(۱) الف)

$$P(y|x;w) \overset{iid}{=} \prod P(y_i|x_i;w) \xrightarrow{\log} \sum \begin{cases} f(x,w) & y=+1 \\ 1-f(x,w) & y=-1 \end{cases} =$$

این دو تا یک چیز هستند

$$\frac{1}{2}\log \sum \left(1+y_n\right) \log f(x_n;w) + \left(1-y_n\right) \log f(x_n;w))$$

این دو تا یک چیز را نشان می‌دهند زیرا اگر $y_n=+1$ یا $y_n=-1$ را جایگذاری کنیم می‌شود رابطه بالا.

سوال ۱ ب

$$\frac{dg}{dx} = \frac{e^x}{1+e^x} \xrightarrow{\text{مشتق بگیریم داریم}} \frac{e^x(e^x+1) - e^x(e^x+0)}{(e^x+1)^2} =$$

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

$e^x \geq 0$ همگی

$(e^x+1)^2 \geq 0$ همگی

$x > 0 \longrightarrow e^x > 0$

$x < 0 \longrightarrow e^{-x} \longrightarrow \frac{1}{e^x} > 0$

$$g\circ f = g(f(x)) \Longrightarrow \frac{dg}{dx} = g'(f(x))\, f'(x) = k$$

$$\frac{dk}{dx} = \underbrace{g'(f(x))}_{\text{متعودی}}\, \underbrace{f''(x)}_{\text{معدب}} + \underbrace{g''(f(x))}_{\text{محدب}}\, \underbrace{\frac{f'(x)\, f'(g)}{(f(x))^2}}_{\geq 0}$$

(۱) (ب۲)

مشتق ۲ بگیریم $K = \frac{dh}{dx} = \sum_{k=1}^{K} c_k f'_k \longrightarrow \frac{d^2h}{dx^2} = \sum_{k=1}^{K} c_k f''_k =$ محدب

می‌دانیم $c_k \geq 0$ ← فرض سوال

$f''_k \geq 0$ ← محدب

(۱) قسمت ۵ $\log(1+e^{w^T x_n})$ محدب است با توجه به این که قضیه ۱ ب که ثابت کردیم.

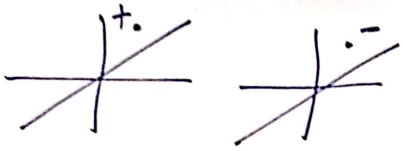
$$\frac{-1}{2}\sum \underbrace{(1+y_n)}_{\text{مثبت}} \underbrace{\log(1+e^{w^T x_n})}_{\text{محدب}} + \underbrace{(1-y_n)}_{\text{مثبت}} \underbrace{\log(1+e^{w^T x_n})}_{\text{محدب}}$$

