

Supervised Classification

TRAINING DATABASE: SUPERVISED DATA

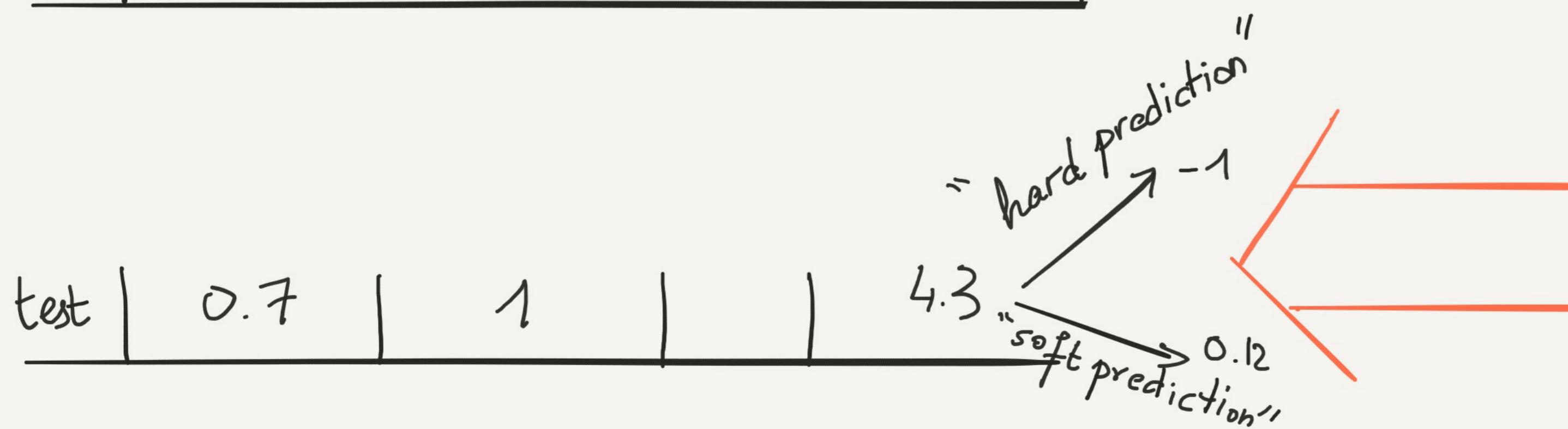
	Feature 1	Feature 2	...	Feature p	Label
obs1	0.8	1		-12.3	1
obs2	0.6	1		5.6	-1
:					:
obs n	0.3	0		8.1	-1

