HW3-1

a)

Names	Email Address
Qi Wan	Jhun0324@berkeley.edu

Description of Team: Best Group Ever

How did I work?

Comments:

b) 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.

Qingyang Zhao

HW2-2

a) $\int_{-\infty}^{y} P_{Y|X}(t|x)\,dt = \mathbb{P}(Y \le y \mid X = x)$ X, Y denote R.V. x, y: the value in Alphabet Set

$= \mathbb{P}(Xw + b + Z \le y \mid X = x)$

$= \mathbb{P}(Z \le y - xw - b) = \int_{-\infty}^{y-wx-b} p(z)\,dz$ $\quad Z \sim \mathcal{N}(0,1)$

$P_{Y|X}(y|x) = \frac{\int_{-\infty}^{y} P_{Y|X}(t|x)\,dt}{dy} = \frac{\int_{-\infty}^{y-wx-b} p_Z(z)\,dz}{dy} = p_Z(y-wx-b) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-wx-b)^2}{2}}$ $\quad Y|X \sim (wx+b, 1)$

b) $\underset{w,b}{\text{maximize}} \prod_{i=1}^{n} P_{Y_i|X_i}(y_i|x_i) \iff \underset{w,b}{\text{maximize}} \log \prod_{i=1}^{n} P_{Y_i|X_i}(y_i|x_i)$

Set $f(w,b) = \sum_{i=1}^{n} \log P_{Y_i|X_i}(y_i|x_i)$

$f(w,b) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-wx-b)^2}{2}} = \left(\sum_{i=1}^{n} \frac{-(y_i - wx_i - b)^2}{2} \right) - n \log \sqrt{2\pi}$

Thus, $\underset{w,b}{\text{maximize}}\ f(w,b)$ is equivalent to $\underset{w,b}{\text{minimize}} \sum_{i=1}^{n} (y_i - wx_i - b)^2$

$$\underset{w,b}{\text{minimize}} \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ w \end{bmatrix} \right\|_2 \quad \text{let } \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \vec{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{bmatrix} \quad \vec{w} = \begin{bmatrix} b \\ w \end{bmatrix}$$

$\text{minimize } \|\vec{y} - \vec{X}\vec{w}\|_2 \qquad \vec{w} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y} = \begin{bmatrix} b \\ w \end{bmatrix}$

c) $Z \sim U[-0.5, 0.5] \qquad p_Z(z) = \begin{cases} 1, & z \in [-0.5, 0.5] \\ 0, & \text{otherwise} \end{cases}$

$\int_{-\infty}^{y} P_{Y|X}(t|x)\,dt = \mathbb{P}(Y \le y \mid X = x) = \mathbb{P}(wX + Z \le y \mid X = x)$

$= \int_{-\infty}^{y-wx} p_Z(z)\,dz = \mathbb{P}(Z \le y - wx)$ $\quad P_{Y|X}(y|x) = \frac{\int_{-\infty}^{y} P_{Y|X}(t|x)\,dt}{dy} = \frac{\int_{-\infty}^{y-wx} p_Z(z)\,dz}{dy}$

$= p_Z(y - wx)$

$$P_{Y|X}(y|x) = P_Z(y-wx) = \begin{cases} 1, & y \in [-0.5+wx,\ 0.5+wx] \\ 0, & \text{otherwise} \end{cases}$$

$$Y|X_w \sim U[-0.5+wx,\ 0.5+wx]$$

(d) criterion: find a "w", such that the number of points lies in $[-0.5+wx_i,\ 0.5+wx_i]$ is maximized.

$$-0.5+wx_i \le y_i \le 0.5+wx_i \qquad y_i-0.5 \le wx_i \le y_i+0.5$$

when $x_i<0$: $\dfrac{y_i+0.5}{x_i} \le w \le \dfrac{y_i-0.5}{x_i}$ | $x_i<0$

$$y_i-0.5 \le wx_i \le y_i+0.5$$

$x_i<0$: $\dfrac{y_i+0.5}{x_i} \le w \le \dfrac{y_i-0.5}{x_i}$ $\qquad \dfrac{y_i}{x_i} - \dfrac{0.5}{|x_i|} \le w \le \dfrac{y_i}{x_i} + \dfrac{0.5}{|x_i|}$

