CS189

1. (a) By myself. consulting piazza.

(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.

Yao Cao

CS 189 HW2

2. (a) minimize $\|y - Xw\|_2^2$ (1)

subject to $\|w\|_2^2 \le \beta^2$

$$\|w\|_2^2 \le \beta^2 \Rightarrow \|w\|_2^2 - \beta^2 \le 0$$

(1) is equivalent to

$$\boxed{\min_{w} \max_{\lambda \geq 0} \|y - Xw\|_2^2 + \lambda(\|w\|_2^2 - \beta^2) = \min_{w} \max_{\lambda \geq 0} L(w, \lambda)}$$

① the argument is that if $\|w\|_2^2 > \beta^2$, to maximize $L(w, \lambda)$ is to set $\lambda \to \infty$, then the inner maximation will blow up and be ignored by outer minimization;

② if $\|w\|_2^2 \le \beta^2$, and $\lambda \geq 0$ then $\lambda(\|w\|_2^2 - \beta^2) \le 0$

$$\max_{\lambda \geq 0} \lambda(\|w\|_2^2 - \beta^2) = 0$$

$\Rightarrow \max_{\lambda \geq 0} L(w, \lambda) = \|y - Xw\|_2^2$, and the problem becomes

$$\min_{w} \|y - Xw\|_2^2$$

In practice if we choose a certain $\lambda > 0$ then the problem becomes:

$$\min_{w} \left(\|y - Xw\|_2^2 + \lambda(\|w\|_2^2 - \beta^2) \right)$$

$$= \min_{w} \left(\|y - Xw\|_2^2 + \lambda \|w\|_2^2 \right)$$

(b) increasing β is to relax the constraint on w, then it is easier to minimize $\|y - Xw\|_2^2$ closer to its unconstrained minimum, which is equivalent to a $\boxed{\text{smaller } \lambda}$, which means that we decrease our penalty on $\|w\|_2^2$ such that the minimization problem becomes more closer to $\min_w \|y - Xw\|_2^2$ and the minimization is closer to its unconstrained minimum.

(c) prediction without noise $w^T x$

prediction with noise $w^T(x+\varepsilon)$

change $|w^T(x+\varepsilon) - w^T x| = |w^T \varepsilon| \le \|w\|_2 \cdot \|\varepsilon\|_2$

so we want small $\|w\|_2$

(d) $\|y - Xw\|_2^2 + \lambda \|w\|_2^2$

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

$= y^T y + w^T X^T X w - 2y^T X w + \lambda w^T w$

$= y^T y + w^T (X^T X + \lambda I) w - 2y^T X w$

$\nabla_w (y^T y + w^T (X^T X + \lambda I) w - 2y^T X w) = (X^T X + \lambda I + X^T X + \lambda I) w - 2X^T y$

$= 2(X^T X + \lambda I) w - 2X^T y = 0$

$\Rightarrow (X^T X + \lambda I) w = X^T y$

$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$

when $\lambda \to \infty \Rightarrow X^T X + \lambda I \to \infty \cdot I \Rightarrow (X^T X + \lambda I)^{-1} \to 0 \cdot I$

$\Rightarrow \hat{w} \to 0$

(e) X^TX has eigenvalues $\sigma_1^2, \cdots \sigma_d^2$ with eigenvectors $V_1, V_2, \cdots V_d$

$$X^TXV_i = \sigma_i^2 V_i \qquad (i = 1, \cdots, d)$$

$$\Rightarrow (X^TX + \lambda I)V_i = X^TXV_i + \lambda V_i$$

$$= \sigma_i^2 V_i + \lambda V_i$$

$$= (\sigma_i^2 + \lambda) V_i$$

so the eigenvalues for $X^TX + \lambda I$ is $\sigma_i^2 + \lambda$ $(i = 1, \cdots d)$

if one of the eigenvalues of X^TX, say σ_i^2 is close to 0, then $\det|X^TX| \to 0$ and X^TX is close to non-full rank situation, reverting such matrix will be numerically unstable;
if we add $\lambda > 0$ to σ_i^2, $\det|X^TX + \lambda I|$ won't get close to zero, thus $(X^TX + \lambda I)$ is solidly full rank, and will be easily inverted.