M
A
C
H
I
N
E

M
O
D
E
L

F
I
T
T
I
N
G

L
E
A
R
N
I
N
G



THE USUAL STEPS

RAW DATA

1.2	BLVD
NaN	AVE
126897	ST
0.4	

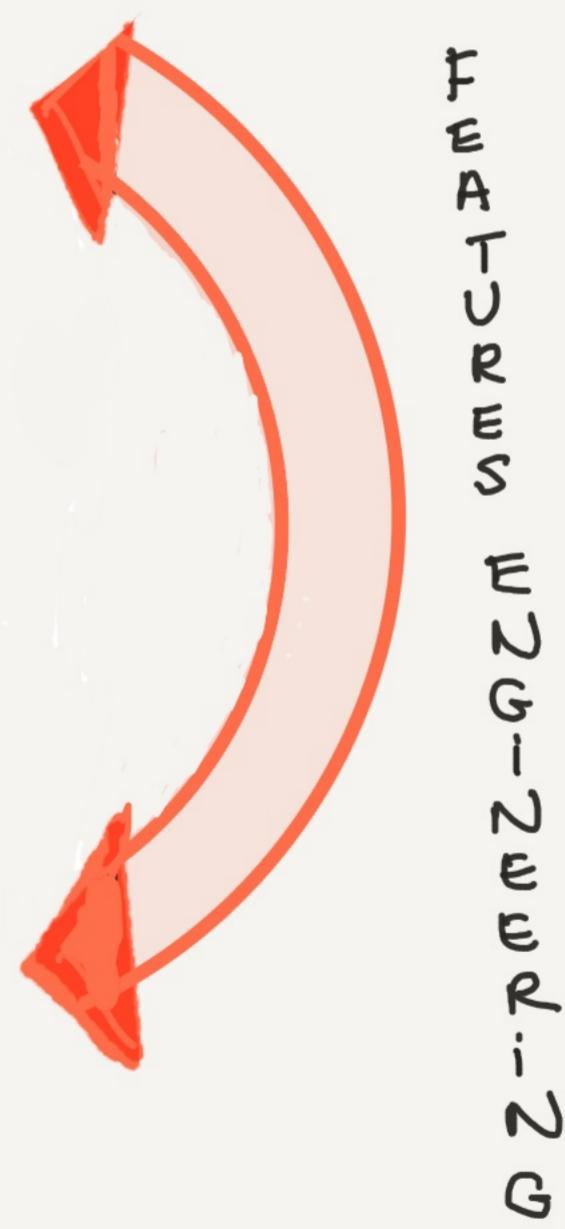
CLEANING

CLEAN DATA

1.2	BLVD
0.8	AVE
0.8	ST
0.4	MISSING

ENGINEERED DATA

1.2	1	0	0	0
0.8	0	1	0	0
0.8	0	0	1	0
0.4	0	0	1	0



REPRESENTATIONS

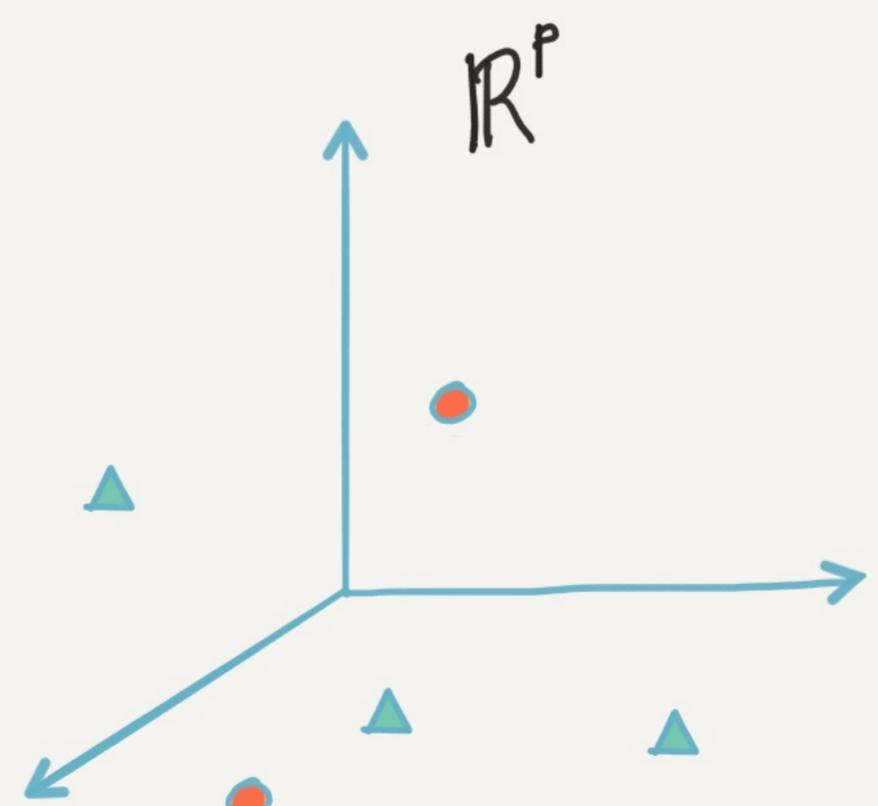
Y

▲	1
●	0
▲	1
▲	1
●	0

X

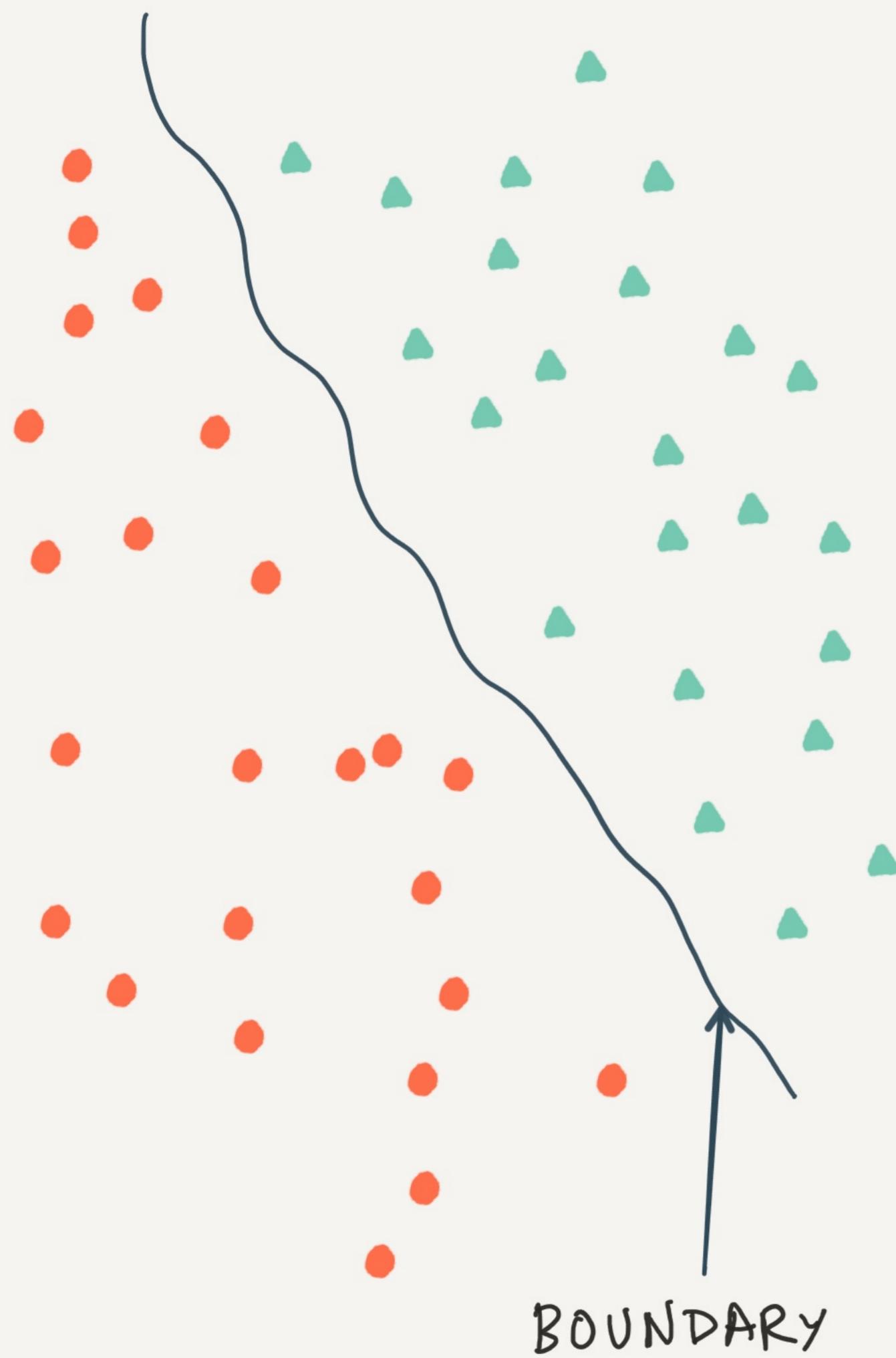
121	1	0	0	12	43
-1	0	1	0	1	1	...	
12.6	0	0	0	12	4		
9.7	0	0	1	12	8		
-1	0	0	0	24	5		

