→ Deep Learning: Learning Theory ∪ Finite hypothesis

Infinite hypothesis → VC-dimension

✓ $m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$

for finite hypothesis ∪ train error = 0

✓ $m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$

finite hypothesis ∪ trainError ≠ 0

$|H| \to \infty$ Model complexity is computed using VC dimension of the hypothesis class.

VC dimension: The maximum number of samples that can be shattered by a hypothesis class is called VC-dimension.

... dimension (VC)

Vapnik-Chervonenkis ( )

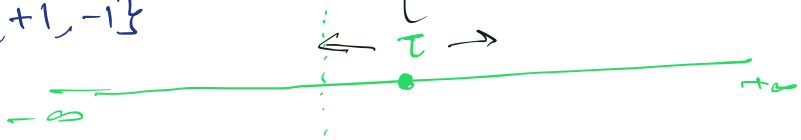
SHATTER: For all possible labellings of the samples, if there exist a hypothesis (model) 'h' with zero error (correctly classify all samples) in the hypothesis class 'H' is called Shattering

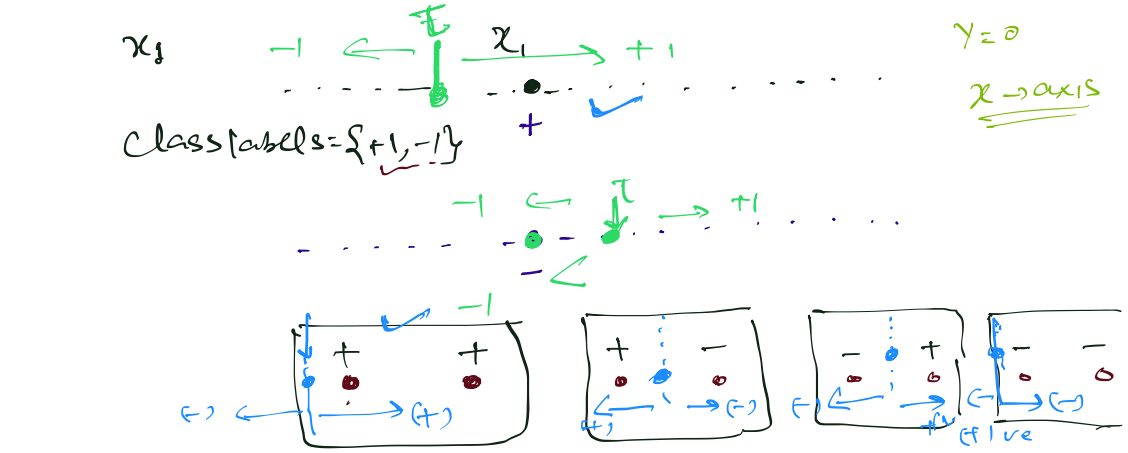
Example:

Let Threshold: τ $\begin{cases} +1 \text{ if } \tau \geq 0 \\ -1 \text{ otherwise} \end{cases}$

Class: {+1, -1}

Samples:



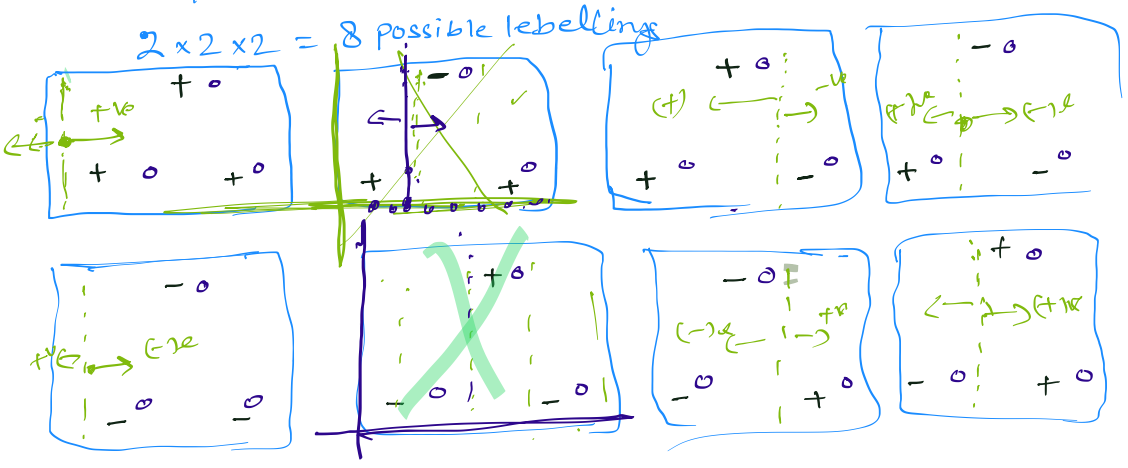


2 - Samples, 2 classes :

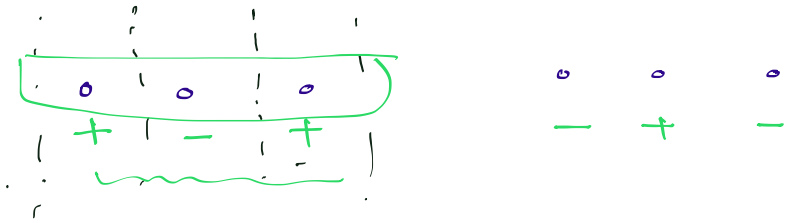
$2 \times 2 = 4$

3 - Samples, 2 classes

$2 \times 2 \times 2 = 8$ possible labellings



threshold does not exist which can shatter.



VC-dimension for threshold hypothesis class is

Q:- It can shatter a maximum of

2 samples.

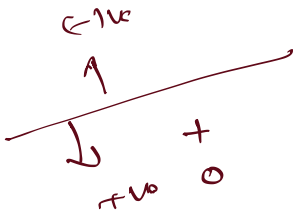
Example 2.

H: Linear models / Linear hyperplane.