(f) the features for this data set seem to be at very different scale, I would expect the w_i to be at very different scale. In this case, I don't want shrink $\|w\|$ too much since that will make w_i more on similar scales and not able to reflect the huge difference in each feature. So I would choose $\lambda_2 = 0.5$

(g) we can decompose any w into $w_0 + X^T\alpha$ for $w_0 \in null(X)$, $\alpha \in \mathbb{R}^n$

① consider the situation without ridge regression

$$X^TXw = X^Ty$$
$$= X^TX(w_0 + X^T\alpha)$$
$$= X^TXw_0 + X^TXX^T\alpha$$
$$= X^TXX^T\alpha$$

$\Rightarrow XX^Ty = XX^TXX^T\alpha$

$\Rightarrow \alpha = (XX^TXX^T)^{-1}XX^Ty$

but there is no way we can solve w_0.

since there are infinite number of w_0, we have infinite number of solutions for w

② for ridge regression

$$w = (X^TX + \lambda I)^{-1}X^Ty$$

since $X^TX + \lambda I$ is always invertible

we only have one solution for w

(h) $(X^TX + \lambda I)w = X^Ty$

$= (X^TX + \lambda I)(w_0 + X^T\alpha)$

$= \lambda w_0 + X^TXX^T\alpha + \lambda X^T\alpha$

since X^Ty, $X^TXX^T\alpha$, $\lambda X^T\alpha \in col(X^T)$

but $\lambda w_0 \in null(X)$

so $X^Ty = X^TXX^T\alpha + \lambda X^T\alpha$ and $\lambda w_0 = 0$

$= (X^TXX^T + \lambda X^T)\alpha$

$\Rightarrow XX^Ty = (XX^TXX^T + \lambda XX^T)\alpha$

$\alpha = (XX^TXX^T + \lambda XX^T)^{-1}XX^Ty$

even if $\lambda \to 0$, $(XX^TXX^T + \lambda XX^T)$ is invertible.

then we have only one solution for α and only one solution for w_0 ($w_0 = 0$)

so we only have one solution for w.

(i) $\frac{1}{2}\|y - Xw\|_2^2 + \lambda\|\Gamma w\|_2^2$

$= \frac{1}{2}(y - Xw)^T(y - Xw) + \lambda(\Gamma w)^T(\Gamma w)$

$= \frac{1}{2}(y^Ty + w^TX^TXw - 2y^TXw) + \lambda w^T\Gamma^T\Gamma w$

$= \frac{1}{2}y^Ty + w^T(\frac{1}{2}X^TX + \lambda\Gamma^T\Gamma)w - y^TXw$

∇_w: $(\frac{1}{2}X^TX + \lambda\Gamma^T\Gamma + \frac{1}{2}X^TX + \lambda\Gamma^T\Gamma)w - X^Ty$

$= (X^TX + 2\lambda\Gamma^T\Gamma)w - X^Ty = 0$

$\Rightarrow \hat{w} = (X^TX + 2\lambda\Gamma^T\Gamma)^{-1}X^Ty$

3. (f) the highest power is D, and number of variates is L.