$x_i>0$: $\dfrac{y_i+0.5}{x_i} \le w \le \dfrac{y_i+0.5}{x_i}$

Thus $w \in \left[\max_{1\le i\le n} \dfrac{y_i}{x_i} - \dfrac{0.5}{|x_i|},\ \min_{1\le i\le n} \dfrac{y_i}{x_i} + \dfrac{0.5}{|x_i|}\right]$

e)

As n gets large, the area of w,b getting smaller

f) $P_{w|\{x_i,y_i\}}(w|\{x_i=x_i, y_i=y_i\}) = \dfrac{P(\{x_i,y_i\},w)}{\int_{-\infty}^{+\infty} P(\{x_i,y_i\}|w)P(w)\,dw}$

$\{x_i, y_i\}$ denote $\{x_1,y_1\cdots x_n,y_n\}$

P_* denote PDF of R.V. "*"

$$P_{\{x_i,y_i\},w}(\{x_i,y_i\},w) = \prod_{i=1}^{n}\left(P_{y_i|x_i,w}\right)P_{x_1,x_2\cdots x_n,w} = \prod_{i=1}^{n}\left(P_{y_i|x_i,w}\right)\left(\prod_{i=1}^{n}P_{x_i}\right)(P_w) \quad (1)$$

(noise are i.i.d.)

$$\int_{-\infty}^{+\infty} P_{\{x_i,y_i\},w}(\{x_i,y_i\},w)\,dw = \prod_{i=1}^{n}P_{x_i}\int_{-\infty}^{+\infty}\prod_{i=1}^{n}P_{y_i|x_i,w}\,P_w\,dw \quad (2)$$

$$P_{w|\{x_i,y_i\}} = \frac{(1)}{(2)} = \frac{\prod_{i=1}^{n}(P_{y_i|x_i,w})P_w}{\int_{-\infty}^{+\infty}\prod_{i=1}^{n}P_{y_i|x_i,w}P_w\,dw} \longleftarrow \text{some constant} \qquad \longrightarrow \text{Gaussian !!!}$$

Thus, [next Page]

HW3-2

f) Thus, $p_{w|\{x_i,y_i\}} = \frac{1}{\sigma_0} e^{-\frac{(w-\mu_0)^2}{2\sigma_0^2}}$

All we need to do is to calculate $\{\mu_0, \sigma_0\}$

which is inside $\prod_{i=1}^{n} (p_{y_i|x_i,w})\, p_w$ (a)

$= A_0 \prod_{i=1}^{n} e^{-\frac{(y_i - x_i w)^2}{2}} = A_0 e^{-\sum_{i=1}^{n} \frac{(y_i - x_i w)^2}{2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{w^2}{2\sigma^2}}$

Some constants A_0 I don't care

Goal:

$$-\frac{(w-\mu_0)^2}{2\sigma_0^2}$$

Given:

$$-\sum_{i=1}^{n} \frac{(y_i - x_i w)^2}{2} - \frac{w^2}{2\sigma^2}$$

$$-\frac{\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} x_i^2 w^2 - 2\sum_{i=1}^{n} x_i y_i w}{2} - \frac{w^2}{2\sigma^2}$$

$$-\frac{(\sum_{i=1}^{n} x_i^2 + \frac{1}{\sigma^2}) w^2 - 2\sum_{i=1}^{n} x_i y_i w}{2} + \text{const}$$

$$\begin{cases} \mu_0 = -\sum_{i=1}^{n} x_i y_i \Big/ \left(\sum_{i=1}^{n} x_i^2 + \frac{1}{\sigma^2}\right) \\ \sigma_0^2 = \frac{1}{\sum_{i=1}^{n} x_i^2 + \frac{1}{\sigma^2}} \end{cases}$$