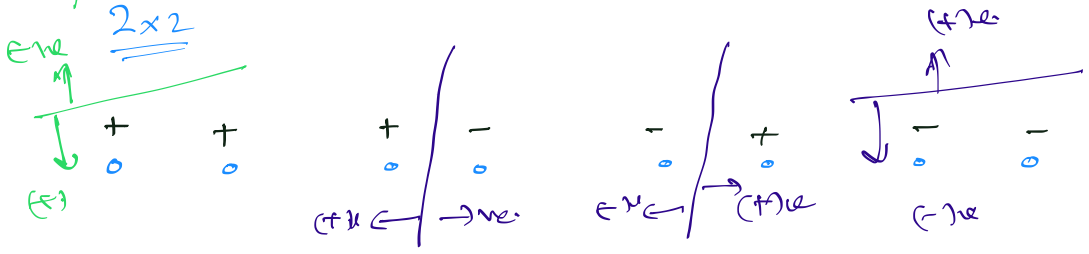
Let example be H: Linear model in 2D (line)

Class = {+1, -1}

1. Sample:



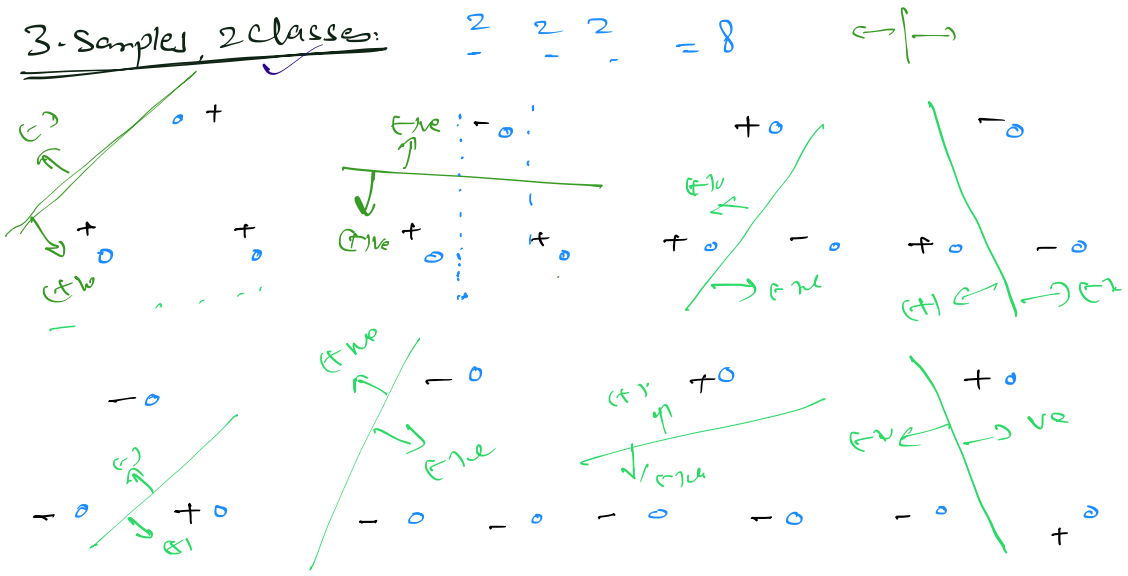
2 Samples, 2 Classes. $2 \quad 2 = 4$

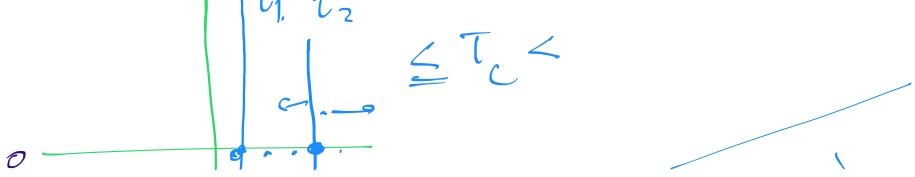


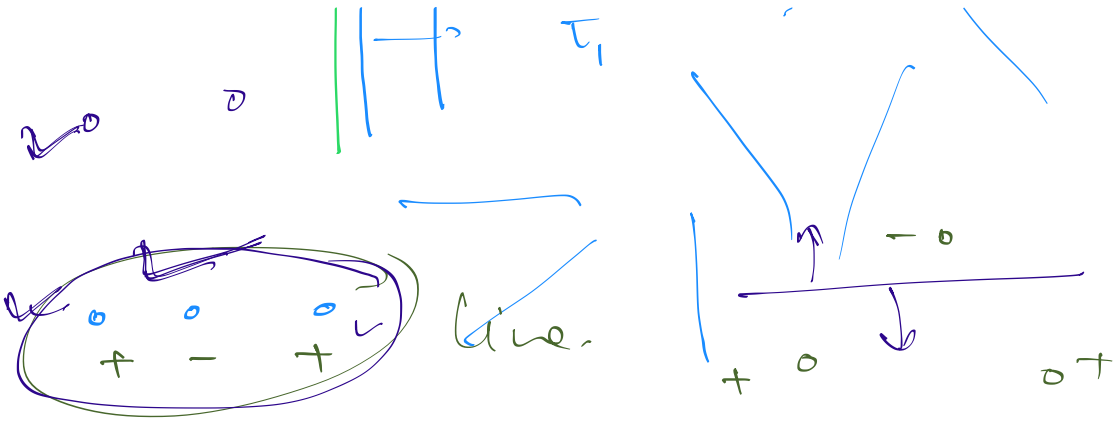
We can shatter 2 points.

3-Samples, 2 classes:

$2 \times 2 \times 2 = 8$



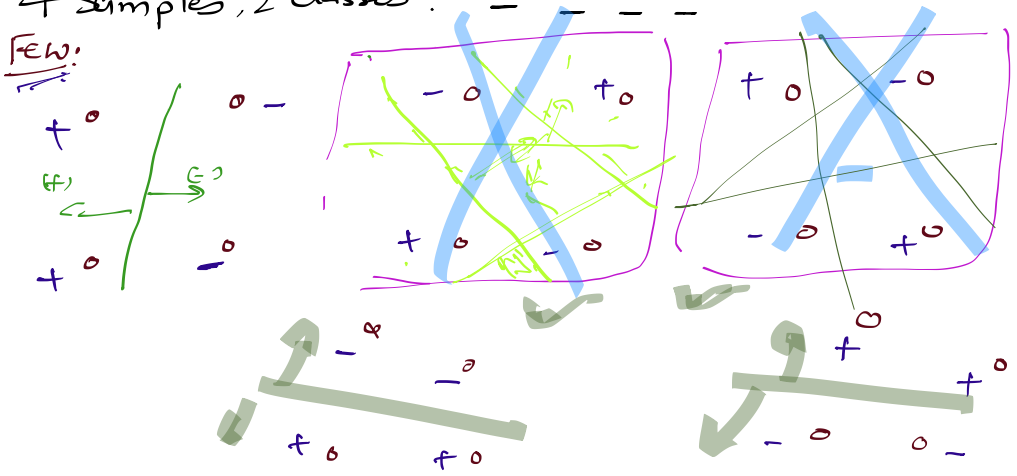




Line can shatter 3 samples.