We can use D bars to represent power, and L stars to represent variates.

the number of bars put behind each star

$\Rightarrow$ variate corresponding to each star have a power of the number of bars put behind it

the number of bars put ahead of all stars

$\Rightarrow$ 1 to the power of number of bars, which is just one

To illustrate, we use $L=2$, $D=3$ as example:

$$\begin{array}{ll} \times\times & \quad ||| \end{array}$$

$$\begin{array}{lcl}
|||\times\times & \rightarrow & 1^3 x_1^0 x_2^0 \rightarrow 1 \\
||\times|\times & \rightarrow & 1^2 x_1^1 x_2^0 \rightarrow x_1 \\
||\times\times| & \rightarrow & 1^2 x_1^0 x_2^1 \rightarrow x_2 \\
|\times|\times| & \rightarrow & 1^1 x_1 x_2 \rightarrow x_1 x_2 \\
|\times||\times & \rightarrow & 1^1 x_1^2 x_2^0 \rightarrow x_1^2 \\
|\times\times|| & \rightarrow & 1^1 x_1^0 x_2^2 \rightarrow x_2^2 \\
\times|\times|| & \rightarrow & 1^0 x_1^1 x_2^2 \rightarrow x_1 x_2^2 \\
\times||\times| & \rightarrow & 1^0 x_1^2 x_2^1 \rightarrow x_1^2 x_2 \\
\times|||\times & \rightarrow & 1^0 x_1^3 x_2^0 \rightarrow x_1^3 \\
\times\times||| & \rightarrow & 1^0 x_1^0 x_2^3 \rightarrow x_2^3
\end{array}$$

then total number of possibility is the number of ways to arrange D bars and L stars, which is to choose D spots out of $(D+L)$ spots, which is $\binom{D+L}{D} = \binom{D+L}{L}$

$L \geq 2$ (multivariate)

(g) $D \geq 1$ for these data to be meaningful, then

the second column is $(x_{1,1}, x_{2,1}, x_{3,1}, \cdots x_{n,1})^T$

$= (\alpha_1, \alpha_2, \alpha_3 \cdots \alpha_n)^T$

the third column is $(x_{1,2}, x_{2,2}, \cdots x_{n,2})^T$

$= (\alpha_1, \alpha_2, \alpha_3 \cdots \alpha_n)^T$

thus F_L always has linearly dependent columns

This shows that we don't need the other $(L-1)$ variates, we only need to first one, so multvariate falls back to univariate.

In (e), it shows that we don't need two equal data points

These are both redundancies that we can remove from our model

(h) We can sample a $\binom{D+L}{L} \times L$ matrix, where each of the elements are randomly generated. This guarantees that $\vec{x}_i \neq \vec{x}_j$ and $x_{i,p} \neq x_{i,q}$ and then this matrix has full column rank. Then the F_L matrix generated from this matrix is also of full column rank.

4. (a) $(e^x)' = e^x$

$(e^x)^{(i)} = e^x$

1st: $e^x \approx e^{x_0} + e^{x_0}(x - x_0) = 1 + (x-0) = 1 + x$

2nd: $e^x \approx e^{x_0} + e^{x_0}(x - x_0) + \frac{1}{2}e^{x_0}(x-x_0)^2 = 1 + x + \frac{1}{2}x^2$

3rd: $e^x \approx 1 + x + \frac{1}{2}x^2 + \frac{1}{6}e^{x_0}(x-x_0)^3 = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3$

4th: $e^x \approx 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4$

$(\sin x)^{(0)} = \sin x$

$(\sin x)^{(1)} = \cos x$

$(\sin x)^{(2)} = -\sin x$

$(\sin x)^{(3)} = -\cos x$

$(\sin x)^4 = \sin x$

1st: $\sin x \approx \sin x_0 + \cos x_0 (x - x_0) = x$

2nd: $\sin x \approx x + \frac{1}{2}(-\sin x_0)(x - x_0)^2 = x$

3rd: $\sin x \approx x + \frac{1}{6}(-\cos x_0)(x - x_0)^3 = x - \frac{1}{6}x^3$

4th: $\sin x \approx x - \frac{1}{6}x^3 + \frac{1}{24}\sin x_0 (x - x_0)^4 = x - \frac{1}{6}x^3$