$$w|\{x_i,y_i\} \sim N\left(-\frac{\sum_{i=1}^{n} x_i y_i}{2} \Big/ \left(\sum_{i=1}^{n} x_i^2 + \frac{1}{\sigma^2}\right), \frac{1}{\sum_{i=1}^{n} x_i^2 + \frac{1}{\sigma^2}}\right)$$

$$E[p_{w|\{x_i,y_i\}}] = -\frac{\sum_{i=1}^{n} x_i y_i}{2}$$

g) $\underset{w}{\text{maximize}} \prod_{i=1}^{n} p_{y_i|x_i}(y_i|x_i) \iff \underset{w}{\text{maximize}} \log \prod_{i=1}^{n} p_{y_i|x_i}(y_i|x_i)$

$$\underset{w}{\text{maximize}} \sum_{i=1}^{n} \log p_{y_i|x_i}(y_i|x_i) \iff \underset{w}{\text{minimize}} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

$$\iff \underset{w}{\text{minimize}} \| y - w^T X \|_2^2$$

$$w = X(X^T X)^{-1} Y$$

h) $P_{w|\{x_i, y_i\}}$ is a multivariable Gaussian

It is just the expansion of f)

$$w \sim N(\mu, \Sigma)$$

Proof: consider $\prod_{i=1}^{n} P_{y|x_i,w} P_w$ $\quad \begin{bmatrix} \frac{1}{\sigma^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_2} \end{bmatrix}$

$$= C_0 e^{\frac{-\sum^{n}(y_i - w^T x_i)^2}{2}} \cdot e^{-\frac{1}{2} w^T \Sigma_w^{-1} w} \qquad -\frac{1}{2}(w-a_0)^T A (w-a_0)$$

$$-\frac{1}{2}(w^T A w - 2a_0^T A w)$$

$$= \frac{\sum_{i=1}^{n}(w^T x_i)^2 - 2\sum_{i=1}^{n} w^T x_i y_i}{2} - \frac{1}{2} w^T \Sigma_w^{-1} w + C_1$$

$$= -\frac{1}{2} \cdot w^T \left(\sum_{i=1}^{n} x_i x_i^T + \Sigma_w^{-1}\right) w + \sum_{i=1}^{n} w^T x_i y_i + C_1$$

$$\vec{\mu} = \left(\sum_{i=1}^{n} x_i x_i^T + \Sigma_w^{-1}\right)^{-1} \sum_{i=1}^{n} x_i y_i \qquad \Sigma = \left(\sum_{i=1}^{n} x_i x_i^T + \Sigma_w^{-1}\right)^{-1}$$

i) As σ^2 becomes large, choice of w becomes large, because more w can be chosen with the same probability,

As training sample becomes large, the estimation becomes better,

But $n \geq 10000$ ~~the true model~~ will have a better estimate.

#10.3-3

a) $E\left[\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right] = E[X]-\mu = 0$

$E\left[\frac{X_1+X_2+\cdots+X_n}{n+1}-\mu\right] = \frac{n}{n+1}E[X]-\mu = -\frac{1}{n+1}\mu$

$E\left[\frac{X_1+X_2+\cdots+X_n}{n+n_0}-\mu\right] = \frac{n}{n+n_0}E[X]-\mu = -\frac{n_0}{n+n_0}\mu$

$E[0-\mu] = -\mu$

b) $\mathrm{var}\left[\frac{X_1+X_2+\cdots+X_n}{n}\right] = \frac{1}{n^2}\cdot n\,\mathrm{Var}[X] = \frac{1}{n}\mathrm{var}[X] = \frac{1}{n}\sigma^2$

$\mathrm{Var}\left[\frac{X_1+X_2+\cdots+X_n}{n+1}\right] = \frac{n}{(n+1)^2}\mathrm{var}[X] = \frac{n}{(n+1)^2}\sigma^2$

