2) A) L1 and L2 regularization

L1 regularization or Lasso

L1 loss function $= \sum_{i=1}^{N} (y^{(i)} - x^{(i)}\theta)^2 + \lambda \|\theta\|_1 \rightarrow \theta = \text{argmin}[(X\theta - Y)^T (X\theta - Y) + \lambda \|\theta\|_1]$

the penalty that prevents further overfitting is $\|\theta\|_1$ in $L1$.

L2 regularization or Ridge

L2 loss function $= \sum_{i=1}^{N} (y^{(i)} - \theta^T x^{(i)})^2 + \lambda \|\theta\|_2^2 \Rightarrow \theta = \text{argmin}[(X\theta - y)^T (X\theta - y) + \lambda \|\theta\|_2^2]$

the penalty is $\|\theta\|_2^2$ or $\theta = (X^T X + \lambda I)^{-1} X^T Y$

Both of them are used to reduce overfitting possibilities.

L2 shrinks all coefs toward zero but not exactly zero.

L1 is more sparse and sets some of the coefs to zero.

B) Ridge Regression

$$L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$$

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

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

$$w^T x^T y = (w^T x^T)(y) \xrightarrow{(AB)^T = B^T A^T} y^T X w$$

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

$$\frac{\partial L(w)}{\partial w} = 0 \Rightarrow -2 y^T X + 2 X^T X w + 2\lambda w = 0$$

$$\Rightarrow (X^T X + \lambda) w = y^T X \xrightarrow{A^T B = AB^T} (X^T X + \lambda) w = y X^T$$

$$\Rightarrow w = (X^T X + \lambda)^{-1} Y X^T = \underbrace{(X^T X + \lambda)^{-1} X^T Y}$$