Week 6 MLT Lecture 1 :

$w^* = \hat{w}^{ML}$ ← max likelihood $= (xx^T)^{\dagger} xy$

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

↳ can be thought of as a Gaussian → $N(w^T x, \sigma^2)$

How good is $\hat{w}^{ML}$ as a guess for the true w ?

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

↓

derived out of an optimization problem

i.e, we maximize the likelihood & get $\hat{w}^{ML}$ as the solution of an optimization problem !!.

We want a way to understand how good $\hat{w}^{ML}$ is in estimating w ?

$$E\left(\| \hat{w}_{ML} - w \|^2\right) = \sigma^2 . \text{trace}\left((xx^T)^{-1}\right)$$

↳ expectation over randomness in y

⇒ The avg deviation between your estimated $\hat{w}_{ML}$ & true w is product of 'variance of the noise that we add to $w^T x$' (σ^2) & ...

DATE :

and ' trace of inverse of cov. matrix '.
trace $\left((XX^T)^{-1}\right)$

It also means that it depends on the data features via the trace of our inverse of covariance matrix!, i.e,

→ depending on how the features themselves are related to each other will also affect Estimators Goodness !!

LECTURE : 2 Cross Validation for Minimizing MSE

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

↳ sum of diagonal entries of matrix

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

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

let the eigenvalues of (XX^T) be $\{\lambda_1, \ldots, \lambda_d\}$

∴ Eigenvalues of $(XX^T)^{-1} = \{1/\lambda_1, \ldots, 1/\lambda_d\}$

Mean Sq Error $(\hat{w}_{ML})$

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

JINDAL
PAGE NO.
DATE :

1 Consider the following estimator :

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

$\in \mathbb{R}_+$; $(xx^T + \lambda I)^{-1} \in \mathbb{R}^{d \times d}$

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

- What are Eigenvalues of $A + \lambda I$?

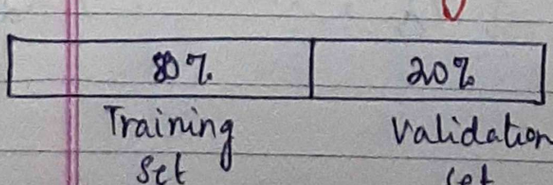
$\{\lambda_1 + \lambda, \ldots, \lambda_d + \lambda\}$

$$\text{trace}\left[(xx^T + \lambda I)^{-1}\right] = \left(\sum_{i=1}^{d} \frac{1}{\lambda_i + \lambda}\right)$$

EXISTENCE Theorem :

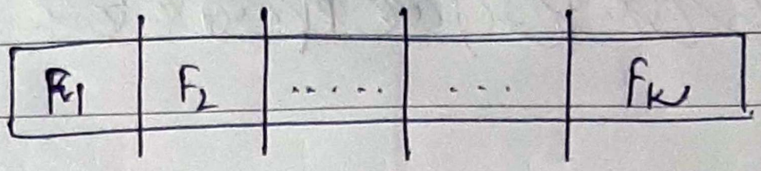
$\exists \lambda \in \mathbb{R}$ such that $\hat{w}_{new} = (xx^T + \lambda I)^{-1} xy$ has lesser mean sq error than $\hat{w}_{ML}$.

In practice, find λ by cross validation



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

K- Fold Cross Validation



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

→ Pick λ that gives least average error !!

Leave One Out Cross Validation

→ not good for large data computationally !!

LECTURE : 3 Bayesian Modelling for Linear Regression

BAYESIAN MODELLING

Need a prior on w ; i.e., $P(w)$ $\rightarrow \mathbb{R}^d$

Likelihood $Y/x \sim N(w^T x, \sigma^2 I)$ ← for simplicity can use σ^2 as well

A Choice for prior :

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

$\vec{0} \in \mathbb{R}^d$

covariance matrix $\mathbb{R}^{d \times d}$

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

1.

As usual,

$$P\left(w \mid \{(x_1,y_1) \dots (x_n,y_n)\}\right) \propto \underbrace{P\left(\{(x_1,y_1) \dots (x_n,y_n)\} \mid w\right)}_{\text{Likelihood}} \cdot \underbrace{P(w)}_{\text{Prior}}$$

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

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

How will the Max. Aposterior estimator look like?

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

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

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

- $\nabla f(w) = (xx^T)w - xy + \frac{w}{\gamma^2}$

$$\hat{w}_{MAP} = \left[(xx^T) + \frac{1}{\gamma^2} I \right]^{-1} xy$$

→ Cross validated in practice

Conclusion : MAP estimation for linear regression with a Gaussian prior $N(0, \gamma^2 I)$ for w is equivalent to "NEW" estimate we used earlier.

