$w^* = \hat{w}_{ML} \leftarrow$ Max likelihood $= \boxed{(XX^T)^{\dagger} Xy}$ random

$y/x = w^T x + \epsilon \rightsquigarrow N(0, \sigma^2)$

$N(w^T x, \sigma^2)$

$\{(x_1, y_1), \ldots (x_n, y_n)\}$

$w^T x_1 + \epsilon_1$

$w \in \mathbb{R}^d \; ; \; \hat{w}_{ML} \in \mathbb{R}^d$

Want a way to understand how good $\hat{w}_{ML}$ is in estimating w

Mean Squared error $\rightarrow \mathbb{E}\left[\|\hat{w}_{ML} - w\|^2\right] = \sigma^2 \cdot \text{trace}\left((XX^T)^{-1}\right)$

over randomness in y

$A = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_d \end{bmatrix}$

$tr(A) = \sum_{i=1}^{d} a_i = \sum_{i=1}^{d} \lambda_i$

i^{th} Eigenvalue of A

$\boxed{\text{trace}\left((XX^T)^{-1}\right)}$

Let Eigenvalues of (XX^T) be $\{\lambda_1, \ldots, \lambda_d\}$

Eigenvalues of $(XX^T)^{-1}$ are $\{\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_d}\}$

Mean sq. error $(\hat{w}_{ML})$

$\mathbb{E}(\|\hat{w}_{ML} - w\|^2) = \sigma^2 \left(\sum_{i=1}^{d} \frac{1}{\lambda_i}\right)$

Consider the following estimator:

$\hat{w}_{new} = (XX^T + \lambda I)^{-1} Xy$

$\lambda \in \mathbb{R}_+$, $I \in \mathbb{R}^{d \times d}$

$\hat{w}_{ML} = (XX^T)^{-1} Xy$

For some matrix A, let Eigen values be $\{\lambda_1, \ldots, \lambda_d\}$

. What are Eigenvalues of $A + \lambda I$? $\{\lambda_1 + \lambda, \ldots, \lambda_d + \lambda\}$

$\text{trace}\left((XX^T + \lambda I)^{-1}\right) = \left(\sum_{i=1}^{n} \frac{1}{\lambda_i + \lambda}\right)$

$A v_1 = \lambda_1 v_1$

$(A + \lambda I) v_1 = A v_1 + \lambda v_1$

$= \lambda_1 v_1 + \lambda v_1$

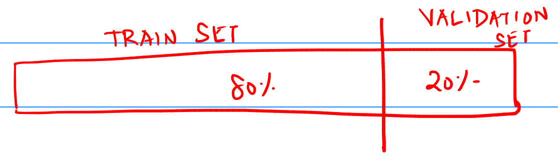
$= (\lambda_1 + \lambda) v_1$

EXISTENCE Thm : (Informal)

$\exists \lambda \in \mathbb{R}$ s.t

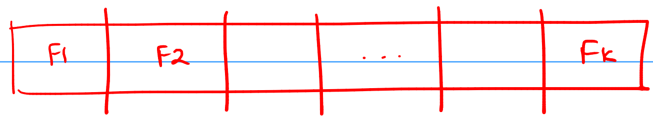
$\hat{w}_{new} = (XX^T + \lambda I)^{-1} Xy$ has lesser mean sq. error than $\hat{w}_{ML}$

In practice, find λ by **CROSS VALIDATION**



- Train on the training set and check for error on validation set.
- Pick λ that gives least error.

K-FOLD CROSS VALIDATION



- Train on folds $\{F_1, \dots F_{i-1}, F_{i+1}, \dots F_k\}$
- Validate on F_i

- Pick λ that gives least average error.

LEAVE ONE OUT CROSS VALIDATION

$\hat{w}_{new} = (XX^T + \lambda I)^{-1} Xy$

IS there an alternate way to understand $\hat{w}_{ML}$?

BAYESIAN MODELING

- NEED A PRIOR on w i.e., $P(w)$ $\mathbb{R}^d$.

For simplicity. Can use σ^2 as well.

LIKELIHOOD $\quad y/x \sim N(w^T x, 1)$

A CHOICE FOR PRIOR

$\in \mathbb{R}^d$

$$w \sim N(0, \gamma^2 I)$$

$$\begin{bmatrix} \gamma^2 & & 0 \\ & \gamma^2 & \\ 0 & & \ddots \; \gamma^2 \end{bmatrix}$$

COVARIANCE MATRIX $\mathbb{R}^{d \times d}$.

As usual,

$$P\left(w \,/\, \{(x_1,y_1) \cdots (x_n,y_n)\}\right) \propto \underline{P\left(\{(x_1,y_1), \cdots, (x_n,y_n)\} / w\right)} \cdot \underline{P(w)}$$

$$\propto \left(\prod_{i=1}^{n} e^{-\frac{(y_i - w^T x_i)^2}{2}}\right) \cdot \underbrace{\left(\prod_{i=1}^{d} e^{-\frac{(w_i - 0)^2}{2\gamma^2}}\right)}_{e^{-\sum_{i=1}^{d} \frac{w_i^2}{2\gamma^2}}}$$

$$\propto \underbrace{\left(\prod_{i=1}^{n} e^{-\frac{(y_i - w^T x_i)^2}{2}}\right) \cdot e^{-\frac{\|w\|^2}{2\gamma^2}}}$$

How will the MAP estimate look like?

$$\hat{w}_{MAP} = \arg\max_{w} \sum_{i=1}^{n} -\frac{(y_i - w^T x_i)^2}{2} - \frac{\|w\|^2}{2\gamma^2}$$

$$\hat{w}_{MAP} = \arg\min_{w} \underbrace{\underbrace{\frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2} + \underbrace{\frac{1}{2\gamma^2} \|w\|^2}} \rightarrow f(w)$$

Take gradient, set it to 0 to solve for $\hat{w}_{MAP}$.

$$\nabla F(w) = (XX^T)w - Xy + \frac{w}{\gamma^2} \quad [\text{verify}]$$

$$\boxed{\hat{w}_{MAP} = \left(XX^T + \frac{1}{\gamma^2} I\right)^{-1} Xy}$$

CROSS VALIDATE in PRACTICE.

CONCLUSION: MAP ESTIMATION for linear regression with a Gaussian prior $\boxed{\mathcal{N}(0, \gamma^2 I)}$ for w is equivalent to "NEW" estimator we used earlier.

LINEAR REGRESSION

$$\hat{w}_{ML} = \arg\min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$\frac{1}{\gamma^2}$

RIDGE REGRESSION

$$\hat{w}_R = \arg\min_{w} \underbrace{\sum_{i=1}^{n} (w^T x_i - y_i)^2}_{\text{LOSS}} + \underbrace{\lambda \|w\|^2}_{\text{REGULARIZER}}$$

RIDGE

---

f_1 height	f_2 weight	f_3 2height + 3weight	label
1	1	1	3height+4weight.
0	c_1	c_2	
3	4	0	