

This homework is due **Friday, September 15 at 12pm noon.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, “HW[n] Write-Up”
2. Submit all code needed to reproduce your results, “HW[n] Code”.
3. Submit your test set evaluation results, “HW[n] Test Set”.

After you've submitted your homework, be sure to watch out for the self-grade form.

- (a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None
Comments : HW is long

- (b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

write up Hank

Problem # 2

$$(a) \quad Y = Xw + b + z$$

$$P(Y|X) = P(Xw + b + z | X)$$

$$\text{We know } P(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}$$

$$\Rightarrow P(Y|X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-f(x))^2}{2\sigma^2}} \quad \text{where } f(x) = Xw + b$$

(b) We want to maximize :

$$\begin{aligned} & P(Y_1, Y_2, \dots, Y_n | X_1, X_2, \dots, X_n; w, b) \\ &= P(Y_1 | X_1; w, b) \times P(Y_2 | X_2; w, b) \times \dots \times P(Y_n | X_n; w, b) \end{aligned}$$

Take the log :

$$\begin{aligned} & \max \log P(Y_1, Y_2, \dots, Y_n | X_1, X_2, \dots, X_n; w, b) \\ &= \max \sum_i^n \log P(Y_i | X_i; w, b) \\ &= \min \sum_i^n \frac{(y_i - f(x_i))^2}{2\sigma^2} - n \log(\sqrt{2\pi}\sigma) \end{aligned}$$

The maximum likelihood estimator (MLE)

$$\arg \min_{w, b} \sum_i^n (y_i - f(x_i))^2$$

$$(c) \quad Z \sim U[-0.5, 0.5]$$

$$\Rightarrow P(Z) = \begin{cases} 1 & -0.5 \leq Z \leq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$P(Y|X) = P(Z = Y - Xw | X) = \begin{cases} 1 & -0.5 \leq Y - Xw \leq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow P(Y|X) = P(-0.5 + Xw \leq Y \leq 0.5 + Xw)$$

(d) we want to maximize :

$$\begin{aligned} & P(Y_1, Y_2, \dots, Y_n | X_1, X_2, \dots, X_n; w) \\ &= \prod_i^n P(Y_i | X_i; w) \\ &= \prod_i^n P(-0.5 \leq Y_i - X_i w \leq 0.5) \end{aligned}$$

The maximum likelihood estimator :

$$\underset{w}{\operatorname{argmax}} \prod_i^n P(-0.5 \leq Y_i - X_i w \leq 0.5)$$

$$\text{we know that } P(-0.5 \leq Y_i - X_i w \leq 0.5) = \begin{cases} 1 & -0.5 \leq Y_i - X_i w \leq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \max \prod_i^n P(-0.5 \leq Y_i - X_i w \leq 0.5) = 1$$

$$\text{iff } -0.5 \leq Y_i - X_i w \leq 0.5 \quad \forall X_i, Y_i$$

$$\frac{Y_i - 0.5}{X_i} \leq w \leq \frac{Y_i + 0.5}{X_i} \quad \forall X_i, Y_i$$

$$\text{assuming that } X_i > 0 \quad \forall i$$

more conservative :

$$\frac{\max(Y) - 0.5}{\min(X)} \leq w \leq \frac{\min(Y) + 0.5}{\max(X)}$$

e.) See appendix

(f) The Bayes' theorem gives:

$$P(w|y, x) = \frac{P(y|w, x) P(w|x)}{\int_w P(y|w, x) P(w|x) dw}$$

$$= \frac{P(y|w, x) P(w)}{\int_w P(y|w, x) P(w) dw}$$

$$w \sim N(0, \sigma^2) \Rightarrow P(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{w^2}{2\sigma^2}}$$

$$z_i \sim N(0, 1) \Rightarrow P(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$y_i = x_i w + z_i \Rightarrow y_i \sim N(x_i w, 1)$$

$$\Rightarrow P(y|w, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-xw)^2}{2}}$$

data

$$\Rightarrow P(w|y, x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-xw)^2}{2}} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{w^2}{2\sigma^2}}}{\int_w \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-xw)^2}{2}} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{w^2}{2\sigma^2}} dw} \quad \text{prior}$$