$\mathrm{var}\left[\frac{X_1+X_2+\cdots+X_n}{n+n_0}\right] = \frac{n}{(n+n_0)^2}\mathrm{var}[X] = \frac{n}{(n+n_0)^2}\sigma^2$

$\mathrm{var}[0] = 0$

c) Expected total Error

$\mathrm{Error}_1 = \mathrm{var}[\hat{X}] + E[\hat{X}-\mu] = \frac{1}{n}\sigma^2$ $\quad \mathrm{Error}_3 = \frac{n}{(n+n_0)^2}\sigma^2 + \frac{n_0^2}{(n+n_0)^2}\mu^2$

$\mathrm{Error}_2 = \frac{n}{(n+1)^2}\sigma^2 + \left(\frac{1}{n+1}\right)^2\mu^2$ $\quad \mathrm{Error}_4 = \mu^2$

d) Simply See the total Error.
Error_1 is when $n_0 = 0$ $\quad \mathrm{Error}_2$ is $n_0 = 1$
Error_4 is $n_0 = n$.

e) $\min_{\alpha}$ $\left(\frac{n\alpha}{n+n\alpha}\right)^2\mu^2 + \frac{n}{(n+n\alpha)^2}\sigma^2$

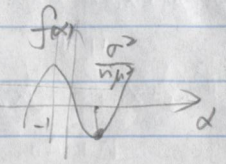
$$f(\alpha) = \left(\frac{\alpha}{1+\alpha}\right)^2\mu^2 + \frac{1}{n}\frac{1}{(1+\alpha)^2}\sigma^2$$

$$\frac{d f(\alpha)}{d\alpha} = \frac{2\mu^2\alpha - \frac{2}{n}\sigma^2}{(1+\alpha)^3}$$

$\alpha \in (-\infty, -1)$ $df/d\alpha > 0$

$(-1, \frac{\sigma^2}{n\mu^2})$ < 0

$(\frac{\sigma^2}{n\mu^2}, +\infty)$ > 0



Thus when $\alpha = \frac{\sigma^2}{n\mu^2}$ $f(\alpha)$ reaches minimum.

f) α becomes rather big. If we know μ should be small, σ large we should put more sample $X=0$ to estimate a better mean, with a low total error.

g) $X' = X - \mu_0$ $\quad E[X'] = E[X] - \mu_0 \to 0$

Or

$X' = \frac{X_1 - \mu_0 + X_2 - \mu_0 + \cdots + X_n - \mu_0}{n+n_0}$ Actually this is the estimator of $X - \mu_0$

h) α can be viewed as a regularization term.

Because when we have prior knowledge about μ to be close to 0, we will set α really big, to drag the estimate back to 0.

Similarly, in Ridge Regression, λ is also use to bound the parameter, if you believe it really small, a large λ should be set to constrain the norm of parameter.

$A\hat{x} - y^*$ $\hat{y} \to A\hat{x}$

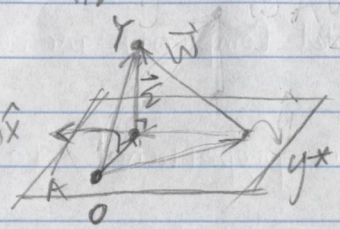
HOW - 4

a)

$$\|A\hat{x} - y^*\|_2^2 = \|A\hat{x} - Y - \vec{w}\|_2^2 = \|\hat{Y} - Y - \vec{w}\|_2^2 = \|\vec{z} - \vec{w}\|_2^2$$

$\to$ Column space of A

$\text{Proj}_A(Y) = \vec{Y} - \vec{z}$ $\quad \vec{z} \perp C(A)$



$\text{Proj}_A(y^* + \vec{w}) = \text{Proj}_A(\vec{w}) + y^* = \vec{Y} - \vec{z}$