numerical

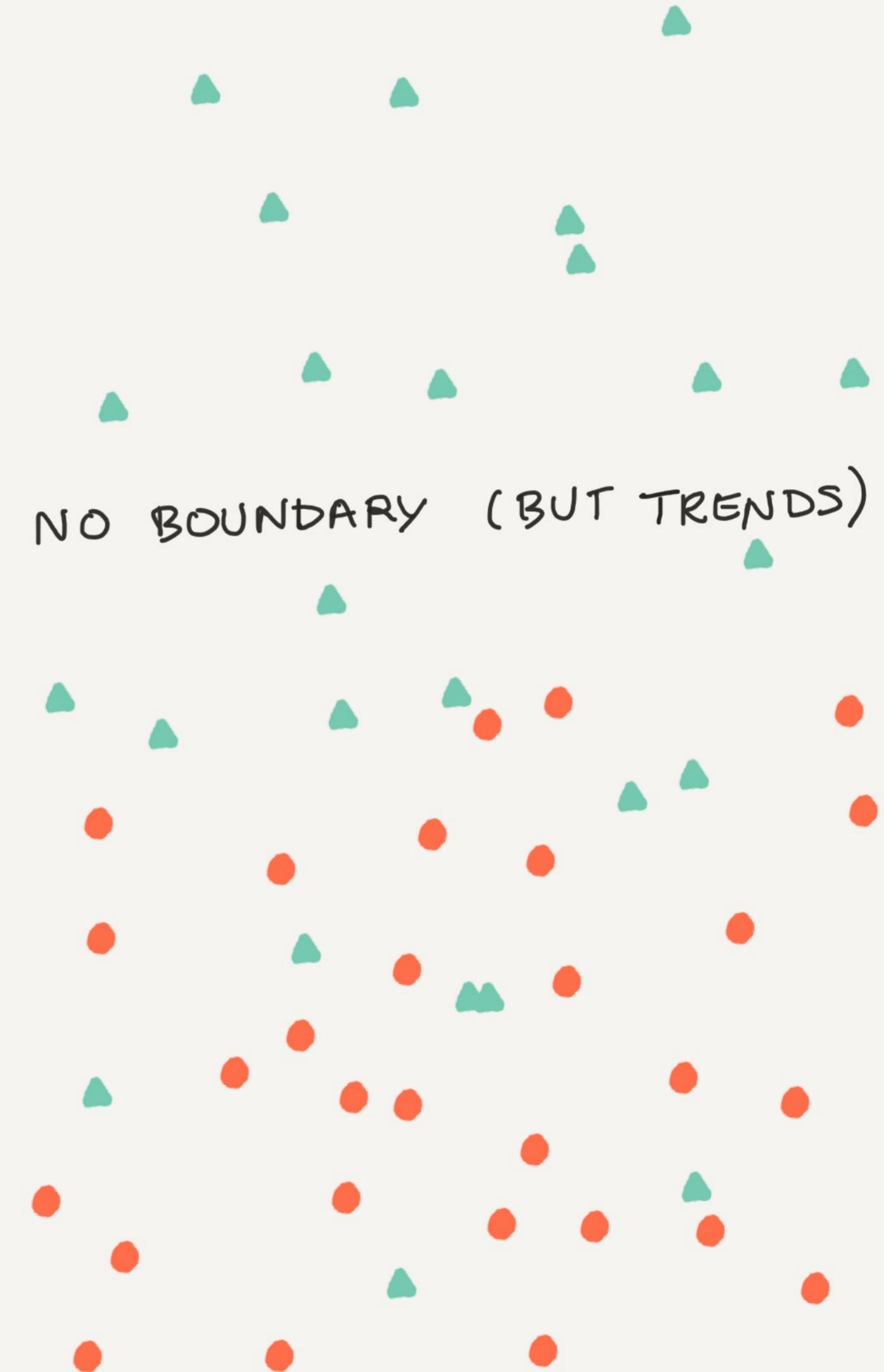


geometrical

IDEAL WORLD



REALITY

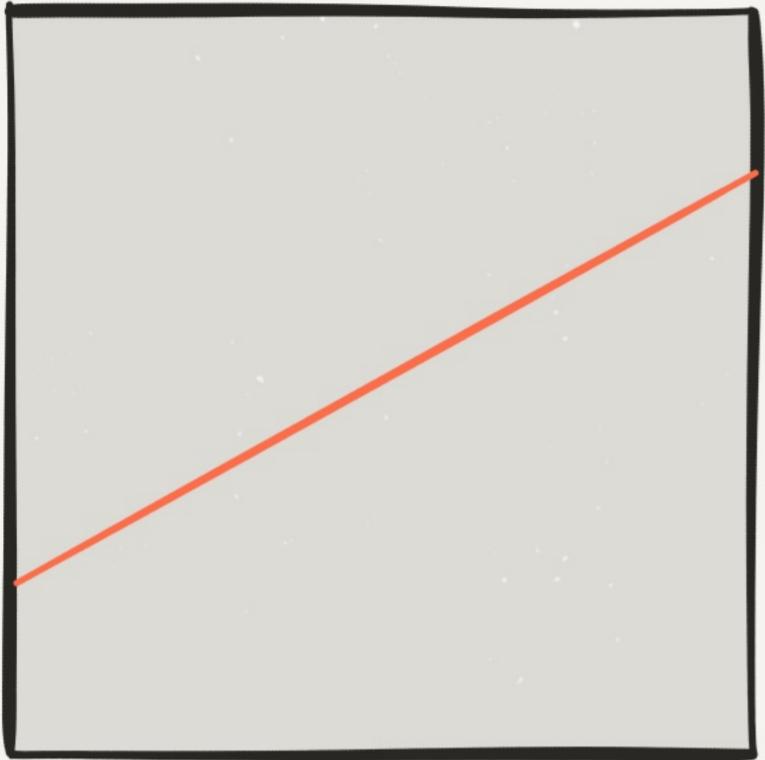


GEOMETRICALLY

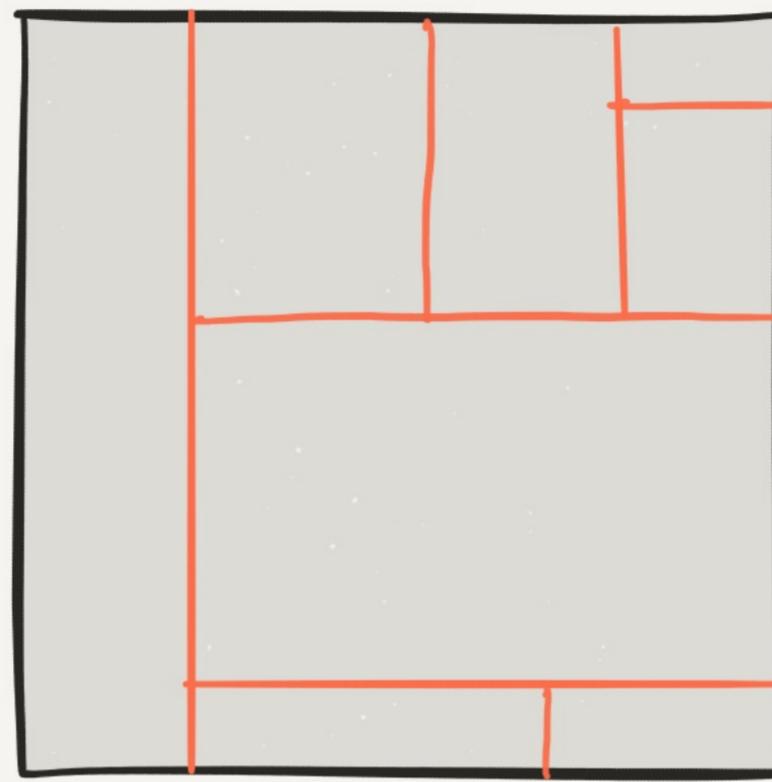
MODELS = PARAMETRIC BOUNDARIES

MODELS

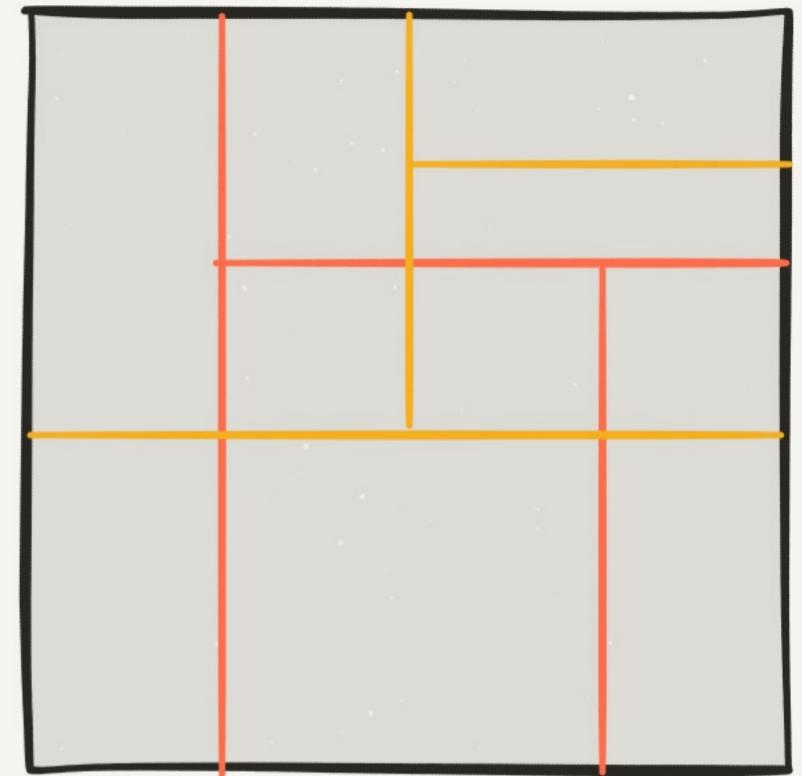
Generalized Linear Models (GLM)