The denominator is constant. $C =$

$$\Rightarrow P(w|y, x) = \frac{1}{2\pi\sigma^2 C} e^{-\frac{(y-xw)^2}{2} - \frac{w^2}{2\sigma^2}}$$

We consider the exponent:

$$\begin{aligned} & -\frac{(y-xw)^2}{2} - \frac{w^2}{2\sigma^2} = -\frac{6^2 y^2 - 26^2 x y w + 6^2 x^2 w^2 + w^2}{2\sigma^2} \\ & = -\frac{w^2 (1 + 6^2 x^2) - 2w \sqrt{1+6^2 x^2} \frac{6^2 x y}{\sqrt{1+6^2 x^2}} + \frac{6^4 x^2 y^2}{1+6^2 x^2} + 6^2 y^2 - \frac{6^4 x^2 y^2}{1+6^2 x^2}}{2\sigma^2} \end{aligned}$$

$$= - \frac{(w\sqrt{1+6^2x^2} - \frac{6^2xy}{\sqrt{1+6^2x^2}})^2 + \text{const}}{26^2}$$

$$\Rightarrow E[w|y, x] = \frac{6^2xy}{1+6^2x^2} \quad \text{where } xy = \frac{1}{n} \sum x_i y_i \\ x^2 = \frac{1}{n} \sum x_i^2$$

(g) $Z_i \sim N(0, 1)$

$$Y_i = w^T X_i + Z_i \Rightarrow Y_i \sim N(w^T X_i, 1)$$

$$\Rightarrow P(Y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - w^T x_i)^2}{2}}$$

MLE is to maximize :

$$\begin{aligned} & P(Y_1, Y_2, \dots, Y_n | X_1, X_2, \dots, X_n; w) \\ &= P(Y_1 | X_1; w) P(Y_2 | X_2; w) \dots P(Y_n | X_n; w) \end{aligned}$$

take log :

$$\begin{aligned} \underset{w}{\operatorname{argmax}} \log P &= \underset{w}{\operatorname{argmax}} \sum_{i=1}^n \log P(Y_i | X_i; w) \\ &= \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \frac{(y_i - w^T x_i)^2}{2} + n \log \sqrt{2\pi} \\ &= \underset{w}{\operatorname{argmin}} \sum_{i=1}^n (y_i - w^T x_i)^2 \\ &= \underset{w}{\operatorname{argmin}} \|y - w^T X\|_2^2 \end{aligned}$$

↑

Least Square Regression

(h) Similar to problem (f) :

$$P(w|y, x) = \frac{P(y|w, x) P(w|x)}{\int_w P(y|w, x) P(w|x) dw}$$

$$= \frac{P(y|w, x) P(w)}{\int_w P(y|w, x) P(w) dw}$$

$$w_i \sim N(0, \sigma^2) \Rightarrow P(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|w\|_2^2}{2\sigma^2}}$$

$$z_i \sim N(0, 1) \Rightarrow P(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|z\|_2^2}{2}}$$

$$y_i = w^T x_i + z_i \Rightarrow y_i \sim N(w^T x_i, 1)$$

$$\Rightarrow P(y|w, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|y - xw\|_2^2}{2}}$$

$$\Rightarrow P(w|y, x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{\|y - xw\|_2^2}{2}}}{\int_w \frac{1}{\sqrt{2\pi}} e^{-\frac{(y - xw)^2}{2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|w\|_2^2}{2\sigma^2}} dw}$$

↑
constant

To calculate the mean, similarly we have:

$$E[w|y, x] = (I + \sigma^2 x^T x)^{-1} \sigma^2 x^T y$$

(i) See appendix

Problem #3

$$(a) 1. \hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\Rightarrow E[\hat{X}] = \frac{E[x_1] + E[x_2] + \dots + E[x_n]}{n} = \frac{n\mu}{n} = \mu$$

$$\Rightarrow \text{bias } E[\hat{X} - \mu] = E[\hat{X}] - \mu = \mu - \mu = 0$$

$$2. \hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n+1}$$

$$\Rightarrow E[\hat{X}] = \frac{n\mu}{n+1}$$

$$\Rightarrow \text{bias } E[\hat{X} - \mu] = E[\hat{X}] - \mu = \frac{n\mu}{n+1} - \mu = -\frac{\mu}{n+1}$$