$\text{Proj}_A(\vec{w}) = \vec{w} - \vec{z}$ $\quad \|\text{Proj}_A(\vec{w})\|_2 = \|\vec{w} - \vec{z}\|_2$

$\|A \cdot (A^TA)^{-1}A^T\vec{w}\|_2 = \|A\hat{x} - y^*\|_2$.

b) $\|A\hat{x} - y^*\|_2^2 = \|A(A^TA)^{-1}A^Tw\|_2$

$= \|U\Sigma V^T(V\Sigma^2V^T)^{-1}V\Sigma U^Tw\|_2$

$= \|U\Sigma V^TV\Sigma^{-2}V^TV\Sigma U^Tw\|_2$

$= \|UU^Tw\|_2$ $\quad w^TUU^TUU^Tw = \|U^Tw\|_2$

$n\times d$ $\quad d\times n$ $\quad n\times 1$ $\quad U^TU = I_{d\times d}$

$= \|U^Tw\|_2$

c) $\frac{1}{n}E[\|A\hat{x} - y^*\|_2^2] = \frac{1}{n}E[\|U^Tw\|_2^2]$.

$= \frac{1}{n}E\left[\left\|\begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_d^T \end{bmatrix} w\right\|_2^2\right] = \frac{1}{n}E\left[(u_1^Tw)^2 + (u_2^Tw)^2 + \cdots + (u_d^Tw)^2\right]$

$\|u_i\|_2^2 = 1$

$u_i^T\frac{w}{\sigma} \sim N(0,1)$

$= \frac{\sigma^2}{n}E\left[\left(\frac{u_1^Tw}{\sigma}\right)^2 + \left(\frac{u_2^Tw}{\sigma}\right)^2 + \cdots + \left(\frac{u_d^Tw}{\sigma}\right)^2\right]$

$= \frac{\sigma^2 d}{n}$

d) $\sigma^2 \frac{d+1}{n} \leq \varepsilon \qquad n \geq \frac{d+1}{\varepsilon}\sigma^2$

when Model Complexity increase, $d\uparrow$, $n\uparrow$

Thus n is proportional to model complexity

e) As D becomes large average error is bigger.

As n becomes large average error becomes smaller.

HW3-5

a) LinAlgError: Singular Matrix

Because X^TX is 2700×2700, samples are only 91.
X^TX is singular.

b)

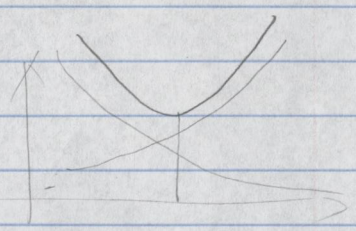
λ =	
0.1	1.2567×10^{-15}
1.0	1.2567×10^{-13}
10	1.2566×10^{-11}
100	1.2556×10^{-9}
1000	1.2460×10^{-7}

c)

λ =	
0.1	3.2557×10^{-7}
1.0	2.9105×10^{-5}
10	1.5904×10^{-3}
100	3.4773×10^{-2}
1000	2.5440×10^{-1}

d) (test data)

λ=0.1	0.86808
1.0	0.86210
10	0.82751
100	0.72465
1000	0.72501



λ goes large. Model Complexity goes down, bias goes up↑, variance goes down↓

e)

without λ	27785230710656 91
λ=100	197781

HW3-6 Own Question

View training error & true error in bias / variance decomposition

training Error:

$E[(h(x;D)-Y)^2]$ $\quad Y = f(x)+N$

$= E[(h(x;D)-\underbrace{E[Y]}_{y^*})^2] + V[h(x;D)] + V[N]$

Bias — Variance — Irreducible Error

True Error:

$E[\|h(x;D)-y^*\|^2]$

$= E[(h(x;D)-y^*)^2] + \|h(x;D)-y^*\|^2$

Variance — Bias

Thus Training Error and True error in difference of Irreducible error

As Model become Complex, models are trained to fit Irreducible error (over-fitting)

That's way training error becomes small but true error increase (variance↑)