4 Samples, 2 classes : $2 \times 2 \times 2 \times 2 = 16$ possible labellin.

FEW:



Line can not Shatter 4-points

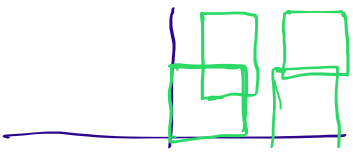
Line can shatter ~~at~~ maximum of 3-points.

VC-dimension (Line (or linear hypothesis in 2D) is 3

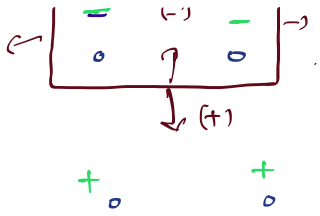
VC-dimension of Linear hyperplane in d-dimension is d+1

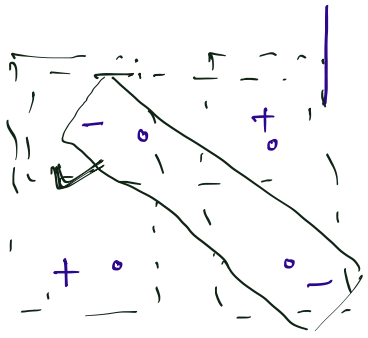
Hypothesis is Axis-aligned Rectange (2D)

4-Simple class: 2



3





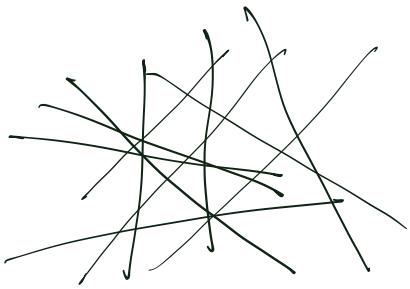
← not axis aligned

Rectangle in 2D → VC-dim = (4)

H = Line in 2D →

$y = mx + c$

$w^T x_i + b = 0$



- b: bias (Distance of the plane from origin of the Vector space)
- w: weight (Vector perpendicular to the plane having x_i)

In a d-dimension — i^{th} sample x_i

$VC(h) \propto$ # parameters

$$x_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^d \end{bmatrix}_{d\times 1} ; \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} ; \quad b \in \mathbb{R}$$

$$h_{w,b}(x_i) = w^T x_i + b$$

$$h(x_i) = w^T x_i + b = 0$$

$x_i \in \mathbb{R}^d$ $\quad w \in \mathbb{R}^d$

$VC(h) = d+1$

Linear model in d-dimensions.

---

m: # samples

$m \geq O\left(\frac{VC(H)}{\epsilon}\right)$

$m \geq O\left(VC(H)\log VC(H)\right)$

$m \geq O\left(VC(H)^2\right)$

Structure

ϵ: Tuned / generalized

$m \geq O\left(\frac{1}{\delta}\right)$ $\quad \tilde{} \; m \geq O\left(\log\frac{1}{\delta}\right)$

$m \geq O\left(\frac{1}{\epsilon}\right)$

Zero Error

$m \geq O\left(\frac{1}{\epsilon^2}\right)$

$m \geq O\left(\ln|H|\right)$

Error / Tolerance

(i) $\epsilon \propto \frac{\ln|H| + \ln\left(\frac{1}{\delta}\right)}{m}$ $\epsilon \propto (1-\delta)$ Prob Conf

(ii) $\epsilon = \sqrt{\frac{\ln|H| + \ln\left(\frac{1}{\delta}\right)}{2m}}$

(iii) $\epsilon = \sqrt{\frac{c_1 \cdot VC(H) + c_2 \ln\left(\frac{c_3}{\delta}\right)}{m}}$

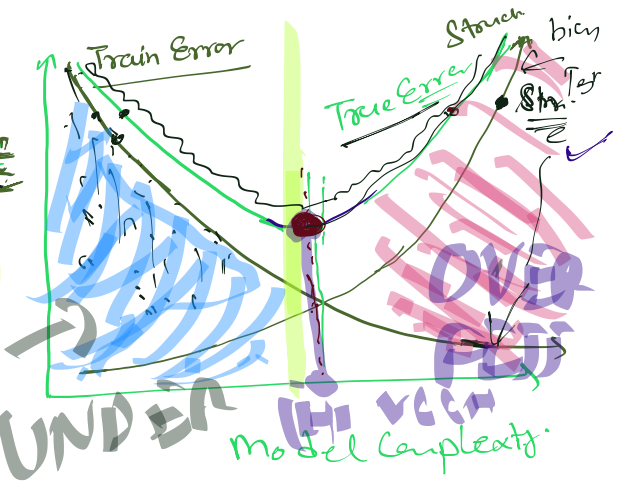
$m \downarrow$ $\delta \downarrow$ $(1-\delta)$ probability ↑

Relative Error + empirical error + Structural error ϵ

True Error $\leq$ Train Error + ϵ

True is Error is high

Train Error is low



overfitting (Variance) VC(H) a |H|

Train Erra → high : Underfitting. (bias problem)

Good hypothesis : Low overfittin Low Underfitt.

high Variance → overfitting.

high bias → Underfitting

Bias - Variance Trade off

Model : low bias & low variance is a good

→ Model is suffering from overfitting (variance) issue:

Remedies: Reducing the model complexity
- ↳ Regularization
- ↳ Dropout
- ↳ Change to less complex model

(ϵ → reduces)

$\epsilon \propto \ln(\frac{1}{\delta})$

- Increase the # samples in the training
  - ↳ Augmentation (Chen method)
- NN: (early stopping) → High Confidence
- "δ": → as $(1-\delta)$ reduced.

$$P(\text{True error} \leq \epsilon) > (1-\delta)$$

Conf.