LECTURE 4 : Ridge Regression

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

Ridge Regression $\hat{w}_R = \underset{w}{\arg\min} \sum_{i=1}^{n} (w^T x_i + y_i)^2 + \lambda \|w\|^2$

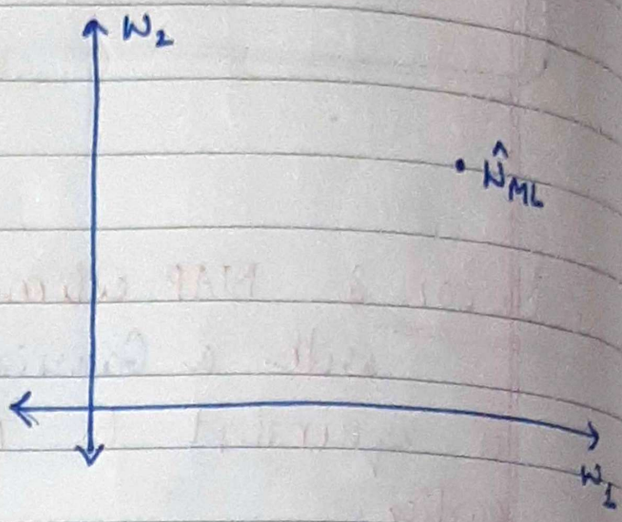
($\hat{w}_R$: Ridge; Loss; Regularizer)

If there are multiple w's which have the same loss, then the w that you would prefer is the one that has the least length — the one that has many small components !

LECTURE : 5 Relation Between Sol^n of L.R & R.R

PARAMETER SPACE

Where is $\hat{w}_R$ - sol^n of the ridge regression problem?



Ridge Regression :

(A) → $\arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2$

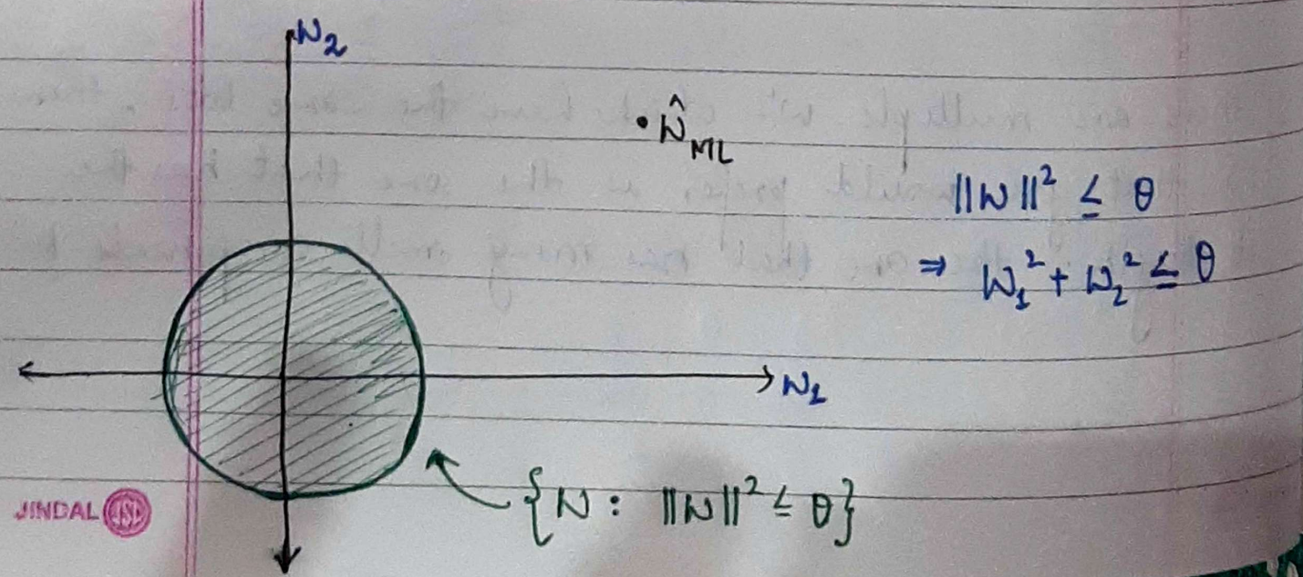
is equivalent to

≡

$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2$ ← (B)

such that $\|w\|^2 \leq \theta$ ← depends on λ

For every choice of $\lambda > 0$, $\exists \theta$ such that the optimal solutions of problems (A) & (B) coincide.



- What is the loss / error / obj function value of linear regression at $\hat{w}_{ML}$?

$$\sum_{i=1}^{n} \left(\hat{w}_{ML}^T x_i - y_i \right)^2 = \underbrace{f\left(\hat{w}_{ML}\right)}_{\text{LOSS FN}}$$

↑ Can the least value among all possible w's.

