7.3) MLE for ridge regression:

Negative ~~likeli~~ likelihood for w

$$J(w,w_0) = NLL(w,w_0) = (y - Xw - w_0 1)^T (y - Xw - w_0 1) + \lambda w^T w$$

Here, $X_{n \times d}$ – input feature vector matrix (n-examples d-features)

$y_{n\times 1}$ – output vector (consider single output) (n-examples)

$w_{d\times 1}$ – weight vector for output y.

$(1w_0) = (1 w_0)_{n\times 1}$ – w_0-add-on coefficient to weight. Its just column-vector of size $n\times 1$ all with w_0.

Compute the gradient w.r.t w_0 and w to get the best estimate.

$$\nabla_{w_0} J(w, w_0) = \frac{\partial}{\partial w_0} J(w, w_0)$$

$$= \cancel{\frac{\partial}{\partial w_0}} \Big[\cancel{(1w_0)^T 1 w_0} +$$

$$= \frac{\partial}{\partial w_0} \Big[(y^T - w^T X^T - (w_0 1)^T)(y - Xw - (w_0 1)) + \lambda w^T w \Big]$$

$$= \frac{\partial}{\partial w_0} \Big[(w_0 1)^T (w_0 1) + 2 (w_0 1)^T X w - 2 (w_0 1)^T y \cancel{\ldots X^T (w_0 1)} \Big] + 0$$

$$= \frac{\partial}{\partial w_0}\left[\sum_{i=1}^{n} w_0^2 + 2w_0 \sum_{i=1}^{n}\sum_{j=1}^{d} x_{ij} w_j + 2w_0 \sum_{i=1}^{n} y_i \right]$$

Set $\nabla_{w_0} J(w, w_0) = 0$ to get estimate

$$\Rightarrow 2w_0 \sum_{i=1}^{n} 1 + 2n \sum_{j=1}^{d} \hat{x}_j w_j - 2\sum_{i=1}^{n} y_i = 0$$

$$2w_0 \cdot n + 0 = 2\sum_{i=1}^{n} y_i$$ [$\because$ each $\hat{x}_j$ is mean of jth feature and it's 0 $\bar{x}_j = 0$ As per given note]

$$w_0 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

Now, compute w.r.t w and set to 0.

$$\nabla_w J(w, w_0) = 0.$$

$$\frac{\partial}{\partial w} J(w, w_0) = 0.$$

$$\frac{\partial}{\partial w}\left[w^T (X^T X) w - 2y^T X w + 2(w_0 1)^T X w + \lambda w^T w \right] = 0.$$

As per equations.

$$\frac{\partial (a^T A a)}{\partial a} = (A + A^T) a$$

$$\frac{\partial (b^T a)}{\partial a} = b.$$

$$2(X^T X) w - 2X^T y + 2X^T (w_0 1) + 2\lambda I w = 0$$

$$\left((X^T X) + \lambda I \right) w = X^T y \quad (\because X^T w_0 = 0)$$ like above.

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

7.4) MLE for σ^2 in linear regression.

~~Negative~~ log likelihood of linear Regression

$$l(\theta) = \frac{-1}{2\sigma^2} RSS(w) - \frac{N}{2} \log(2\pi\sigma^2)$$

To, get the best estimate of σ^2, take partial derivative w.r.t σ^2 and set 0.

$$\nabla l_{\sigma^2}(\theta) = \frac{\partial}{\partial \sigma^2} l(w, \sigma^2) = 0$$

$$\frac{\partial}{\partial \sigma^2}\left(\frac{-1}{2\sigma^2} RSS(w) - \frac{N}{2}\log(2\pi\sigma^2)\right)$$

$RSS(w)$ = Residual square errors

$$= \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

$$\frac{-1}{\cancel{2}} \cdot \frac{-\cancel{2}}{\sigma^3} RSS(w) - \frac{N}{\cancel{2}} \cdot \frac{1}{\cancel{2\pi}\sigma^2} \cdot \cancel{2\pi} \cdot \cancel{2}\sigma = 0$$

$$\frac{RSS(w)}{\sigma^3} = \frac{N}{\sigma}$$

$$\sigma^2 = \frac{1}{N} RSS(w)$$

Best estimate for error-variance. $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{w}^T x_i)^2$

Here, we substitute MLE for $w = \hat{w}$ and to get MLE for $\sigma^2 = \hat{\sigma}^2$