(True = 1)

$\delta = 0.01$

$1-\delta = .99$

$\delta = 0.05$

$1-\delta = .95$

=1

Confidence interval, p-value, Significance, Null hypothesis

→ Distribution Shift:

Training data, Validation data, Test data

If there is a mismatch (all these three does not follow same distribution: distribution shift)

↳ Collecting data from different sources

↳ Data splitting during training

↳ Cross-validation (

Small-data: Leave-one-out (LOOCV)

medium data: 80-20, 70-30, 60-40, (Train, validation) + K-fold Cross-Validation

Large dataset: 1 to 10% data for validation

↳ low K-fold (K = 3 to 5)

↳ Random sampling for fixing the hyper-parameters.

Hyper-parameters: Set by the user.

(Cross-validation, Splitting)

Validation

Step 1: Split data into 2-parts → Training | Testing

Step 2: Training data → Train + Validation

Step 3: Fit the model w hyperparameters with Good validation accuracy. (CV, splitting)

Step 4: Fixed Hyperparameters in the previous step

Train the model on Training Data and

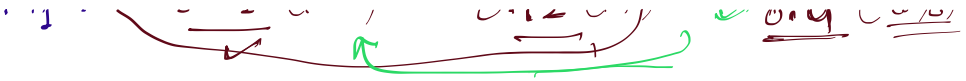
Same steps

Evaluate model on Test Data

Increase the validation % size

Real T.

	Train error	Validation	Test error
M_1 :	0.1 (10%)	0.12 (12%)	0.11 (11%)



Distribution Shift is found:

- Try to Increase the % age of Validation

Normalized of feature

Random Sample the samples from Test Distribution and Try to fine tune the large model on this random samples. (If you are not able to retrain

Otherwise Augment the Training Data with random Samples from Test distribution and retrain the model and fix the hyper-parameter using validation data

→ Model is suffering from Underfitting (bias)

Issue: (Training error is high)

↳ Increase the model Complexity

↳ Somehow try to increase the VC-dim
↳ Introduces non-linearity

↳ Change to complex model.

↳ Increase the epochs / iterations

↳ Look in to the Data

↳ Preprocessing (lossy techniques) Reduce

↳ Enhancement

Biased Data

Hard

Winter

Blurry Image

Example:

	Training Error	Validation Error	Test Error	Problem ??
	#s		#s	
M_1	10%	9%	11%	Low bias, low var, No dist shift
M_2	8%	20%	18%	Low bias, High variance, ??
M_3	30%	40%	45%	High bias, ?? ??
→ M_4	13%	14%	20%	Low bias, low variance, Dist.

≈ 3.5%

overfitting

→ . These remedies doesn't guarantee the perfect model :

. In practice It works

. Moderate size data is needed ≥ 10 k samples

$\epsilon \propto O\left(\frac{1}{m}\right)$: Problem.

. Experience Matters (Expert)

+ Comparing the models: Statistical Test are required

errors: $(\bar{\mu}_{m_1} \pm CI_{m_1}), (\mu_{m_2} \pm CI_{m_2})$ - - - - - ( )

Model with low mean error and low Confidence Interval of an error is a GOOD model

i^{th} model: Repeatable training 'n' times with different seeds:

$$\mu_{m_i} = \frac{1}{n} \cdot \sum_{j=1}^{n} \text{valid/test}_j(e_j)$$

$$CI_{m_i} = \frac{\sigma^2}{\sqrt{n}}$$

σ^2 ← Variance

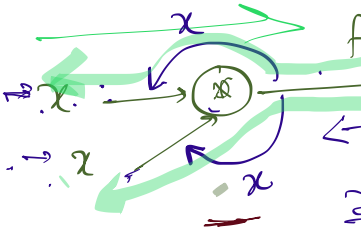
e_1 ← 1
e_2 ← 2
e_n ← n
Testers

→ Computation Graph: Pictorial Rep

→ Pictorial Representation of a Computation

$\Rightarrow f(x) = x^2 = x * x$

$f(x,y) = x * y$



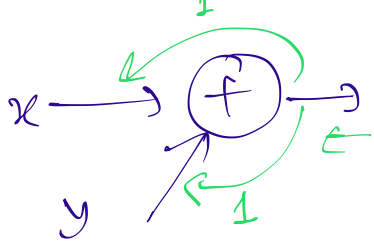
$\frac{\partial f}{\partial x} = 1 \times x$

$\frac{\partial f}{\partial x} = x + x = 2x$

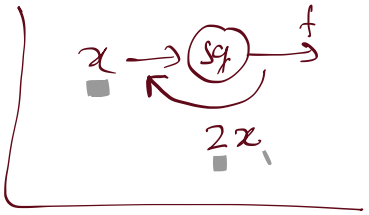
$\frac{\partial f}{\partial x} = x \cdot \frac{\partial y}{\partial x} + \frac{\partial x}{\partial x} \cdot y = y$

$\frac{\partial f}{\partial y} = x$

- $f(x,y) = x+y$



$\leftarrow \frac{\partial f}{\partial f} = 1$

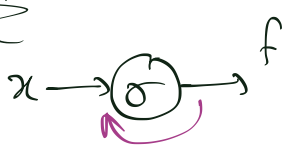


↳ Chain Rule

$f(x) = \log(x)$

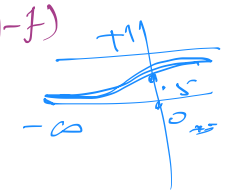
$\frac{1}{x} \leftarrow 1/\text{input}$ $\quad \frac{\partial f}{\partial f} = 1$