$$3. \hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n+n_0}$$

$$\Rightarrow E[\hat{X}] = \frac{n\mu}{n+n_0}$$

$$\Rightarrow \text{bias } E[\hat{X} - \mu] = E[\hat{X}] - \mu = \frac{n\mu}{n+n_0} - \mu = -\frac{n_0\mu}{n+n_0}$$

$$4. \hat{X} = 0$$

$$\Rightarrow E[\hat{X}] = 0$$

$$\Rightarrow \text{bias } E[\hat{X} - \mu] = E[\hat{X}] - \mu = 0 - \mu = -\mu$$

(b)

$$1. \hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\Rightarrow \text{Var}[\hat{X}] = \frac{1}{n^2} \text{Var}[x_1 + x_2 + \dots + x_n] =$$

$$= \frac{\text{Var}(x_1) + \text{Var}(x_2) + \dots + \text{Var}(x_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

2.

$$\hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n+1}$$

$$\Rightarrow \text{Var}[\hat{X}] = \frac{1}{(n+1)^2} \text{Var}[x_1 + x_2 + \dots + x_n] = \frac{n\sigma^2}{(n+1)^2}$$

3.

$$\hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n+n_0}$$

$$\Rightarrow \text{Var}[\hat{X}] = \frac{1}{(n+n_0)^2} \text{Var}[x_1 + x_2 + \dots + x_n] = \frac{n\sigma^2}{(n+n_0)^2}$$

4.

$$\hat{X} = 0$$

$$\Rightarrow \text{Var}[\hat{X}] = 0$$

(c)

$$1. E[\hat{X} - \mu] = 0 \quad \text{Var}[\hat{X}] = \frac{\sigma^2}{n}$$

$$\Rightarrow \text{ETE} = 0^2 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n}$$

$$2. E[\hat{X} - \mu] = -\frac{\mu}{n+1} \quad \text{Var}[\hat{X}] = \frac{n\sigma^2}{(n+1)^2}$$

$$\Rightarrow \text{ETE} = \frac{\mu^2}{(n+1)^2} + \frac{n\sigma^2}{(n+1)^2} = \frac{\mu^2 + n\sigma^2}{(n+1)^2}$$

3.

$$E[\hat{X} - \mu] = -\frac{n_0\mu}{n+n_0} \quad \text{Var}[\hat{X}] = \frac{n\sigma^2}{(n+n_0)^2}$$

$$\Rightarrow \text{ETE} = \frac{n_0^2\mu^2}{(n+n_0)^2} + \frac{n\sigma^2}{(n+n_0)^2} = \frac{n_0^2\mu^2 + n\sigma^2}{(n+n_0)^2}$$

$$4. \quad E[\hat{X} - \mu] = -\mu \quad \text{Var}[\hat{X}] = 0$$

$$\Rightarrow \text{ETE} = \mu^2$$

(d) The 3rd estimator is the general case:

$$\hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n + n_0}$$

$$\text{with bias } E[\hat{X} - \mu] = -\frac{n_0 \mu}{n + n_0} \quad \text{Var}[\hat{X}] = \frac{n \sigma^2}{(n + n_0)^2}$$

$$n_0 = 0 : \quad \hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$E[\hat{X} - \mu] = 0 \quad \text{Var}[\hat{X}] = \frac{\sigma^2}{n}$$

→ case 1

$$n_0 \rightarrow \infty : \quad \hat{X} = \lim_{n_0 \rightarrow \infty} \frac{x_1 + x_2 + \dots + x_n}{n + n_0} = 0$$

$$E[\hat{X} - \mu] = \lim_{n_0 \rightarrow \infty} -\frac{n_0 \mu}{n + n_0} = \lim_{n_0 \rightarrow \infty} -\frac{\mu}{\frac{n}{n_0} + 1} = -\mu$$

$$\text{Var}[\hat{X}] = \lim_{n_0 \rightarrow \infty} \frac{n \sigma^2}{(n + n_0)^2} = 0$$

→ case 4

$$n_0 = 1 : \quad \hat{X} = \frac{x_1 + x_2 + \dots + x_n}{n + 1}$$

$$E[\hat{X} - \mu] = -\frac{1 \times \mu}{n + n_0} \quad \text{Var}[\hat{X}] = \frac{n \sigma^2}{(n + 1)^2}$$