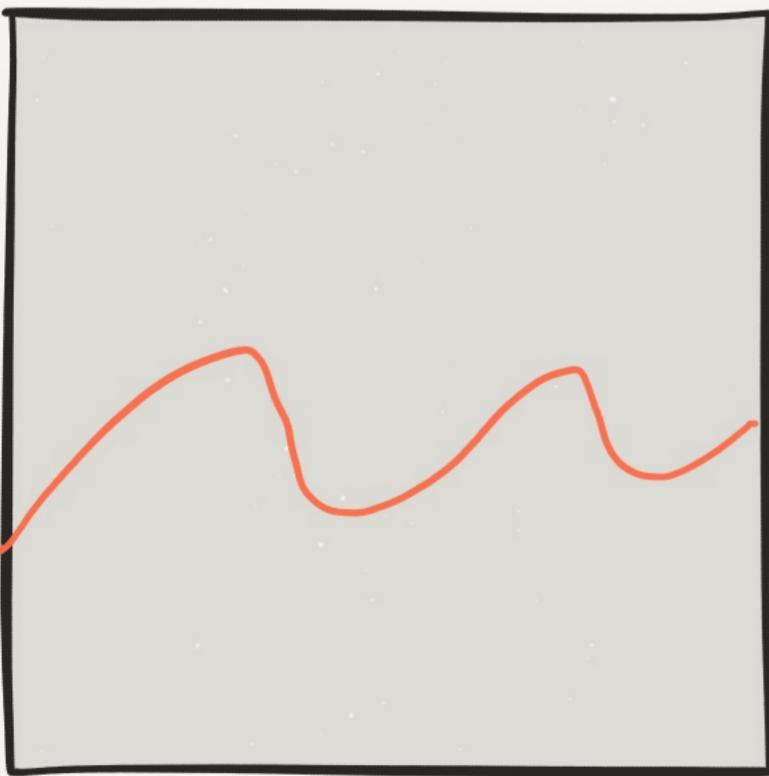
Decision Tree



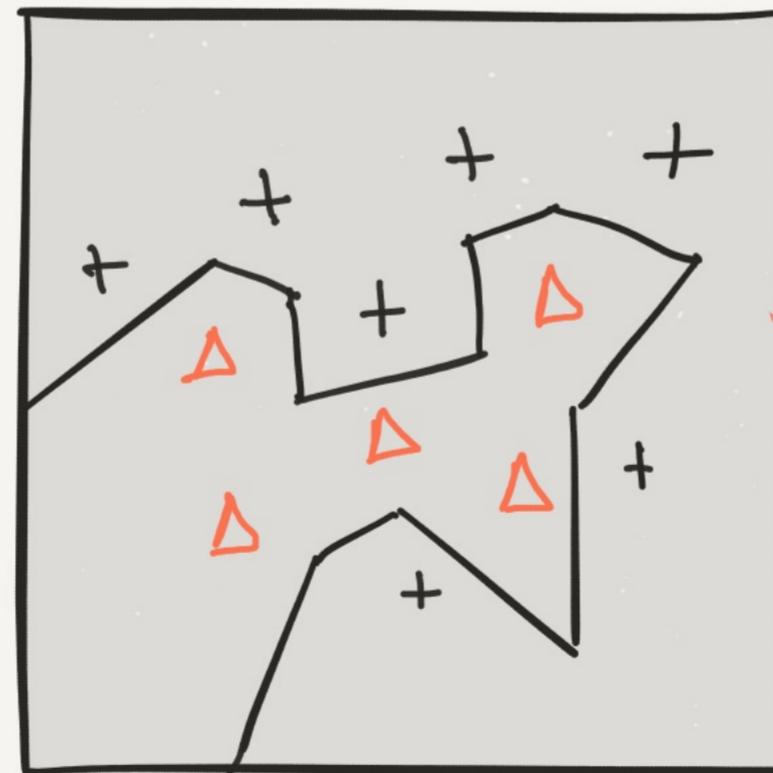
Forest



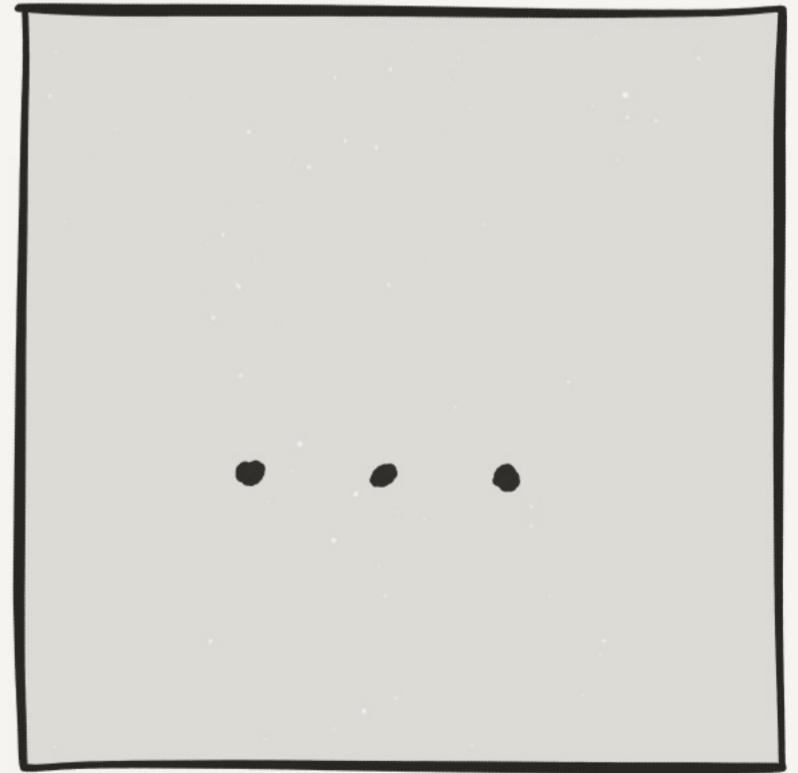
k. SVM



k Nearest Neighbors



etc .



SESSION OUTLINE

- 1) Intro ✓
- 2) Linear Models
- 3) Generalized Linear Models
- 4) Support Vector Machines
- 5) Kernel Methods
- 6) Decision Trees
- 7) Random Forests
- 8) Boosting

LINEAR REGRESSION

1. Problem Statement

Input: n data points (x_i, y_i) , $1 \leq i \leq n$ où $x_i \in \mathcal{X}$ et $y_i \in \mathbb{R} \forall i$

Expected Output: $w, b = \arg \min \sum_{i=1}^n (y_i - \langle x_i, w \rangle - b)^2$

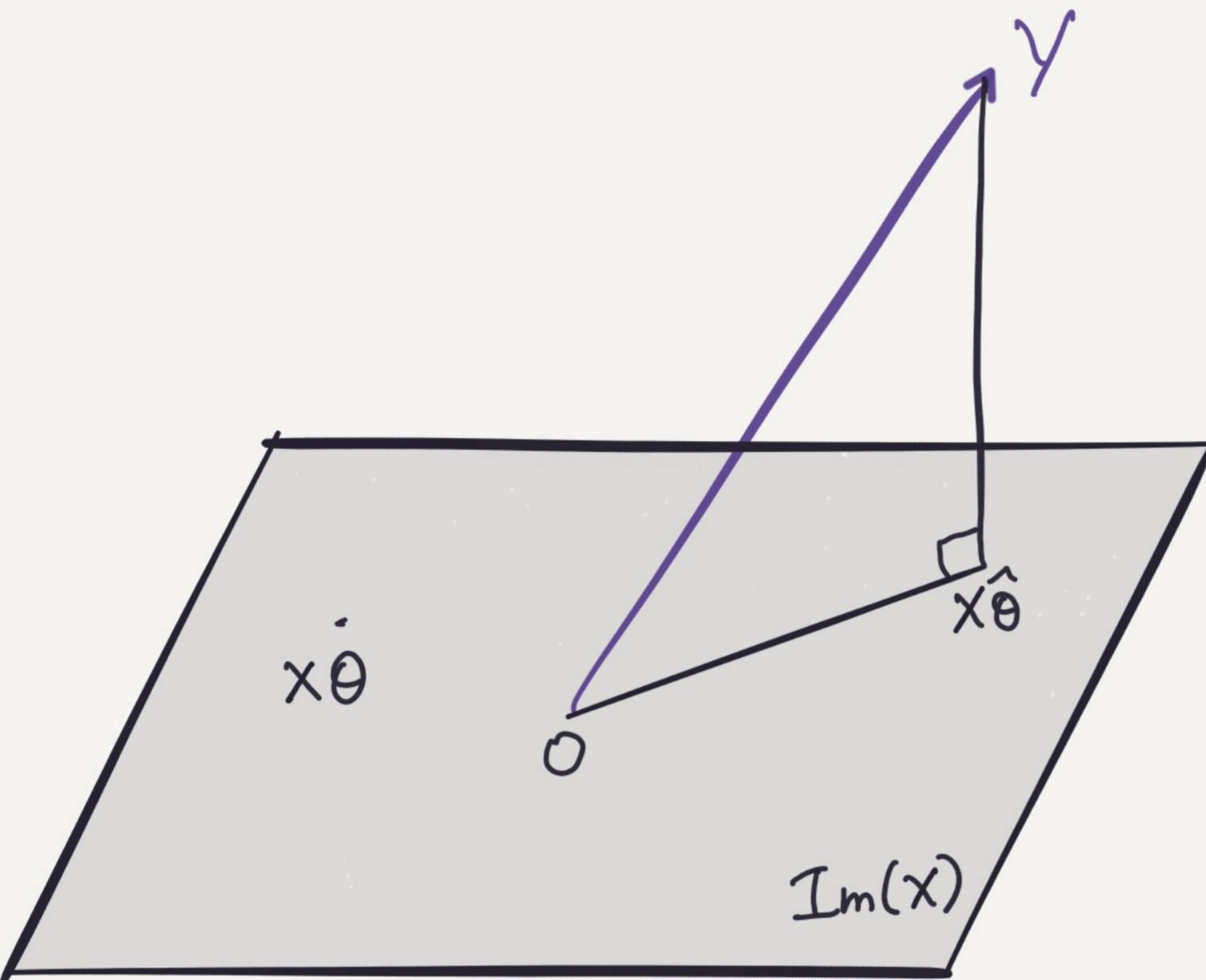
These are Ordinary Least Squares

2. Geometrically

Let $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, $X = \begin{pmatrix} 1^T x_1 \\ 1^T x_n \end{pmatrix}$ and $\theta = \begin{pmatrix} b \\ w \end{pmatrix}$