$$\sigma(x) = \frac{1}{1+e^{-x}} = \cancel{\sigma(x)\cdot(}$$



$$\frac{\partial\, \sigma(x)}{\partial x} = \sigma(x)\,(1-\sigma(x))$$

$f\cdot(1-f)$



$$f(x) = \frac{1}{x} \checkmark$$

$$f(x,y) = x \div y \ (\text{or } x/y) \Rightarrow$$

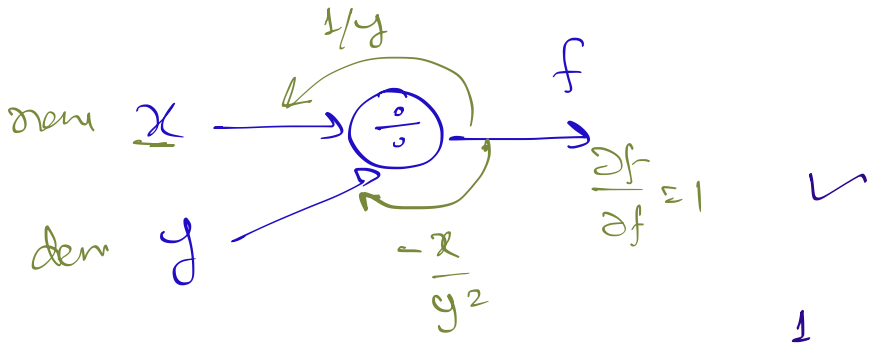
$$\frac{\partial f}{\partial x} \cdot \frac{\partial x}{\partial x} = \frac{1}{y} \Rightarrow \frac{1}{y}$$

nom / den

$$\frac{\partial f}{\partial \text{ nominator}} = \frac{1}{\text{denominator}}$$

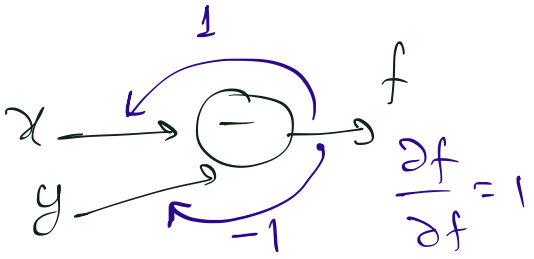
$$\frac{\partial f}{\partial \text{den.}} = \frac{-\text{nominator}}{(\text{denominator})^2} \checkmark$$

$$\frac{\partial f}{\partial y} = \frac{y \cdot \frac{\partial x}{\partial y} - x \cdot \frac{\partial y}{\partial y}}{(y)^2} = \frac{-x}{y^2} \checkmark$$

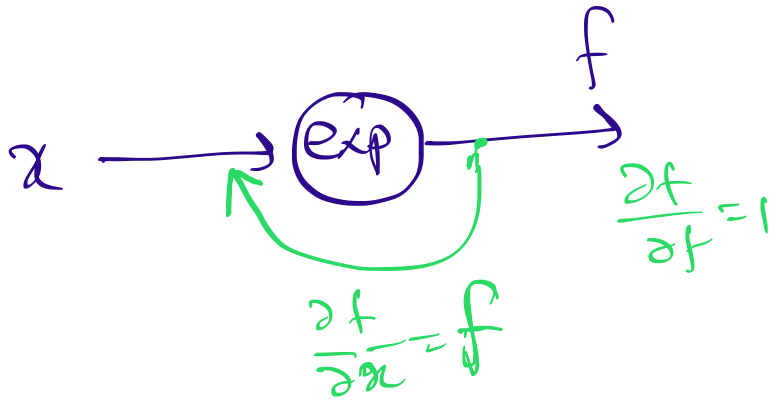


$f(x,y) = x - y$

$\frac{\partial f}{\partial x} = 1 \; ; \; \frac{\partial f}{\partial y} = -1$



$f(x) = e^x$; $\frac{\partial f}{\partial x} = e^x$



$f(x) = \sin(x)$

$\frac{\partial f}{\partial x} = \cos(x)$

$f'(x) = \cos(x)$

$\frac{\partial f}{\partial x} = -\sin(x)$

$f(x) = \tan(x) = \frac{\sin(x)}{\cos(x)}$

$\frac{\partial f}{\partial x} = \sec^2(x) = 1 + \tan^2 x$

$= \frac{\cos(x) \cdot \frac{\partial \sin(x)}{\partial x} - \sin(x) \cdot \frac{\partial \cos(x)}{\partial x}}{(\cos(x))^2}$

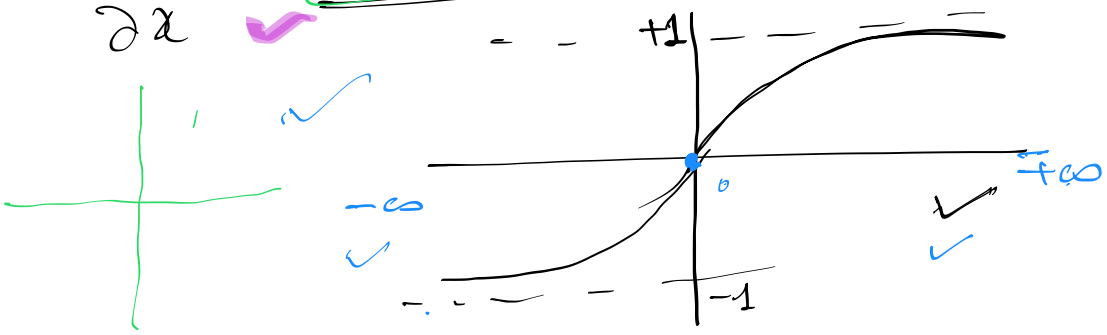
$\Leftarrow = \frac{\cos^2(x) + \sin^2(x)}{\cos^2 x}$

$\frac{\cos^2 x}{\cos^2 x} + \frac{\sin^2(x)}{\cos^2 x}$

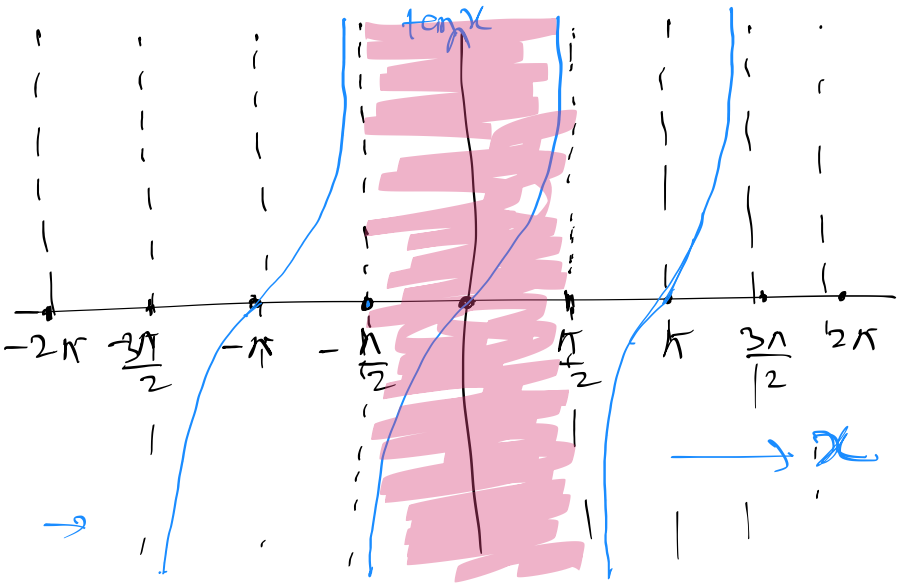
$f(x) = \tanh(x)$ (hyperbolic tangent)