محدب است

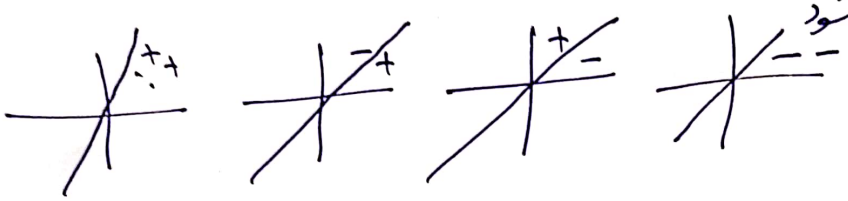
مقعر

زیرا ترکیب خطی هم متقاعدیم محدب است هر متقی محدب مقعر می شود

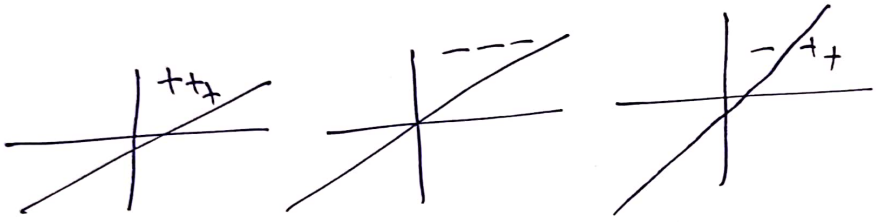
(۲) الف) خب با یک نقطه که ۲ حالت می‌شود



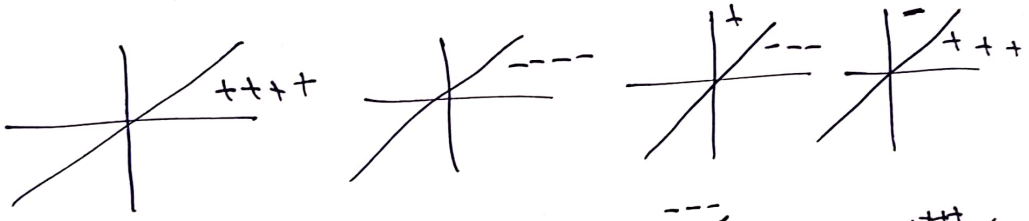
دو نقطه را ۴ حالت می‌شود

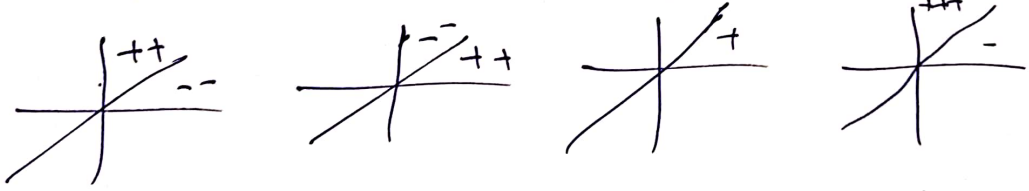


۳ نقطه می‌شود ۸ حالت



۴ نقطه ۱۶ حالت است





پس می‌شود 2^N تابع رشد

(۲) (ب) زیرا توابع ثابت داریم هرچیکه باشد هردوتایی که داشته باشد

بهرحال ۲ تا فرضیه بیشتر اضافه نمی‌توانند بکنند که نهایتا اگر شود ۲ تا اضافه می‌کنیم.

با یک نقطه که دوحالت بیشتر نمی‌تواند بشود

با دو نقطه که ۴ حالت بیشتر نمی‌تواند بشود

با ۳ نقطه ۶ حالت بوجود و ۲ حالت اضافه می‌شود که می‌شود ۸

با ۴ نقطه ۸ حالت بوجود و ۲ حالت اضافه می‌شود که می‌شود ۱۰ حالت.

(۲) (ب) پس بعد VC این سوال می‌شود ۳ و برای بهینت ۴ می‌شود $k=4$

پس داریم.

$$4 m_H(2N) e^{-\frac{1}{8}\varepsilon^2 N} \le \delta$$

$$4 (2N)^3 e^{-\frac{1}{8}\times\frac{1}{2}N} = 4\times 8 N^3 e^{-\frac{1}{16}N}$$

و احتمال می‌شود

$$1 - 32 N^3 e^{-\frac{N}{16}}$$

(2) (ت) ندو واقع همان سوال ۲ ت است یعنی وقتی مشتق بگیریم داریم.

$$\frac{\partial \sum |h-y_n|}{\partial y_n} = \sum \operatorname{sgn}(h-y_n)$$

$$\frac{\partial |h-y_n|}{\partial y_n} = \begin{cases} -1 & h < y_n \\ \infty & h = y_n \\ 1 & h > y_n \end{cases}$$

که ندو واقع دارد میانه را حساب می‌کند.

۳) الف) زیرا مجموع مربعات در واقع فاصله را اندازه می‌گیرد و در مسئله کلاس‌بندی فاصله مهم نیست مثلا
۱+ و ۵+ هردو در یک کلاس دسته‌بندی می‌شوند و در واقع فاصله مثلاً ۱+ و ۱- باید بیشتر از ۱+ و ۱۰+
باشد که مجموع مربعات خطا این را خوب انجام نمی‌دهد.
البته فکر می‌کنم در سؤال اشتباه تایپی وجود دارد و باید ~~[illegible]~~ مجموع مربعات در نظر گرفته شود.»
~~[illegible]~~

