} x^{(i)} y^{(i)}\right)$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \rightarrow w_{OPT} = (X^T X)^{-1} X^T y$$ ✓

5.2) by adding the regularization term we would have

$$J(w) = \|y - Xw\|^2 + \lambda \|w\|^2$$

we know that $\|X\|^2 = X^T X$, therefore:

$$J(w) = (y - Xw)^T (y - Xw) + \lambda w^T w$$

$$\Rightarrow \nabla_w J = -2X^T y + 2X^T X w + 2\lambda w = 0$$

continued

$$\Rightarrow w_{OPT} = (\lambda I + X^T X)^{-1} X^T y$$ ✓

Subject:
Year. Month. Date. ( )

5.3) continued

5.3) Define matrix F such as

$$F = \begin{bmatrix} F_1 & & 0 \\ & F_2 & \\ 0 & & \ddots \ F_n \end{bmatrix} \Rightarrow F = F^T$$

$$\Rightarrow J(w) = \underset{1\times n}{(y-xw)^T} \underset{n\times n}{F} \underset{n\times 1}{(y-xw)}$$

$$\Rightarrow J(w) = y^T F y - y^T F x w - w^T x^T F y + w^T x^T F x w$$

$$\Rightarrow \nabla_w J(w) = -2x^T F y + 2x^T F x w = 0$$

$$\Rightarrow w_{OPT} = (x^T F x)^{-1} x^T F y$$

---

6) because we can choose $y(x)$ independent of x, the minimum of $E[L_q]$ can be found by minimizing this: $\int |t-y(x)|^q p(t|x)\,dt$ for each x value.

set derivative to 0

$$0 = \int q|t-y(x)| \operatorname{sign}(t-y(x))\, p(t|x)\,dt$$

$$\Rightarrow \int_{-\infty}^{y(x)} q|t-y(x)|^{q-1} p(t|x)\,dt = \int_{y(x)}^{\infty} q|t-y(x)|^{q-1} p(t|x)\,dt \quad \circledast$$

and $\circledast$ is the desired condition.

if $q=1 \Rightarrow \int_{-\infty}^{y(x)} p(t|x)\,dt = \int_{y(x)}^{\infty} p(t|x)\,dt \Rightarrow$ $y(x)$ must be the conditional median of t

continued

Subject:
Year. Month. Date. ( )

6_continued

if $q \to 0$ $\Rightarrow |t-y(x)|^q \to 1$

in every point except around $t=y(x)$ which is 0

$\Rightarrow$ the value of $\int |t-y(x)|^q p(t|x)\,dt$ is close to 1 but decreased close to $t=y(x)$.

most reduction is in the biggest value of $p(t)$

$\simeq$ conditional mode. ✓