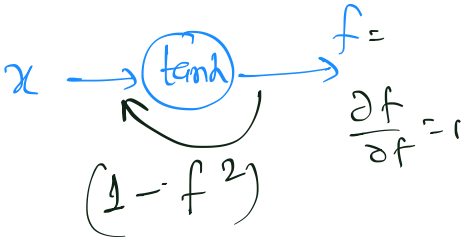
$$\frac{\partial f}{\partial x} = 1 - \tanh^2(x)$$

$$\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$



tan(x)





Rem $f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$

$\frac{\partial f}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ \end{cases}$

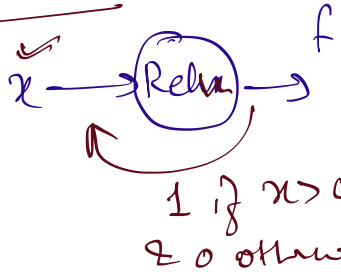
$\frac{\partial x}{} = \{1 \text{ otherwise}\}$

1 if inp > 0
Otherwise '0'

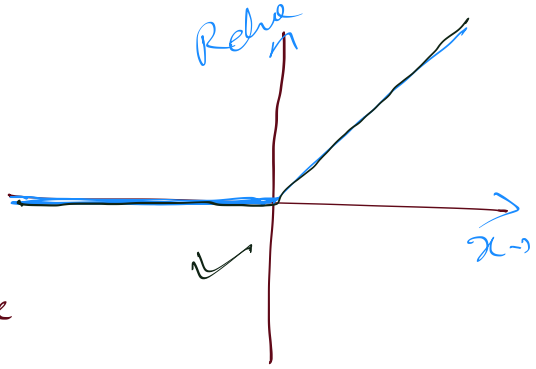
$f(x) = \max(x, 0) = \begin{cases} x, \text{if } x > 0 \\ 0 \text{ otherwise} \end{cases}$

Relu(x) =>

Rectified Linear Unit



1 if $x > 0$
& 0 otherwise



$$\sigma(x) = \frac{1}{1+e^{-x}} \; ; \quad \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

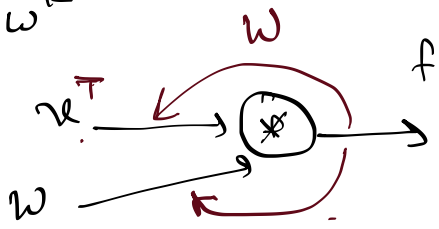
Exercise: $\tanh(x) = 2\sigma(2x) - 1$

Vectors v, w $\quad v \in \mathbb{R}^d, \; w \in \mathbb{R}^d$

$$f(v, w) = v^T \cdot w \qquad \begin{bmatrix} v^T \\ \hline \end{bmatrix} \begin{bmatrix} w \\ \vdots \end{bmatrix} =$$

$$= v^1 \cdot w^1 + v^2 \cdot w^2 + \dots v^d \cdot w^d$$

$$= \sum_{k=1}^{d} v^k \cdot w^k$$



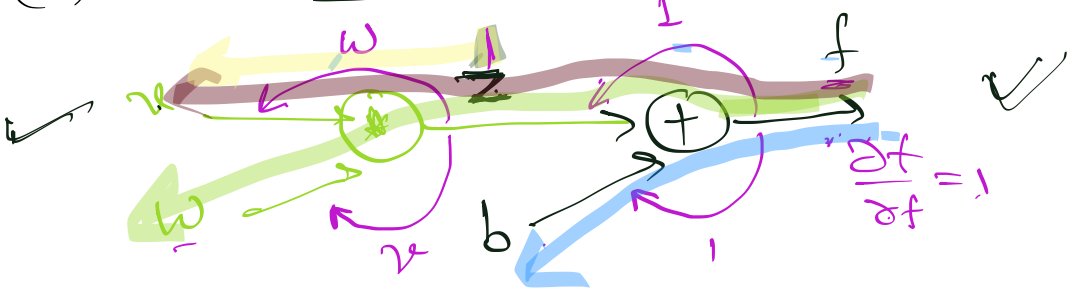
v^T

$v, w \in \mathbb{R}^d$

$b \in \mathbb{R}^1$

$v = \begin{bmatrix} v^1 \\ v^2 \\ \vdots \\ v^d \end{bmatrix}$; $w = \begin{bmatrix} w^1 \\ w^2 \\ \vdots \\ w^d \end{bmatrix}$

$f(v, w, b) = v^T w + b$



$$\frac{\partial f}{\partial w} = 1 \times 1 \times v = v$$

$$\frac{\partial f}{\partial v} = 1 \times 1 \times w = w$$

$$\frac{\partial f}{\partial b} = 1 \times 1 = 1$$

$$\frac{\partial (v \cdot w)}{\partial v} = w$$

$$\frac{\partial^2 (wv+b)}{\partial w} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial w} = 1 \times 1 \times v = v$$

$$\frac{\partial (wv+b)}{\partial v} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial v} = 1 \times 1 \times w$$

$$\frac{\partial (wv+b)}{\partial b} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial b} = 1$$