۳) ب) در واقع این روش یک soft thresholding در نظر گرفته که خیلی بهتر از hard threshold یا همان sgn(x) است
علت این موضوع آن است که در رگرسیون لاجستیک ما نیاز به احتمال تعلق به هریک از دسته‌ها داریم و روش‌های sgn(a) و غیره صرفا کلاس را برمی‌گردانند و بصورت کلی بهتر است از cross entropy استفاده کنیم.
البته مشتق‌پذیری یا smooth بودن تابع هم مهم است.

$$\frac{-2}{N}\sum_{n=1}^{N}(x_n-\theta)=0 \Rightarrow \frac{-2}{N}\sum_{n=1}^{n}x_n+\frac{2N\theta}{N}=0$$

$$\theta^* = \frac{\sum_{n=1}^{N} x_n}{N}$$

(۳) (ب) (۱) بیشتر تمرین.

(۳) (ب) (۱) مشتق

$$\frac{\partial |x_n-\theta|}{\partial \theta} = \begin{cases} 1 & x_n>\theta \\ \infty & x_n=\theta \\ -1 & x_n<\theta \end{cases} = sgn(x_n-\theta)$$

$$\frac{\partial L_{MAE}(\theta)}{\partial \theta} = \frac{1}{N}\sum sgn(x_n-\theta)$$

در واقع داریم به دنبال مقداری می گردیم که دنباست ما دوقسمت

σ^2 پس همان میانه θ^* ما خواهد بود

(3) (ب) (2) روش MSE ← در واقع روش خوبی است برای این که مدل آموزش داده شده داده outlier نداشته باشد زیرا توان آن خیلی زیاد می شود.

بدی آن نیز آن است که اگر یک پیش بینی خیلی بد داشته باشیم آن توان ۲ خطا را خیلی زیاد می کند

MAE دقیقاً همان مشکلات MSE را حل می کند یعنی خطا را به اندازه خود خطا تأثیری دهد.

و مزایای MSE را هم ندارد یعنی اگر داده outlier داشته باشیم آن را حذف نمی کند

۳ ب ۲

برای خواهیم MLE

$$L(w) = \prod_{i=1}^{N} P(y_i | x_i, w) \Rightarrow \log L(w) = \sum_{i=1}^{N} P(y_i | x_i, w)$$

در این سوال دو حالت می‌توان گفت اما سوال ۱ الف است که باید دستی بگیریم و سعی
صفر قرار دهیم که وقت ست!

۴ (الف) اگر تعداد دیتا کم باشد بهتر است از روش‌های ساده‌تر (مدل‌های ساده‌تر) استفاده کنیم و اگر تعداد دیتا زیاد باشد می‌توان با مدل‌های پیچیده‌تر نیز جواب‌های خوب گرفت اما باید N عدد نسبتاً بزرگ باشد.

۴ (ب)

$$x_1 w_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} -3 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 5 \end{bmatrix} \qquad x_2 w_2 = \begin{bmatrix} -2 \\ 3 \end{bmatrix} \begin{bmatrix} -2 & 2 \\ 3 & -1 \end{bmatrix} =$$

$$w_1 x_1 = \begin{bmatrix} -3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 5 \end{bmatrix} \qquad w_2 x_2 = \begin{bmatrix} -2 & 2 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ -9 \end{bmatrix}$$

$$\max(w_1 x_1, 0) = \begin{bmatrix} 0 \\ 5 \end{bmatrix} \qquad \max(w_2 x_2, 0) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}$$

$$\tanh\left(\begin{bmatrix} 0 \\ 05 \end{bmatrix} + \begin{bmatrix} 10 \\ 0 \end{bmatrix}\right) = \tanh\left(\begin{bmatrix} 10 \\ 0.5 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 0.46 \end{bmatrix}$$

$P_1 = W_1 x_1$ $\quad \max P_1 = \max(P_1, 0) \longrightarrow$ Fraction $= \frac{1}{\max P_1}$