Scanned by CamScanner

7.5) Linear regression is modeled as

$$P_{w, w_0, \sigma^2}(y/x) = \mathcal{N}(w^T x + w_0, \sigma^2)$$

w_0 - offset coefficient (scalar)

w - weight vector

y - output variable

x - input vector

Negative log likelihood for linear regression

$$NLL(\theta) = J(w_0, w, \sigma^2) = \frac{1}{2\sigma^2}(y - Xw - 1w_0)^T(y - Xw - 1w_0) + \frac{N}{2}\log(2\pi\sigma^2).$$

Minimize NLL to estimate the parameters:

$$\frac{\partial J}{\partial w_0} = 0 \Rightarrow \frac{\partial}{\partial w_0}\left[(y - Xw - 1w_0)^T (y - Xw - 1w_0)\right] = 0$$

Here, y - output vector of all examples $(n \times 1)$

X - input examples x features matrix $(n \times d)$

w - weight vector $(d \times 1)$

$1w_0$ - Column vector with all w_0's. $(n \times 1)$

$$\frac{\partial}{\partial w_0}\left[(1w_0)^T \cdot (1w_0) + 2(1w_0)^T Xw - 2(1w_0)^T y\right] = 0$$

$$\frac{\partial}{\partial w_0}\left[\sum_{i=1}^{n} w_0^2 + 2w_0 \sum_{i=1}^{n} x_i^T w - 2w_0 \sum_{i=1}^{n} y_i\right] = 0$$

y_i - one example of o/p vector

x_i - one example (d-features) of i/p.

Scanned by CamScanner

$$2w_0 \cdot n + 2 \cdot \sum_{i=1}^{n} x_i^T \cdot w - 2\sum_{i=1}^{n} y_i = 0$$

$$\cancel{w_0 = \bar{y} - \bar{x}^T}$$

$$w_0 = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} x_i^T \cdot w$$

$$\boxed{w_0 = \bar{y} - \bar{x}^T w.}$$

Now, estimate for W

$$\frac{\partial J}{\partial w} = 0 \Rightarrow \frac{\partial}{\partial w}\left[(y - Xw - 1w_0)^T (y - Xw - 1w_0) \right] = 0$$

$$\frac{\partial}{\partial w}\left\{ w^T (X^T X) w - 2y^T X w + 2(1w_0)^T X w \right] = 0.$$

Using the matrix derivation equations used in (7.3) calc

$$2(X^T X) w - 2X^T y + 2X^T(1w_0) = 0$$

we'll write these equations ~~as~~ in terms of (x_i, y_i) i.e for each example. after substituting $\boxed{1w_0 = \bar{y} - \bar{x}^T w}$

$$\left(\sum_{i=1}^{N} x_i x_i^T\right) w - \sum_{i=1}^{N} x_i y_i + \sum_{i=1}^{N} x_i \bar{y} - w\sum_{i=1}^{N} x_i \bar{x}_i^T = 0$$

$$\cancel{\left(\sum_{i=1}^{N} x_i x_i^T - \sum_{i=1}^{N} x_i \bar{x}_i^T\right) w - \sum_{i=1}^{N} (x_i y_i - x_i \bar{y})} = 0$$

$$w \sum_{i=1}^{N} (x_i x_i^T - x_i \bar{x}_i^T) - \left(\sum_{i=1}^{N} x_i y_i - x_i \bar{y}\right) = 0$$

$$(X_c^T X_c) w = X_c^T y_c$$

Here X_c = Matrix with each row centred around its mean $(x_i - \bar{x})$

y_c = centred output vector $(y - \bar{y})$

$$\boxed{w = (X_c^T X_c)^{-1} X_c^T y_c}$$

This is equivalent to

$$w = \left[\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T\right]^{-1} \left[\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})\right]$$

7.6) Here, number of input features are 1 i.e $D = 1$

So, we have only one input variable x_1 and one output variable y.

From above form we can easily see that

$$w_0 = \bar{y} - \bar{x} w_1$$

(As we have only one feature we'll have only one weight) w_1.

$$w_0 \simeq E[y] - w_1 E[x]$$

And from equation of w_1 we can see

$(X_c^T X_c)^{-1}$ is one entry with inverse equal to

$$\frac{1}{\sum_i (x_i - \bar{x})^2}, \text{ so}$$

$$w_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

7.7) we don't have original data but we have sufficient statistics.