Chain Rule

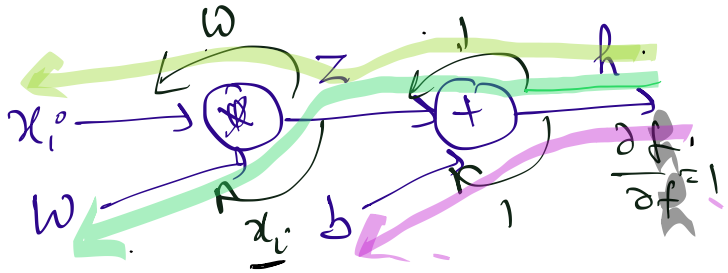
$x_i \in \mathbb{R}^d, \quad w \in \mathbb{R}^d, \quad b \in \mathbb{R}$

$$h(w, x_i, b) = w^T x_i + b$$

$\frac{\partial h}{\partial w} = ? \quad 1 \times 1 \times x_i = x_i$

$\frac{\partial h}{\partial x_i} = ? \quad 1 \times 1 \times w = w$

$\frac{\partial h}{\partial b} = ? \quad 1 \times 1 = 1$



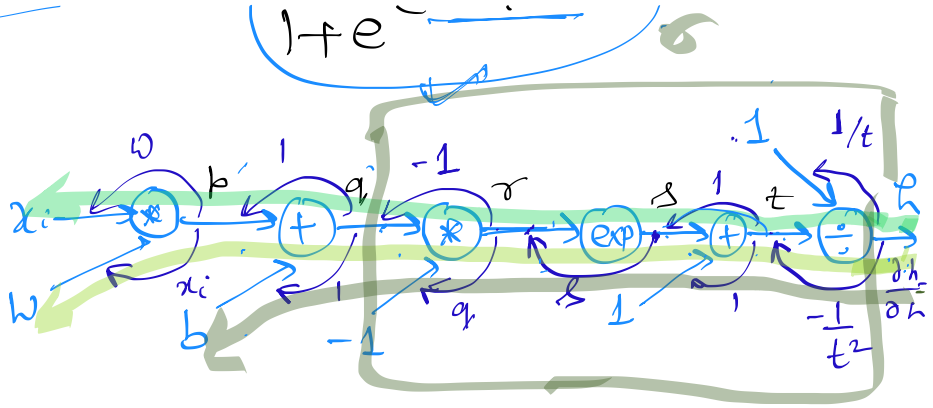
$$h(w, x_i, b) = \frac{1}{-(w^T x_i + b)}$$

$\frac{\partial e^{x_1}}{\partial x} = e^x$

$\frac{\partial h}{\partial x_i} =$

$\frac{\partial h}{\partial w} =$

$\frac{\partial h}{\partial b} =$



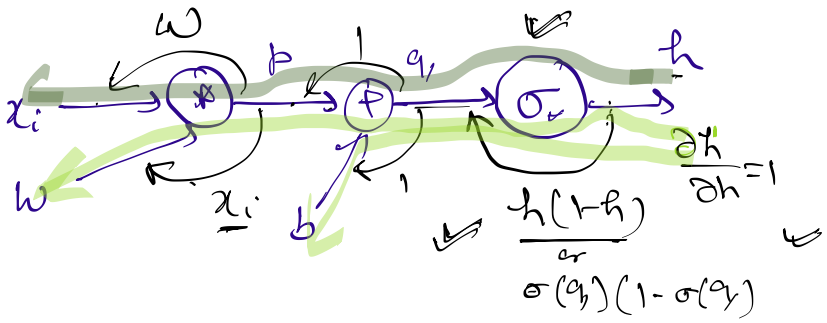
$$\frac{\partial h}{\partial x_i} = 1 * \left(\frac{-1}{t^2}\right) \cdot 1 \times s \times (-1) \times 1 \times w = \frac{s \times w}{t^2}$$

$$\frac{\partial h}{\partial w} = 1 \times \left(\frac{-1}{t^2}\right) \times 1 \times s \times (-1) \times 1 \times x_i = \frac{s\, x_i}{t^2}$$

$$\frac{\partial h}{\partial b} = 1 \times \left(-\frac{1}{t^2}\right) \times 1 \times s \times (-1) \times 1 = \frac{s}{t^2}$$

$$\rightarrow \quad h(w, x_i, b) = \sigma\left(\underbrace{w^T x_i + b}_{z}\right) \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$



$$\frac{\partial h}{\partial x_i} = 1 \times \sigma(q_1).(1-\sigma(q_1)) \times 1 \times w$$

$$\frac{\partial h}{\partial w} = 1 \times \sigma(q_1).(1-\sigma(q_1)) \times 1 \times x_i$$

$$\frac{\partial h}{\partial b} = 1 \times \sigma(q_1)\{1-\sigma(q_1)\} \times 1$$

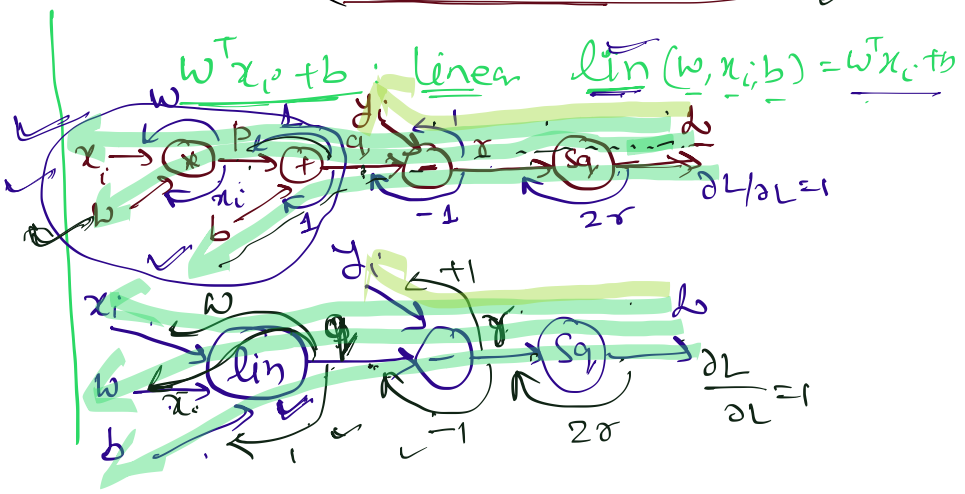
$$L_0(w, x_i, b, y_i) = (y_i - (w^T x_i + b))^2$$