$P_2 = W_2 x_2$ $\quad \max P_2 = \max(P_2, 0)$

Sum = Fraction + $\max P_2$, tanResult = $\tanh(\text{sum})$

$$\frac{\partial f}{\partial \text{sum}} = 1 - \tanh^2 h = \begin{bmatrix} 0 \\ 0.75 \end{bmatrix}$$

$$\frac{\partial f}{\partial \text{Fraction}} = \frac{\partial f}{\partial \text{sum}} \times \frac{\partial \text{sum}}{\partial \text{fraction}} = 1 - \tanh^2 x = \alpha$$

$$\frac{\partial f}{\partial P_1} = \frac{\partial f}{\partial \max P_1} \times \frac{\partial \max P_1}{\partial P_1} = \alpha \begin{cases} 0 & P_1 \le 0 \\ 1 & P_1 > 0 \end{cases}$$

$$\frac{\partial f}{\partial \max P_2} = \frac{\partial f}{\partial \text{sum}} \times \frac{\partial \text{sum}}{\partial \max P_2} = 1 - \tanh^2(x) = \beta$$

$$\frac{\partial f}{\partial P_2} = \frac{\partial f}{\partial \max P_2} \times \frac{\partial \max P_2}{\partial P_2} = \beta \begin{cases} 0 & P_2 \le 0 \\ 1 & P_2 \ge 0 \end{cases}$$

$$\frac{\partial f}{\partial \max P_1} = \frac{\partial f}{\partial \text{fraction}} \times \frac{\partial \text{fraction}}{\partial \max P_1} = (1 - \tanh^2 x) \frac{1}{\max P_1} = \alpha$$

$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial P_1} \times \frac{\partial P_1}{\partial x_1} = \frac{\partial f}{\partial P_1} \times w_1 \qquad \frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial P_1} \times \frac{\partial P_1}{\partial w_1} = \frac{\partial f}{\partial P_1} \times x_1$$

$$\frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial P_2} \times \frac{\partial P_2}{\partial w_2} = \frac{\partial f}{\partial P_2} \times x_2 \qquad \frac{\partial f}{\partial x_2} = \frac{\partial f}{\partial P_2} \times \frac{\partial P_2}{\partial x_2} = \frac{\partial f}{\partial P_2} \times w_2$$

$$\frac{\partial f}{\partial \max P_1} = \begin{bmatrix} 0 \\ 0.75 \end{bmatrix} \qquad \frac{\partial f}{\partial \max P_2} = \begin{bmatrix} 0 \\ 0.75 \end{bmatrix}$$

$$\frac{\partial f}{\partial P_1} = \begin{bmatrix} 0.075 \\ 0 \end{bmatrix} \qquad \frac{\partial f}{\partial P_2} = \begin{bmatrix} 0.75 \\ 0 \end{bmatrix}$$

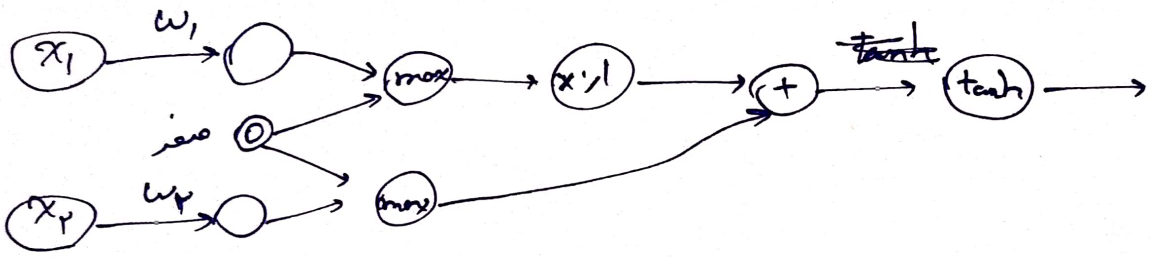
$$\frac{\partial f}{\partial x_1} = \begin{bmatrix} -3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0.075 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.22 \\ 0.075 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_1} = \begin{bmatrix} 0 \\ 0.75 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0.075 & 0.15 \end{bmatrix}$$