Problem becomes:

$$\inf_{\theta} \| Y - X\theta \|^2$$



$\hat{x_\theta}$ is orthogonal projection
of y over $\text{Im}(x)$:

$$\langle y - \hat{x_\theta}, x_\theta \rangle = 0 \quad \forall \theta$$

Therefore, $\|y - \hat{x_\theta}\|^2 = \inf_{\theta} \|y - x_\theta\|^2$. Indeed:

$$\|y - x_\theta\|^2 = \|y - \hat{x_\theta} + x(\hat{\theta} - \theta)\|^2 = \|y - \hat{x_\theta}\|^2 + \|x(\hat{\theta} - \theta)\|^2 \geq \|y - \hat{x_\theta}\|^2$$

3. $\hat{\theta}$ computation

From $\langle y - \hat{x}\hat{\theta}, x\theta \rangle = 0 \quad \forall \theta$, we get $\langle {}^t x(y - \hat{x}\hat{\theta}), \theta \rangle = 0 \quad \forall \theta$

Then, for $\theta = {}^t x(y - \hat{x}\hat{\theta})$: $\| {}^t x(y - \hat{x}\hat{\theta}) \| ^2 = 0$

Hence: $\underline{{}^t x(y - \hat{x}\hat{\theta})} = 0$

Giving: ${}^t x x \hat{\theta} = {}^t x y$

Now, if $\text{Ker}(x) = \{0\}$, which is indeed the case if $(x_i)_{1 \leq i \leq n}$ are linearly independent: ${}^t x x$ is invertible, provided $n \geq p+1$ and,

$$\hat{\theta} = ({}^t x x)^{-1} {}^t x y$$

It is a nice closed-form formula. However it is not used in practice even for moderate p because matrix ${}^t x x$ is ill-conditioned, yielding numerical instability.

4. Penalized Linear Models

Sometimes we have an *a priori* over θ and we know that it should belong to some set C . So we can formulate the problem:

$$\inf_{\theta \in C} \|y - X\theta\|^2$$

A mathematically related problem is:

$$\inf_{\theta} \|y - X\theta\|^2 + \lambda J(\theta)$$

where $J(\theta)$ denotes a non-negative function called *penalization function* taking large values where θ should not be a priori.

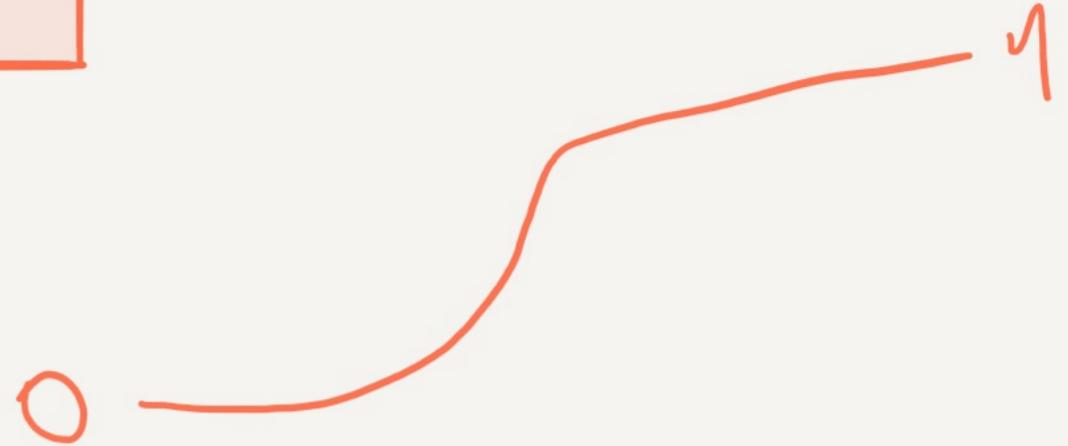
Two penalization functions are mainly used:

— the LASSO function: $J(\theta) = \|\theta\|_1$

— the RIDGE function: $J(\theta) = \|\theta\|_2^2$

To each $\lambda \geq 0$ corresponds a $\hat{\theta}_\lambda$. Function $\lambda \in [0, +\infty[\rightarrow \hat{\theta}_\lambda$ is called regularization path. In the case of Ridge/Lasso, we see $\hat{\theta}_0 = \hat{\theta}$ and $\hat{\theta}_\infty = 0$

LOGISTIC REGRESSION



1. Problem Statement

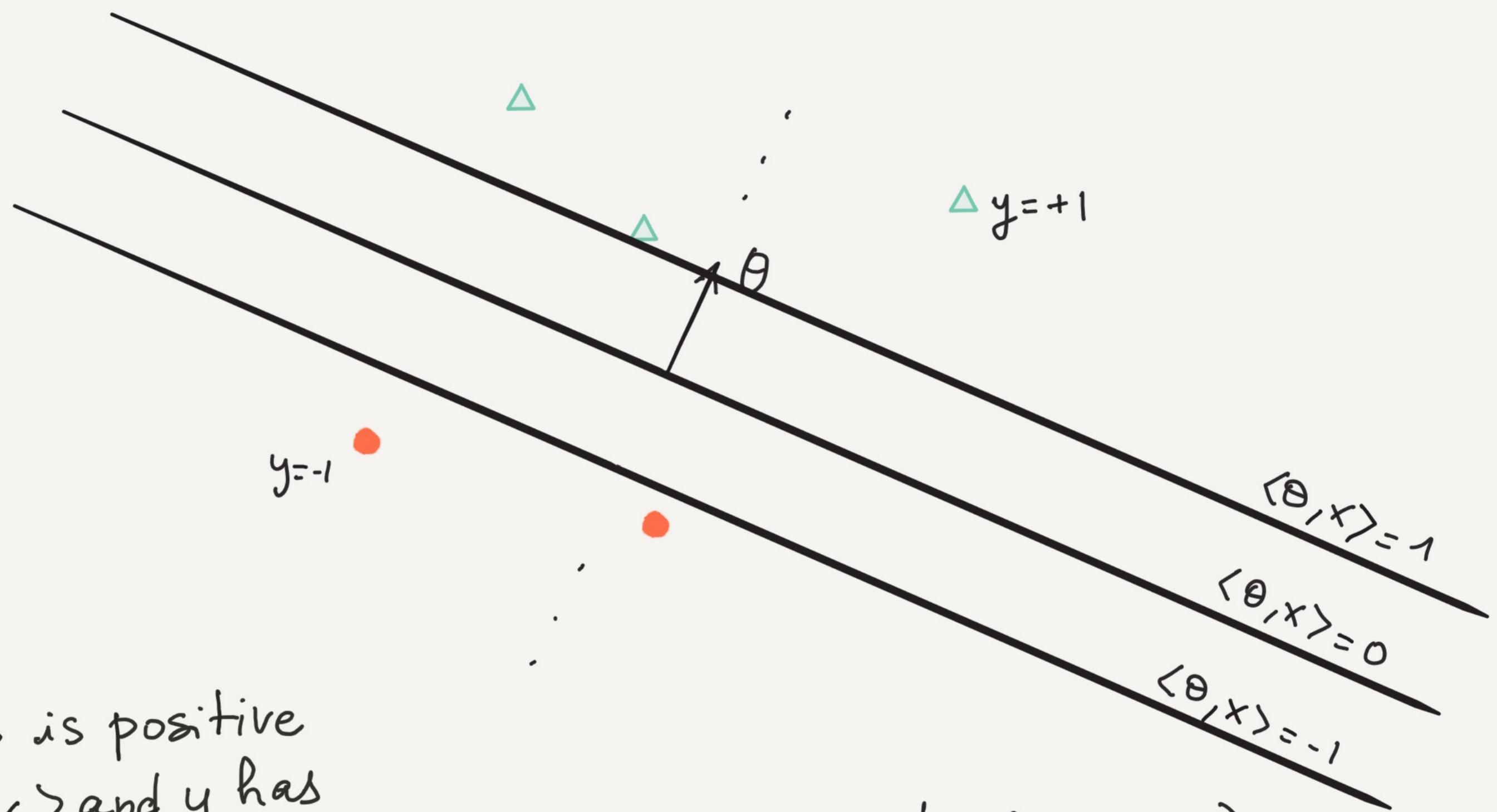
Input: n data points (x_i, y_i) , $1 \leq i \leq n$
with $y_i \in \{-1, +1\}$

$$P_{\theta}[y=1|x] = \frac{1}{1+e^{-\langle \theta, x \rangle}} \Rightarrow \log P_{\theta}[y=1|x] = -\log(1+e^{-\langle \theta, x \rangle})$$