$$\frac{\partial L}{\partial w} =$$

$$\frac{\partial L}{\partial x_i} =$$

$$\frac{\partial L}{\partial b} =$$

$$\frac{\partial L}{\partial y_i} =$$



$$\frac{\partial L}{\partial w} = 1 \times 2r \times (-1) \times (+1) \times x_i$$

$$\frac{\partial L}{\partial x_i} = 1 \times 2r \times (-1) \times 1 \times w$$

$$\frac{\partial L}{\partial b} = 1 \times 2r \times (-1) \times 1$$

$$\frac{\partial L}{\partial y_0} = 1 \times 2r \times 1$$

$\frac{\partial L}{\partial w}$

$1 \times 2r \times (-1) \times x_i$

$\frac{\partial L}{\partial x_i}$

$w \times 1 \times 2r \times (-1)$

$\frac{\partial L}{\partial b}$

$1 \times (2r) \times (-1) \times 1$

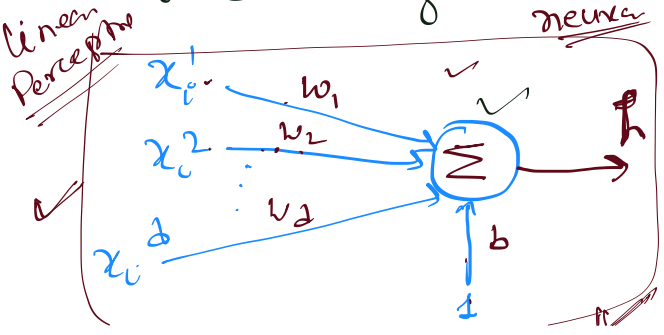
$\frac{\partial L}{\partial y_0}$

$1 \times 2r \times 1$

$$h(w, x_i, b) = w^T x_i + b$$

Linear Regression, loss fn = Mean Square Error.

Linear Perceptron

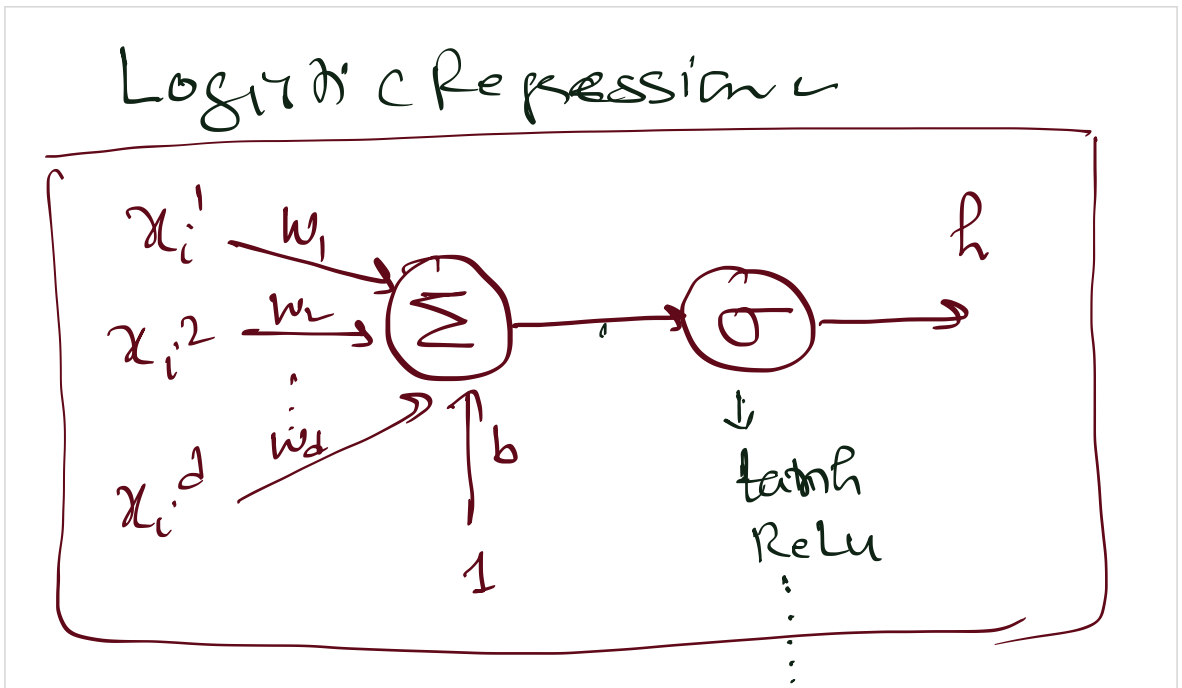


$x_i \in \mathbb{R}^d$, $b \in \mathbb{R}$

$w \in \mathbb{R}^d$

$h = w_1 \cdot x_i^1 + w_2 \cdot x_i^2 + w_d \cdot x_i^d + b \times 1$

$h = w^T x_i + b$


Logistic Regression
x_i^1
w_1
x_i^2
w_2
x_i^d
w_d
Σ
b
1
σ
h
tanh
ReLu

$$\mathcal{L}(w, x_i, b, y_i) = -y_i \log(\sigma(w^T x_i + b)) - (1 - y_i) \cdot \log(1 - \sigma(w^T x_i + b))$$

Binary Cross Entropy

$\frac{\partial \mathcal{L}}{\partial w}$: Exercise

$\frac{\partial \mathcal{L}}{\partial x_i}$ =

$\frac{\partial \mathcal{L}}{\partial b}$ =

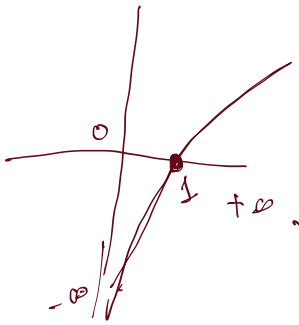
$\frac{\partial \mathcal{L}}{\partial y_i}$ =

$\hat{y}_i = \sigma(w^T x_i + b)$ $\quad h(w, x_i, b)$

Predicted

$h($ $\sigma(wx+b) \in \{0, 1\}$

$y_i = \{0, 1\}$



y_i	$\hat{y}_i$	loss	
0	0	0	$-(1-y_i) \cdot \log(1-\hat{y}_i) \to +\infty$
0	1	→ huge	
1	0	→ huge	$-y_i \log \hat{y}_i \to +\infty$
-1	-1	0	