(b) $|f(x) - \phi_2(x)|$

$= \left| \frac{1}{6} e^{a(x)} x^3 \right| = \frac{1}{6} e^{a(x)} |x^3|$

when $x = 0$ $\quad |f(x) - \phi_2(x)| = 0$

when $x > 0$ $\quad |f(x) - \phi_2(x)| = \frac{1}{6} e^{a(x)} x^3 \Rightarrow$ monotonically increasing

when $x < 0$ $\quad |f(x) - \phi_2(x)| = -\frac{1}{6} e^{a(x)} x^3 \Rightarrow$ monotonically decreasing

so the sup is at two ends

$|f(3) - \phi_2(3)| = |e^3 - 1 - 3 - \frac{1}{2} \cdot 9| = 11.59$

$|f(-4) - \phi_2(-4)| = |e^{-4} - 1 + 4 - \frac{1}{2} \cdot 16| = 5.98$

so $\|f - \phi_2\|_\infty = 11.59$

- arbitrary m

$$|e^x - \phi_m(x)| = \left| \frac{1}{(m+1)!} e^{a(x)} x^{m+1} \right| \qquad a(x) \in [-4, 3], \quad x \in [-4, 3]$$

$$= \frac{1}{(m+1)!} e^{a(x)} |x|^{m+1}$$

$$= \frac{1}{\sqrt{2\pi(m+1)} \left(\frac{m+1}{e}\right)^{m+1}} e^{a(x)} |x|^{m+1}$$

$$= \frac{e^{a(x)}}{\sqrt{2\pi(m+1)}} \left(\frac{e|x|}{m+1}\right)^{m+1}$$

$$\leq \frac{e^3}{\sqrt{2\pi(m+1)}} \left(\frac{4e}{m+1}\right)^{m+1}$$

when $m \to \infty$, $\frac{1}{\sqrt{m+1}} \to 0$

$$\left(\frac{1}{m+1}\right)^{m+1} \to 0$$

$$|e^x - \phi_m(x)| \leq 0$$

- Outside of I, the approximation error becomes larger than the upper bound derived only for I

(c) $D > O(\log(1/\varepsilon))$

$$(D+1)\log\frac{1}{4e} + (D+\tfrac{1}{2})\log(D+1) > O(\log(1/\varepsilon)) = \log(1/\varepsilon)$$

$$\left(\frac{1}{4e}\right)^{D+1} \times (D+1)^{D+1/2} > 1/\varepsilon$$

$$(4e)^{D+1}\left(\frac{1}{D+1}\right)^{D+1/2} < \varepsilon$$

$$\left(\frac{4e}{D+1}\right)^{D+1}\frac{1}{\sqrt{D+1}} < \varepsilon$$

ignoring constants, we have proved that the upper bound of $|e^x - \phi_D(x)|$ is $< \varepsilon$

so for all $x \in I$, $|e^x - \phi_D(x)| \leq \varepsilon$

for $\sin x$, we have

$$|f(x) - \phi_D(x)| \leq \frac{1}{(D+1)!}\cdot|x|^{D+1} \leq \frac{1}{(D+1)!}4^{D+1} = \frac{1}{\sqrt{2\pi(D+1)}}\left(\frac{4e}{D+1}\right)^{D+1}$$

ignoring the constant, this upper bound error for $\sin x$ is the same as e^x, so the same conclusion can be drawn for $\sin x$

(d) $|f(x) - \phi_D(x)| \leq \frac{1}{(D+1)!}T\cdot|x|^{D+1}$ has the same form as these in (c) if ignoring the constant

so we can always find $D > O(\log(1/\varepsilon))$ to make the sup-norm less than ε