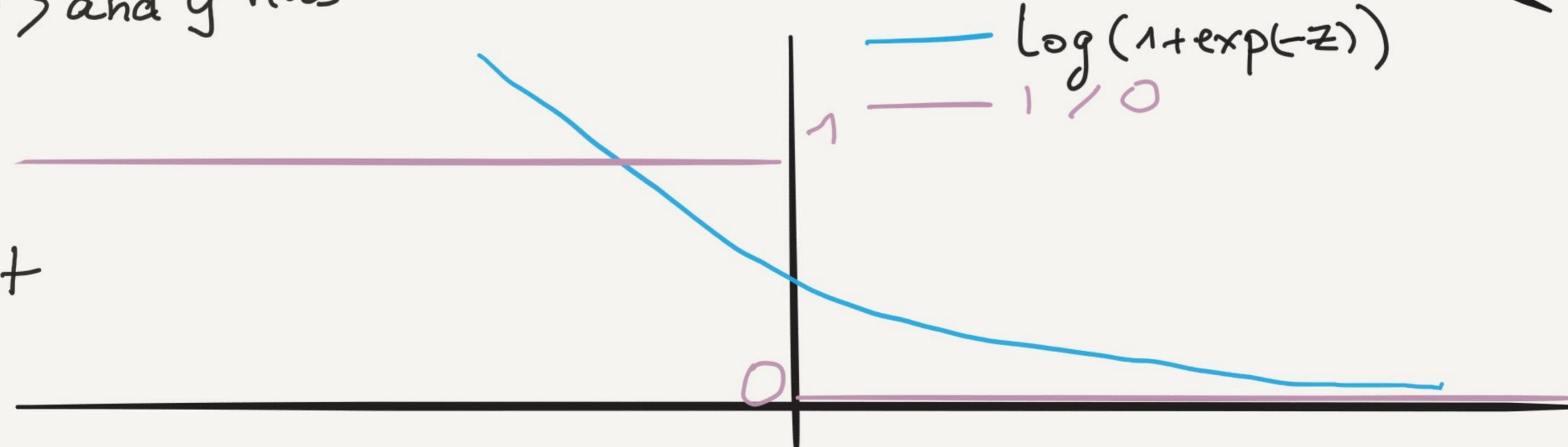
Maximum Likelihood estimator writes:

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \frac{\sum_{i=1}^n \left[y_i \log P_{\theta}[y=1|x](x_i) + (1-y_i) \log (1-P_{\theta}[y=1|x](x_i)) \right]}{} \\ &= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n \log (1+e^{-y_i \langle \theta, x_i \rangle})\end{aligned}$$

2 Geometrical Interpretation



\Rightarrow correctly
labeled point



3 Computation of $\hat{\theta}$

Contrarily to Ordinary Least Squares, there is no closed-form here. So numerical optimisation is performed

We start with:

$$\inf_{\theta} F(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^n f(y_i; \langle x_i, \theta \rangle) \quad \text{where} \quad f(z) = \log(1 + \exp(-z))$$

$$\nabla F(\theta) = \sum_{i=1}^n y_i \varphi'(y_i; \langle x_i, \theta \rangle) x_i$$

$$f'(z) = \frac{-1}{1 + \exp(z)}$$

$$\nabla^2 F(\theta) = \sum p_i (1 - p_i) x_i^T x_i$$

$$f''(z) = \frac{\exp(z)}{(1 + \exp(z))^2}$$

Gradient Descent gives:

$$\theta_{n+1} = \theta_n + \gamma \sum_{i=1}^n \left[\left(\frac{1+y_i}{2} \right) - p_i \right] x_i$$

Newton-Raphson:

$$\theta_{n+1} = \theta_n - \nabla^2 F(\theta_n)^{-1} (\nabla F(\theta_n))$$

4. Penalized Logistic regression

Let us replace

with

$$\inf_{\theta} \sum_{i=1}^n \log (1 + \exp (-y_i \langle x_i, \theta \rangle))$$

$$\inf_{\theta} \sum_{i=1}^n \log (1 + \exp (-y_i \langle x_i, \theta \rangle)) + \lambda J(\theta)$$

For example:

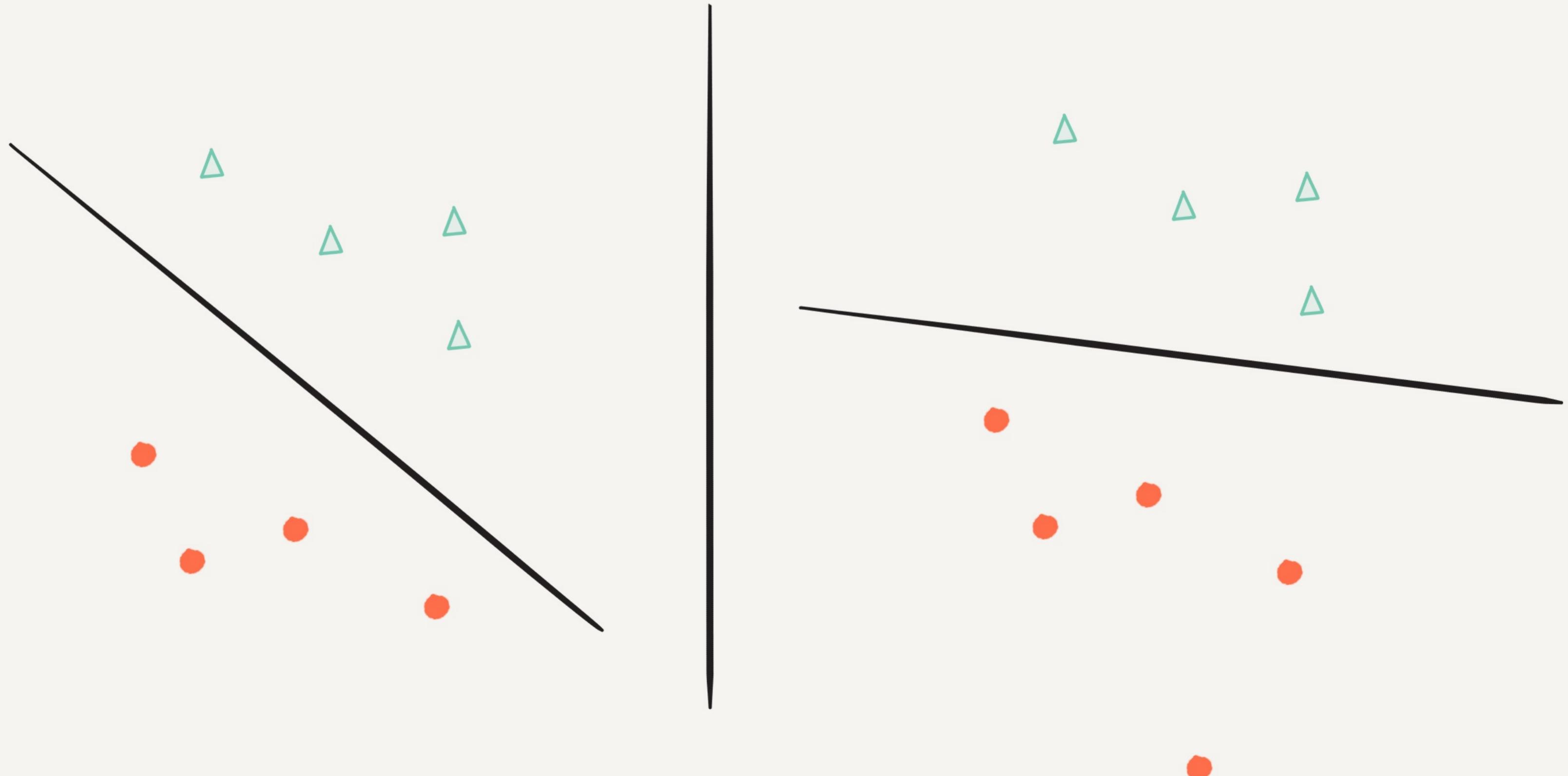
$$* J(\theta) = \|\theta\|_1$$

$$* J(\theta) = \frac{1}{2} \|\theta\|_2^2$$

SUPPORT VECTOR MACHINES

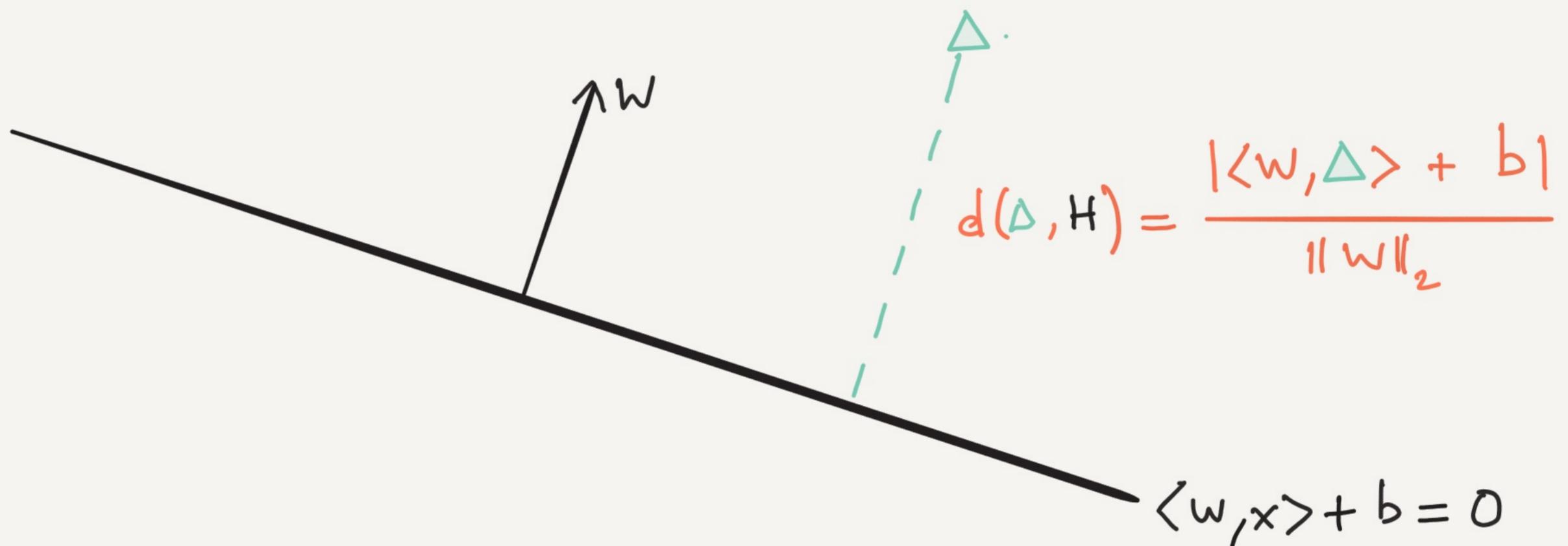
Towards the notion of "margin"

Logistic Regression gives :



2 Geometrical viewpoint

Linear Separation: $H = \{x \in X : \langle w, x \rangle + b = 0\}$



$$\text{Margin} = \min_{1 \leq i \leq n} \frac{y_i (\langle w, x_i \rangle + b)}{\|w\|_2}$$

Maximum Margin:

$$\sup_{w, b} \min_i \frac{y_i (\langle w, x_i \rangle + b)}{\|w\|_2}$$

3. convexification

$$\sup_{w,b} \min_i \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|_2}$$

is a difficult problem to solve, invariant to $(w, b) \rightarrow (\lambda w, \lambda b)$, $\lambda > 0$

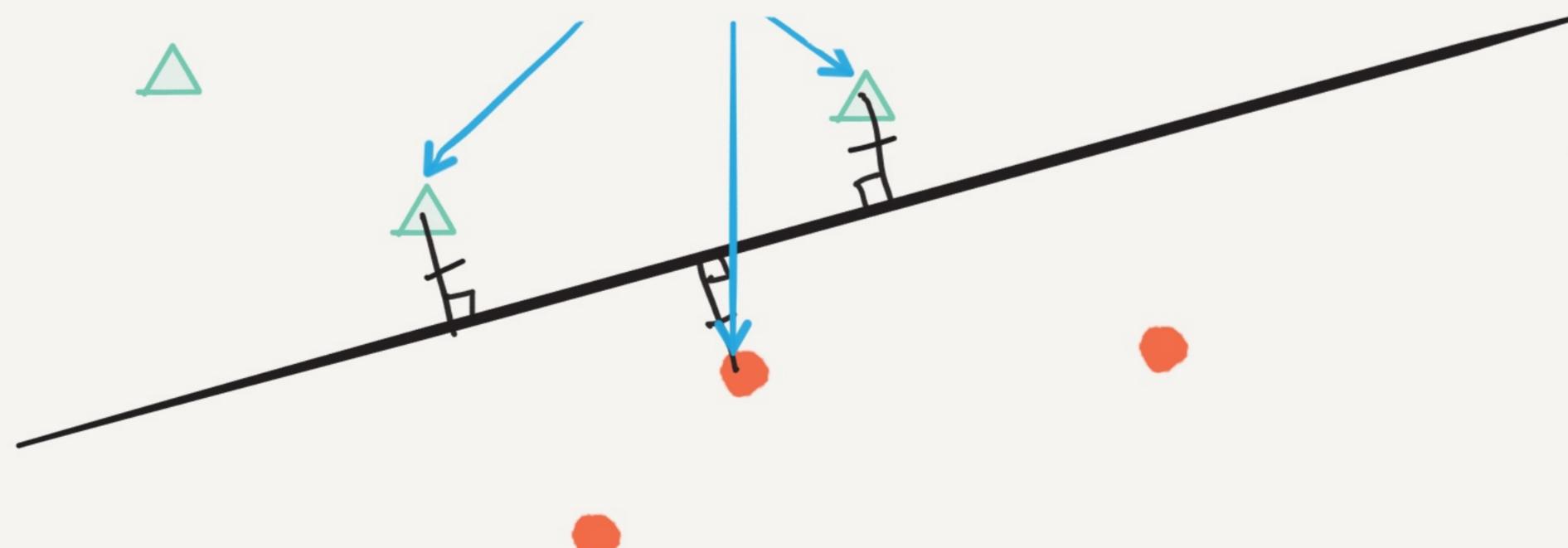
Assume $(x_i, y_i)_{1 \leq i \leq n}$ are **linearly separable**:

$$\min_i y_i(\langle w, x_i \rangle + b) / \|w\|_2 > 0.$$

We can always reparametrize to have: $y_i(\langle w, x_i \rangle + b) \geq 1$

points (x_i, y_i) such that $y_i(\langle w, x_i \rangle + b) = 1$ are called

support vectors



Which gives:

$$\sup_{w, b} \frac{1}{\|w\|_2} \\ \text{s.t. } \forall i: y_i(\langle w, x_i \rangle + b) \geq 1$$

And, finally, the convex problem:

