L2 Regularization

overfitting and underfitting

$$w^* = \arg\min \sum_{i=1}^{d} \log\left(1 + \exp\left(-y_i w^T x_i\right)\right)$$

$\exp(-x)$

plot e^{-x}

↑

always +ve

but $z = y_i w^T x_i$

$$\arg\min \sum_{i=1}^{d} \log\left(1 + \exp(-z_i)\right)$$

↑

$\exp(-z_i) \geq 0$

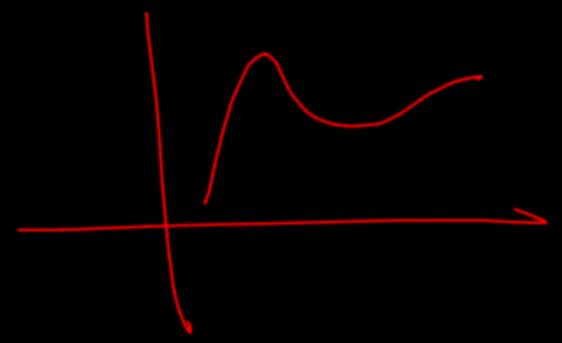
=

↑

ie $\log(1 + \exp(-z_i)) \geq 0$

ie. $\sum_{i=1}^{n} \log(1 + \exp(-z_i)) \geq 0$

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1+\exp(-z_i)\right) \geq 0$$



Minimal → '0' ← Best

But when '0' → occurs

$z_i = y_i \, w^T \, x_i \rightarrow \uparrow \rightarrow \infty$

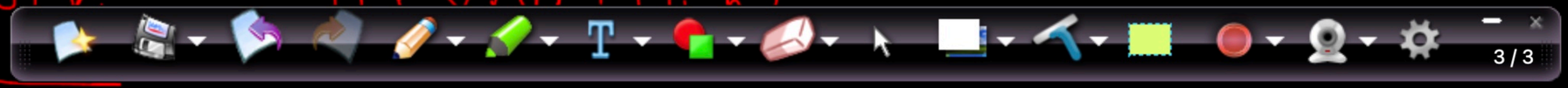
w^T

$w \rightarrow \infty$ big val

$e(\infty) \rightarrow 0$

$z_i \rightarrow \infty$

$z_i = y_i w^T x_i$



$z_i \to \infty$

↳ modify my (w) such way that $z_i \to \infty$

If $z_i = +ve$ → correct classification

2) $z_i \to +\infty$ (minimum value)

Regularization

↑

"overfitting"

→ If I pick my (w) s.t.

a) all the training pts are correctly classified

b) $z_i \to \infty$

$$w^* = \underset{w}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \log(1+\exp(-z_i))}_{\text{loss-term}} + \underbrace{\lambda \, w^T \cdot w}_{x}$$

$G =$

L1 L2

$\lambda = 0$

$\lambda \uparrow$

$\|w\|$

$\lambda \cdot \underbrace{\|w\| \cdot \|w\|}$

$\lambda = 0$

overfitting $\uparrow$

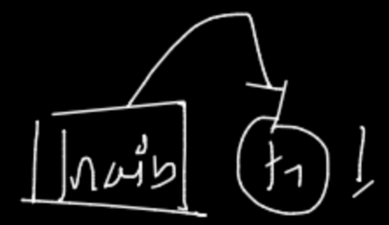
$w \to \infty$

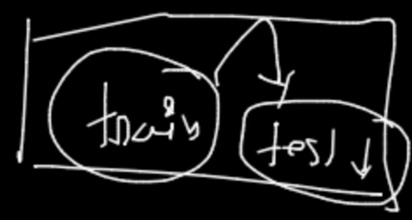
train

test $\downarrow$

train

test $\downarrow$





Hyper-parameter tuning : Search / optimisation

if $\begin{cases} \lambda = 0 \longrightarrow \text{overfitting} \\ \lambda = \infty \longrightarrow \text{underfitting} \end{cases}$

(i) GridSearchCV

(ii) RandomSearchCV

Sklearn

(iii) Bayesian optimisation → "Hyperopt"

$\lambda = 0.001$

$M(x)$

one p

$\lambda = [0.1, 0.01, 0.001, 0.0001]$

$\rightarrow M(\cancel{0.01}, \cancel{0.1})$

loss

$M(D_x)$

Note → 85 %

86 %

loss

$M(\lambda_1, \lambda_2)$

$KNN(K = ?, \text{metrics} = ?)$

$\lambda_1 = [0.1, 0.01, 0.001, 0.0001]$

$\lambda_2 = [0.1, 0.01, 0.001, 0.0001]$

$\lambda_3 = [\lambda_1 = 0.001, \lambda_2 = 0.01]$ → min

100

200 min

get pm

1st

$M1(0.1, \cancel{0.1})$ → loss/error → 0.54

loss/error → 0.55

error/loss → 0.54

loss →

0.001

~~0.001~~ → 2 min

~~0.01~~

Grid search

λ_1 and λ_2

$\uparrow$

$\lambda_1 \longrightarrow m^1$

$\lambda_1 \lambda_2 \longrightarrow m_1 \times m_2 \longrightarrow m^2$

$\vdots$

$K \longrightarrow m^K$

$K\uparrow \longrightarrow m^K$

$\uparrow\downarrow$

101 times ?

$\Downarrow$

(Big data) $\longrightarrow$ Random search CV

$\uparrow$

$\uparrow$