Consider the set of all w's such that

$$f(w) = f(\hat{w}_{ML}) + c \qquad \boxed{c > 0}$$

$$S_c = \{ w : f(w) = f(\hat{w}_{ML}) + c \}$$

→ i.e, every $w \in S_c$ satisfies

$$\underbrace{\| x^T w - y \|^2}_{f(w)} = \underbrace{\| x^T \hat{w}_{ML} - y \|^2}_{f(\hat{w}_{ML})} + c$$

On simplification : we get,

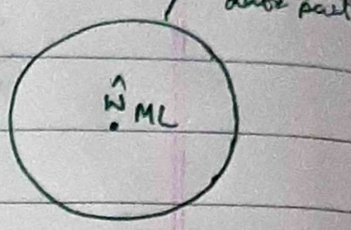
$$(w - \hat{w}_{ML})^T (xx^T) (w - \hat{w}_{ML}) = c'$$

↑ some constant that depends on $c, (xx^T), \hat{w}_{ML}$ and not on w.

If $XX^T = I$,

$$(w - \hat{w}_{ML})^T I (w - \hat{w}_{ML}) = C'$$

$$\Rightarrow \quad \| w - \hat{w}_{ML} \|^2 = C'$$



Conclusion :

1. Ridge regression pushes weight values towards 0. but does not necessarily make it 0.

LECTURE 6 : Relationship b/w soln of L.R & Lasso Regression

- An alternate way to regularize would then be using L_1 norm instead of L_2 norm

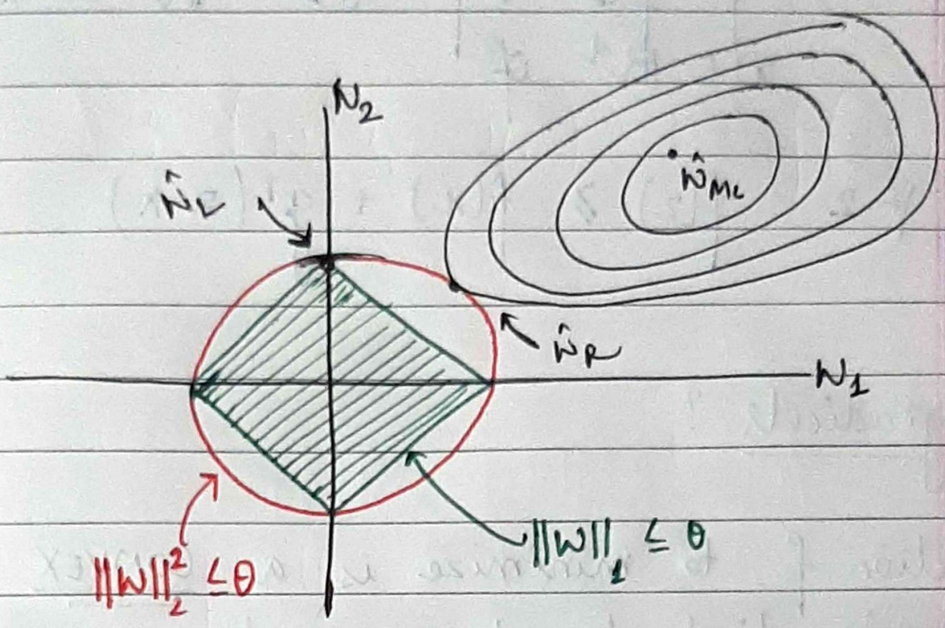
L_1 norm : $\|w\|_1 = \sum_{i=1}^{d} |w_i|$

L_1 Regularization :

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} (w^T x_i - y_i)^2 + \lambda \|w\|_1$$

It is equivalent to,

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \quad \text{such that} \quad \|w\|_1 \leq \theta$$



- Instead of L_2 regularizer if you used an L_1 penalty or a regularizer you will get more sparse solution.

↑ This is known as LASSO Regression

LASSO : Least Absolute Shrinkage & Selection Operator

LECTURE 7 : Characteristics of Lasso Regression

- Lasso does not have a closed form solution.
- Sub-gradient methods are usually used to solve LASSO.

Subgradient : A vector $g \in \mathbb{R}^d$ is a subgradient of $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ if

$$\forall z \quad f(z) \geq f(x) + g^T(z - x)$$

Why subgradients ?

If function f to minimize is a CONVEX fn, then subgradient descent converges !

- Lasso problem is a convex optimization problem !!

There are other special purpose methods for LASSO
Eg : IRLS (Iterative reweighted least squares)