(e) $f(\vec{v}) = f(\vec{v}_0) + (\vec{v}-\vec{v}_0)^T \underbrace{[\nabla_{\vec{v}} f(\vec{v}_0)]}_{①} + \frac{1}{2!}(\vec{v}-\vec{v}_0)^T \underbrace{\nabla_{\vec{v}\vec{v}} f(\vec{v}_0)}_{②}(\vec{v}-\vec{v}_0)$

$$= f(x_0, y_0) + (x-x_0, y-y_0)\underbrace{\begin{pmatrix} f_x(x_0,y_0) \\ f_y(x_0,y_0) \end{pmatrix}}_{①} +$$

$$\frac{1}{2!}(x-x_0, y-y_0)\underbrace{\begin{pmatrix} f_{xx}(x_0,y_0) & f_{xy}(x_0,y_0) \\ f_{yx}(x_0,y_0) & f_{yy}(x_0,y_0) \end{pmatrix}}_{②}\begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix}$$

$$= f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0) +$$

$$\frac{1}{2!}\Big[f_{xx}(x_0,y_0)(x-x_0)^2 + f_{xy}(x_0,y_0)(x-x_0)(y-y_0) +$$
$$f_{yx}(x_0,y_0)(x-x_0)(y-y_0) + f_{yy}(x_0,y_0)(y-y_0)^2\Big]$$

- $f(\vec{v}) = e^x y^2$

$f_x = e^x y^2 \qquad f_{xx} = e^x y^2 \qquad f_{xy} = f_{yx} = 2e^x y$

$f_y = 2e^x y \qquad f_{yy} = 2e^x$

$$e^x y^2 = e^{x_0} y_0^2 + (\vec{v}-\vec{v}_0)^T \begin{pmatrix} e^x y^2 \\ 2e^x y \end{pmatrix} + \frac{1}{2!}(\vec{v}-\vec{v}_0)^T \begin{pmatrix} e^x y^2 & 2e^x y \\ 2e^x y^2 & 2e^x \end{pmatrix}(\vec{v}-\vec{v}_0)$$

(f) suppose $\vec{v} = \begin{pmatrix} x \\ y \end{pmatrix}$

$$g(t) = f(\vec{v}(t)) = f(t\vec{v}) = f(tx, ty) = e^{tx}(ty)^2$$

$$g'(t) = e^{tx} \cdot x\,(ty)^2 + e^{tx} \cdot 2ty^2$$

$$= e^{tx} y^2 (xt^2 + 2t) \qquad g'(t=0) = 0$$

$$g''(t) = e^{tx} \cdot x y^2 (xt^2 + 2t) + e^{tx} \cdot y^2 (2tx + 2) \qquad g''(t=0) = 2y^2$$

$$g(t) \approx g(0) + g'(0)t + \tfrac{1}{2} g''(0)(t)^2$$

$$= \tfrac{1}{2} \cdot 2y^2 t^2 = y^2 t^2$$

5. (a) make a matrix $X \in \mathbb{R}^{n\times(D+1)}$, two vectors $\vec{y} \in \mathbb{R}^{n\times 1}$, $\vec{w} \in \mathbb{R}^{(D+1)\times 1}$

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^D \\ 1 & x_2 & x_2^2 & \cdots & x_2^D \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^D \end{pmatrix} \qquad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \vec{w} = \begin{pmatrix} w_0 \\ \vdots \\ w_D \end{pmatrix}$$

then the problem becomes

$$\arg\min_{\vec{w}} \| X\vec{w} - y \|_2^2$$

(c) the larger D, the smaller training error.
larger D means larger degree of freedom in fitting, thus fits better.

If $D = n$, then $X \in \mathbb{R}^{n\times(n+1)}$ $\quad$ $\text{rank}(X) = n$.

$\text{rank}(X^TX) = \text{rank}(X^T) = n$. $\quad$ $X^TX \in \mathbb{R}^{(n+1)\times(n+1)}$

$X^TX\vec{w} = X^T\vec{y}$

(X^TX) is not full rank and cannot be inverted.

(d) for the same D, generally test error are larger than training error. (Training error → bias, test error → bias + variance)
Training error generally decreases as D gets larger (bias ↓)
test error first decreases and then increases as D gets larger (as D ↑, bias ↓, variance ↑)

(e) we should choose D of the smallest test error since test error reflects true error, then $D = 4$, this is reasonable since number of training data is $n = 20$.