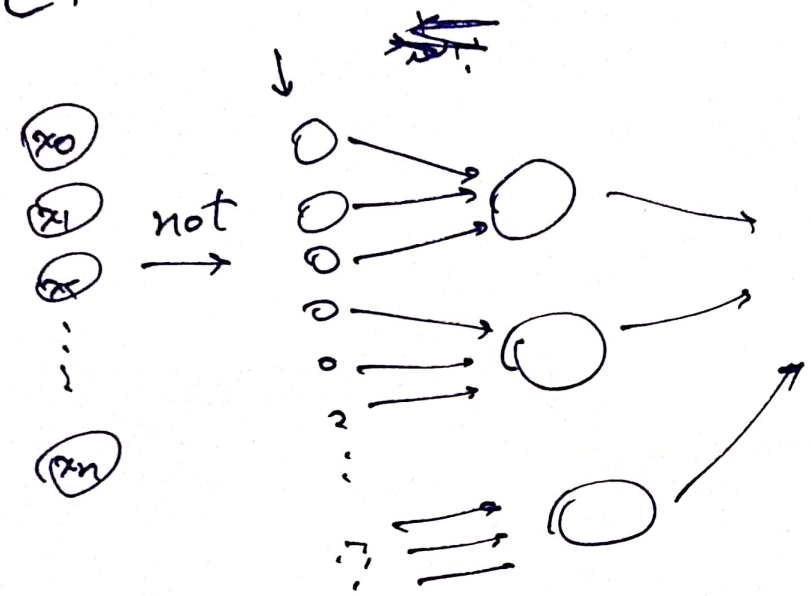
$$\frac{\partial f}{\partial x_2} = \begin{bmatrix} -2 & 2 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 0.75 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.15 \\ 0 \end{bmatrix}$$

$$\frac{\partial f}{\partial w_2} = \begin{bmatrix} 0 \\ 0.75 \end{bmatrix} \begin{bmatrix} -2 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -0.15 & 2.25 \end{bmatrix}$$

٤) ب) سؤال



ب) با یک لایه Not ها را ایجاد می‌کنیم با لایه بعد ورودن‌ها ضرب می‌کنیم در لایه بعد هم جمع را انجام می‌دهیم.

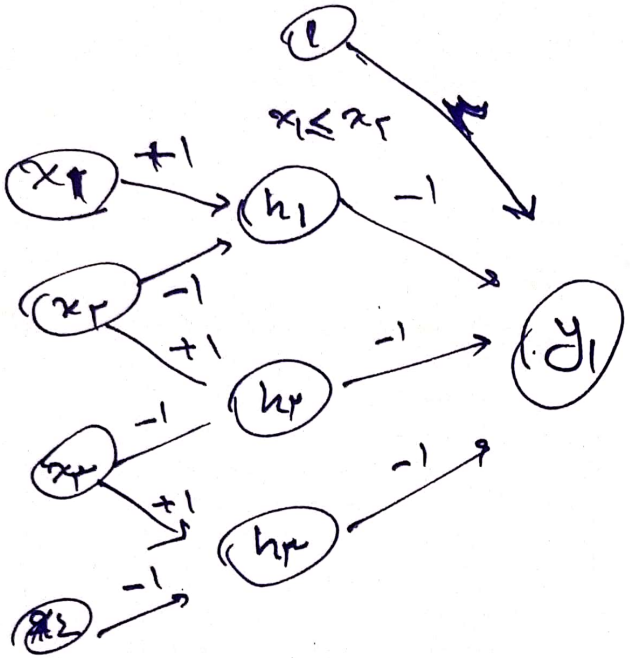


لایه اول Not ها را داریم

در لایه دوم ضرب‌ها انجام می‌شود

و در لایه سوم می‌توان جمع‌ها را انجام داد.

البته می‌توان خیلی گسترده‌تر فکر کرد و هر ضرب را به تعداد المان‌های آن گسترش داد مثلا $x_1x_2x_3$ را در ۳ لایه مختلف ضرب کرد و جمع‌ها را نیز در لایه‌های مختلف انجام داد.