→ case 2

$$(e) n_0 = \alpha n$$

$$ETE = \frac{n_0^2 \mu^2 + n \delta^2}{(n+n_0)^2} = \frac{\alpha^2 n^2 \mu^2 + n \delta^2}{(n+\alpha n)^2}$$

$$\frac{d ETE}{d \alpha} = \frac{2 \alpha n^2 \mu^2 (n+\alpha n)^2 - (2n^2 + 2\alpha n^2)(\alpha^2 n^2 \mu^2 + n \delta^2)}{(n+\alpha n)^4} = 0$$

$$\Rightarrow 2 \alpha n \mu^2 (1+\alpha)^2 - (1+\alpha)(\alpha^2 n \mu^2 + \delta^2) = 0$$

$$\alpha n \mu^2 (1+\alpha) - \alpha^2 n \mu^2 - \delta^2 = 0$$

$$\cancel{\alpha^2 n \mu^2} + \alpha n \mu^2 - \cancel{\alpha^2 n \mu^2} - \delta^2 = 0$$

$$\Rightarrow \alpha = \frac{\delta^2}{n \mu^2}$$

(f) μ small & δ large

$\Rightarrow \alpha \nearrow^\infty$ to minimize ETE

$$(g) \mu = \mu_0 \Rightarrow \alpha = \frac{\delta^2}{n \mu_0^2} \Rightarrow n_0 = \alpha n = \frac{\delta^2}{\mu_0^2}$$

We have two approaches:

1. Estimate the mean of the sampling distribution of mean

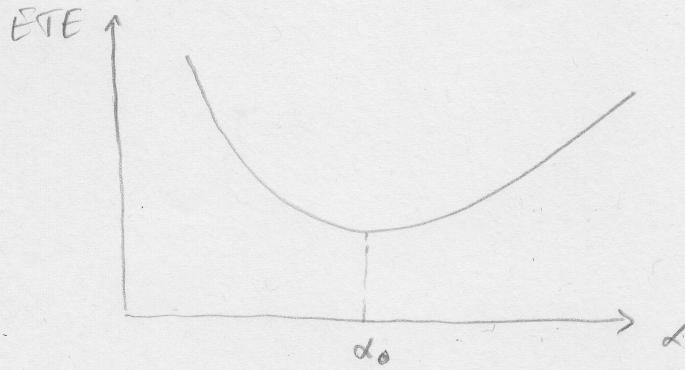
$$\frac{x_1 + x_2 + \dots + x_n}{n+n_0} \quad \text{where } n_0 = \frac{\delta^2}{\mu_0^2}$$

by doing this, the mean has minimum value.

2. Refine x before estimating mean, that is

$$x' = x - \mu_0$$

(h) From question (e), we have the diagram of ETE vs α

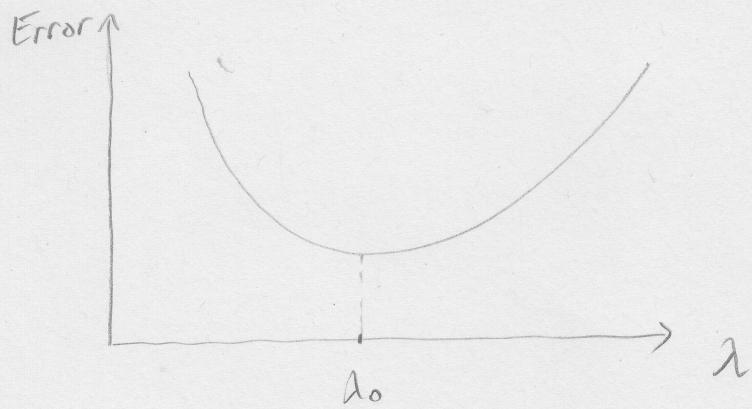


That is there is a value α_0 minimizing the expected total error ETE

$\alpha < \alpha_0$: under training, # of samples is not enough

$\alpha > \alpha_0$: over training, # of sample is too large

From lecture : Error v.s. λ in Ridge regression



$\lambda < \lambda_0$: over-fitting

$\lambda > \lambda_0$: under fitting, the regularization term dominates

$\Rightarrow \alpha$ and λ share the same meaning in opposite direction.