$\bar{x}^{(n)}, \bar{y}^{(n)}$, - Mean of i/p, o/p's of n examples.

$C_{xx}^{(n)}, C_{xy}^{(n)}, C_{yy}^{(n)}$ - Variance among x, Covariance between x,y Variance among y.

a) From exercise 7.6, for w_1, we need minimal set of statistics as $C_{xy}^{(n)}, C_{xx}^{(n)}$

$$w_1 = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}}.$$

b) From exercise 7-6, for w_0 we need ~~and~~ minimal set of statistics as $\bar{y}^{(n)}, \bar{x}^{(n)}$ and $C_{xy}^{(n)}, C_{xx}^{(n)}$

$$w_0 = \bar{y} - w_1\bar{x} = \bar{y}^{(n)} - \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}} \cdot \bar{x}^{(n)}$$

c) For online learning, to update our sufficient statics without looking at old data.

$$\bar{x}^{(n+1)} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}\left(n\bar{x}^{(n)} + x_{n+1}\right)$$

$$= \frac{n+1-1}{n+1} \cdot \bar{x}^{(n)} + \frac{1}{n+1} x_{n+1}$$

$$= \bar{x}^{(n)} + \frac{1}{n+1}\left(x_{n+1} - \bar{x}^{(n)}\right)$$

In the same way, we can update $\bar{y}$ ... like

~~this~~ $\bar{y}^{(n+1)} = \frac{1}{n+1}\sum_{i=1}^{n+1} y_i = \frac{1}{n+1}\left(n\bar{y}^{(n)} + y_{n+1}\right)$

$$\bar{y}^{(n+1)} = \bar{y}^{(n)} + \frac{1}{n+1}\left(y_{n+1} - \bar{y}^{(n)}\right)$$

This is nothing but new estimate equals to old estimate plus correction.

d) Like given expression for $C_{xy}^{(n+1)}$ we can update our online learning for $C_{xx}^{(n+1)}$ too.

$$C_{xx}^{(n+1)} = \frac{1}{n+1}\left[\sum_{i=1}^{n+1}\left(x_i - \bar{x}^{(n+1)}\right)^2\right]$$

$$= \frac{1}{n+1}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}^{(n+1)}\right)^2 + \left(x_{n+1} - \bar{x}^{(n+1)}\right)^2\right]$$

$= \frac{1}{\cancel{n+1}}\Big[$ Substitute $\bar{x}^{(n+1)} = \bar{x}^{(n)} + \frac{1}{n+1}\left(x_{n+1} - \bar{x}^{(n)}\right)$

$$= \frac{1}{n+1}\left[\sum_{i=1}^{n}\left[x_i - \bar{x}^{(n)} - \frac{1}{n+1}\left(x_{n+1} - \bar{x}^{(n)}\right)\right]^2 + \left[x_{n+1} - \bar{x}^{(n)} - \frac{1}{n+1}\left(x_{n+1} - \bar{x}^{(n)}\right)\right]^2\right]$$

Consider $\frac{x_{n+1} - \bar{x}^{(n)}}{n+1} = K$

$$= \frac{1}{n+1}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}^{(n)} - \frac{K}{\cancel{n+1}}\right)^2 + \frac{(nK)^2}{\cancel{(n+1)^2}}\right]$$

$$= \frac{1}{n+1}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}^{(n)}\right)^2 + \sum_{i=1}^{n}\frac{K^2}{\cancel{(n+1)^2}} - 2K\sum_{i=1}^{n}\left(x_i - \bar{x}^{(n)}\right) + n^2K^2\right]$$

Scanned by CamScanner

$$= \frac{1}{n+1}\left[n C_{xx}^{(n)} + nk^2 + (0) + n^2k^2 \right]$$

since $C_{xx}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}^{(n)})^2$ & $\sum_{i=1}^{n} x_i - \bar{x}^{n} = 0$

$$= \frac{1}{n+1}\left[n C_{xx}^{(n)} + nk^2 (n+1) \right]$$

$$k = \frac{1}{n+1}\left(x_{n+1} - \bar{x}^{(n)} \right) = \bar{x}^{(n+1)} - \bar{x}^{(n)}$$

So $$C_{xx}^{(n+1)} = \frac{1}{n+1}\left[n C_{xx}^{(n)} + n(n+1)\left(\bar{x}^{(n+1)} - \bar{x}^{(n)} \right)^2 \right]$$