1
$x_1 \le x_2$
x_1
+1
h_1
−1
x_2
−1
+1
y_1
−1
h_2
x_3
−1
+1
h_3
−1
x_4
−1

(۵) (آ) وقتی ضریب λ یا رگولاریزاسیون کم شود، دچار overfitting زیادی شود. اگر وجود دارد کم می‌شود که نتیجه‌ی می‌شود که خطای داده‌های آموزش کم شود یعنی در آموزش بهتر شود و خطای داده‌های تست زیادی شود.

(۵) (ب) زیرا هر هدف مسئله منظم ساز آن فرق می‌کند و ما منظم ساز بهینه اصلا نداریم حتی هم یک منظم ساز ممکن در یک مسئله خوب باشد و در یک مسئله دیگر بد باشد و هر بدترین حالت نیز حتی اگر از منظم ساز بد استفاده کنیم λ آن را ۰ می‌گذاریم.

(۵) (ج)

نرم ۱ $= L_1 = \sum x_i = 7$ $L_1(x_1) = 7$ $L_1(x_2) = 5$

$L_2^{(x_1)} = \sqrt{4+4+9} = \sqrt{17} = 4.1$ $L_2(x_2) = \sqrt{16+9} = 5$ $L_3(x_3) = \sqrt[3]{25} = 5$

L_1 این طور انجام می‌دهد زیرا نرم ۱ برای بردار ۳ کمتر است و در واقع از آن بردار بیشتر استفاده می‌شود و تشخیص رگولاریزاتور L_1 این است که x_2 بیشتر از x_2 و x_1 حضور داشته باشد.

نرم ۲ برای بردار ۱ کمتر است و در واقع منظم ساز بیشتر از بردار x_1 استفاده می‌کند.

به صورت کلی هم L_1 تمایل دارد تعداد صفرها را بیشتر می‌کند.

$\varepsilon \sim N(0, \beta^{-1})$ $\quad \beta^{-1} = \sigma^2$

$P(w;\alpha) = N(0, \alpha^{-1} I)$

Posterior $\propto$ likelihood $\times$ Prior

(ت) (ه)

$w \sim N(0, \alpha^{-1} I)$ $\quad P(y_k|w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_k - x^T w)^2\right)$

$L(w) = \prod_{k=1}^{N} P(y_k|w) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_k - x^T w)^2\right)$ *

رابطه بیزین ما به این شکل است

$P(w|\text{Dataset}) = \frac{P(\text{Dataset}|w)\, P(w)}{P(D)} \xrightarrow{\log} \log P(w|D) = \underbrace{\log P(D|w) + \log P(w)}_{\text{مشتق}} - \underbrace{\log P(D)}_{\text{ثابت}}$

حال باید لگ بگیریم پس مشتق می‌گیریم.

$w^* = \text{argmax}_w \log P(D|w) + \log P(w)$

$\underbrace{\log P(D|w)}_{\text{توزیع مشخص}} + \underbrace{\log P(w)}_{\text{توزیع اولیه}}$

$w \sim N(0, \alpha^{-1} I) \Rightarrow P(w) = \frac{\alpha^{D/2}}{(2\pi)^{D/2}} \exp\left(-\frac{\alpha}{2} w^T w\right)$

$\log P(w) = \log \alpha^{D/2} - \log (2\pi)^{D/2} - \frac{\alpha}{2} w^T w$

۲۱) الف) حل باید لایکلی هود را max کنیم که در w بدست اورد

$\log P(D|w) =$

$$\sum \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_k - x^T w)^2\right) =$$

$$\sum_{k=1}^{D} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(y_k - x^T w)^2$$

در مشتق مهم نیست

پس داریم

$$\hat{w} = \arg\max_w \left(\frac{-1}{2\sigma^2}\sum (y_k - x^T w)^2 - \frac{\alpha}{2} w^T w\right)$$

$$\hat{w} = \arg\min_w = \frac{-1}{2\sigma^2}\sum \left(y_k - x^T w\right)^2 + \frac{\alpha}{2} w^T w)$$

$$= \frac{-1}{2} B \sum (y_k - x^T w)^2 + \frac{\alpha}{2} w^T w$$