Problem #4

$$\begin{aligned}
 (a) \quad \hat{x} &= \arg \min_x \|y - Ax\|_2^2 = \arg \min_x \|Ax^* + w - Ax\|_2^2 \\
 &= \arg \min_x \|w - A(x - x^*)\|_2^2 \\
 \Rightarrow \hat{x} - x^* &= (A^T A)^{-1} A^T w \Rightarrow \hat{x} = (A^T A)^{-1} A^T w + x^* \\
 \|A\hat{x} - y^*\|_2^2 &= \|A(A^T A)^{-1} A^T w + Ax^* - \underbrace{y^*}_{y^*}\|_2^2 \\
 &= \|A(A^T A)^{-1} A^T w\|_2^2
 \end{aligned}$$

(b) From (a) :

$$\begin{aligned}
 \|A\hat{x} - y^*\|_2^2 &= \|A(A^T A)^{-1} A^T w\|_2^2 \\
 A &= U\Sigma V^T \nearrow \\
 \|A(A^T A)^{-1} A^T w\|_2^2 &= \|U\Sigma V^T (\sqrt{\Sigma^T U^T U \Sigma V^T})^{-1} \sqrt{\Sigma^T U^T w}\|_2^2 \\
 &= \|U \underbrace{\Sigma V^T (\Sigma V^T)^{-1}}_I \underbrace{(U^T U)^{-1}}_I \underbrace{(\sqrt{\Sigma^T})^{-1} \sqrt{\Sigma^T U^T w}}_I\|_2^2
 \end{aligned}$$

$$U^T U = I \text{ because } U_{full} U_{full}^T = U_{full}^T U = I, \quad U_{full} = [u_1 \ u_2 \ \dots \ u_n]$$

U is U_{full} reduced by omitting some column

$$\Rightarrow U^T U = I \text{ but } UU^T \neq I$$

$$\begin{aligned}
 \Rightarrow \|A\hat{x} - y^*\|_2^2 &= \|UU^T w\|_2^2 \\
 &= (UU^T w)^T (UU^T w) \\
 &= \underbrace{w^T U U^T U U^T w}_I = w^T U U^T w \\
 &= (U^T w)^T U^T w = \|U^T w\|_2^2
 \end{aligned}$$

$$(c) \quad w_i \sim N(0, \sigma^2) \Rightarrow U^T w = \begin{bmatrix} -u_1^T \\ -u_2^T \\ \vdots \\ -u_n^T \end{bmatrix} w \sim N\left(0, \begin{bmatrix} \|u_1\|_2^2 \sigma^2 \\ \|u_2\|_2^2 \sigma^2 \\ \vdots \\ \|u_n\|_2^2 \sigma^2 \end{bmatrix}\right)$$

but U is a reduced orthonormal matrix $\Rightarrow \|u_i\|_2^2 = 1 \forall i$

$$\Rightarrow U^T w \sim \left(0, \begin{bmatrix} \sigma^2 \\ \sigma^2 \\ \vdots \\ \sigma^2 \end{bmatrix}\right) \Rightarrow E[\|U^T w\|_2^2] = \mu(U^T w)^0 = \underbrace{\sigma^2 + \sigma^2 + \dots + \sigma^2}_{d \text{ rows}} = d\sigma^2$$

$$\frac{1}{n} E[\|A\hat{x} - y^*\|_2^2] = \frac{1}{n} E[\|U^T w\|_2^2] = \frac{1}{n} d\sigma^2$$

d) The problem set up is:

$$\min \|Ax - y\|_2^2$$

where $A \in \mathbb{R}^{n \times (D+1)}$ is Vandermonde matrix

The average squared error:

$$\frac{1}{n} E[\|A\hat{x} - y^*\|_2^2] = \sigma^2 \frac{D+1}{n} \leq \epsilon$$

$$\Rightarrow n \geq \frac{\sigma^2(D+1)}{\epsilon}$$

Conclusion: $D \uparrow \rightarrow n \uparrow$ proportionally w/ slope $\frac{\sigma^2}{\epsilon}$

e.) See appendix

Problem # 5

See appendix