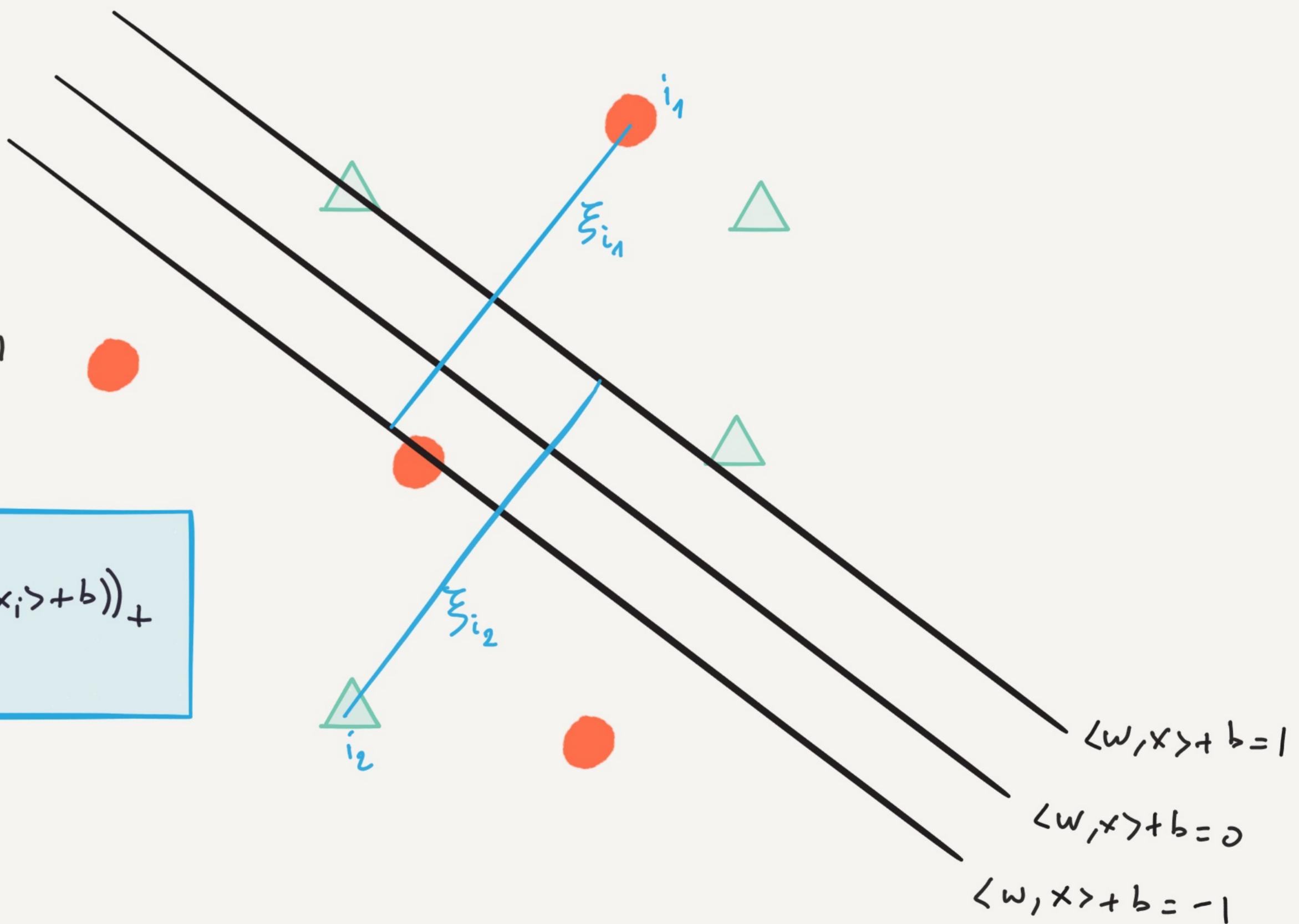
$$\inf \frac{1}{2} \|w\|_2^2 \\ \text{s.t. } \forall i: y_i(\langle w, x_i \rangle + b) \geq 1$$

4. Non linearly separable case

We use what is called **Slack Variables** to penalize each misclassified point with a price $C\xi_i$ (C is a parameter)

Optimization Problem
becomes:

$$\inf \frac{1}{2} \|\omega\|_2^2 + C \sum_{i=1}^n (1 - y_i(\langle \omega, x_i \rangle + b))_+$$



5. Formal Comparison with Logistic Regression

SVM

$$\inf_{w, b} \sum_{i=1}^n (1 - y_i \langle x_i, w \rangle - y_i b)_+ + \lambda \frac{\|w\|^2}{2}$$

RIDGE LOGISTIC REGRESSION

$$\inf_{w, b} \sum_{i=1}^n \log \left[1 + \exp(-y_i(\langle w, x_i \rangle + b)) \right] + \lambda \frac{\|w\|^2}{2}$$

Comparons alors

$$(1-z)_+$$

$$\log(1 + \exp(-z))$$

$$— (1-z)_+$$

$$— \log(1 + \exp(-z))$$



6. Dual Formulation

Start with:

Lagrangian Duality

$$\inf_{\text{sc. } \forall i g_i(x) \leq 0} f(x)$$

which becomes:

$$\inf_x \sup_{\alpha \geq 0} f(x) + \sum_i \alpha_i g_i(x)$$

Define:

$$\mathcal{L}(x, \alpha) = f(x) + \sum_i \alpha_i g_i(x)$$

The following holds:

$$\inf_x \sup_{\alpha \geq 0} \mathcal{L}(x, \alpha) \geq \sup_{\alpha \geq 0} \inf_x \mathcal{L}(x, \alpha)$$

Quantity:

$$\inf_x \sup_{\alpha \geq 0} \mathcal{L}(x, \alpha) - \sup_{\alpha \geq 0} \inf_x \mathcal{L}(x, \alpha)$$

is called "Duality Gap". If it is 0, we speak of strong duality

For SVM, we admit the result that strong duality holds.
It comes from convexity.

Lagrangian function writes:

$$\mathcal{L}(w, b, \xi; \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^n \beta_i \xi_i$$

Dual function writes:

$$F(\alpha, \beta) = \inf_{w, b, \xi} \mathcal{L}(w, b, \xi; \alpha, \beta)$$

KKT conditions are:

$$\nabla_w \mathcal{L} = 0 \implies w = \sum_i \alpha_i y_i x_i$$

$$\nabla_b \mathcal{L} = 0 \implies \sum_i \alpha_i y_i = 0$$

$$\nabla_{\xi_i} \mathcal{L} = 0 \implies \alpha_i + \beta_i = C$$

which gives:

$$\begin{aligned} F(\alpha, \beta) &= \frac{1}{2} \left\| \sum_i \alpha_i y_i x_i \right\|^2 - \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_i \alpha_i y_i b + \sum_i \alpha_i \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \end{aligned}$$

Finally:

$$\sup_{\alpha \geq 0, \beta \geq 0} F(\alpha, \beta) = \sup_{0 \leq \alpha \leq C} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

We realize that α^*, β^* do not depend on x_i but on $\langle x_i, x_j \rangle$.
The problem is therefore equivariant to isometries.

This remark opens the way to **kernel methods**

KERNEL METHODS

1. Definitions

Déf. A positive kernel over a set \mathcal{X} is a function $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ satisfying:

- $\forall x \in \mathcal{X}, \forall x' \in \mathcal{X} \quad K(x, x') = K(x', x)$
- $\forall n \geq 1, \forall x_1, \dots, x_n, \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}$ is a positive semi definite matrix.

Examples.

$K(x, x') = \langle x, x' \rangle$ est un noyau (sym. pos.)

$K(x, x') = (\langle x, x' \rangle + 1)^2$ est un noyau (sym. pos.)

$K(x, x') = \langle \phi(x), \phi(x') \rangle$ est un noyau (sym. pos)

Plus généralement

We have the following fundamental result (Aronszajn, 1950)

Reproducing Kernel Hilbert Space.

Let K be a positive kernel over \mathcal{X} , there exists a Hilbert Space H and a map

$$\Phi: \mathcal{X} \rightarrow H$$

called the "feature map" such that:

$$\forall x \in \mathcal{X} \quad \forall x' \in \mathcal{X} \quad K(x, x') = \langle \phi(x), \phi(x') \rangle_H$$

and:

$$\forall h \in H \quad \forall x \in \mathcal{X} \quad h(x) = \langle h, K(x, \cdot) \rangle.$$

For instance, with $K(x, x') = (\langle x, x' \rangle + 1)^2$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$

one may take

$$\phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{pmatrix}$$

and we check that: $\langle \phi(x), \phi(x') \rangle = (\langle x, x' \rangle + 1)^2$

We also check that after applying Φ , the following problem

$$\Delta a = (-1, 1) \quad \bullet b = (1, 1) \quad \Delta \downarrow \phi(a) = (1, 1, \sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1) \quad \Delta \downarrow \phi(b) = (1, 1, \sqrt{2}, \sqrt{2}, \sqrt{2}, 1)$$

$$\bullet d = (-1, -1) \quad \Delta c = (1, -1) \quad \bullet \phi(d) = (1, 1, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 1) \quad \bullet \phi(c) = (1, 1, -\sqrt{2}, \sqrt{2}, -\sqrt{2}, 1)$$

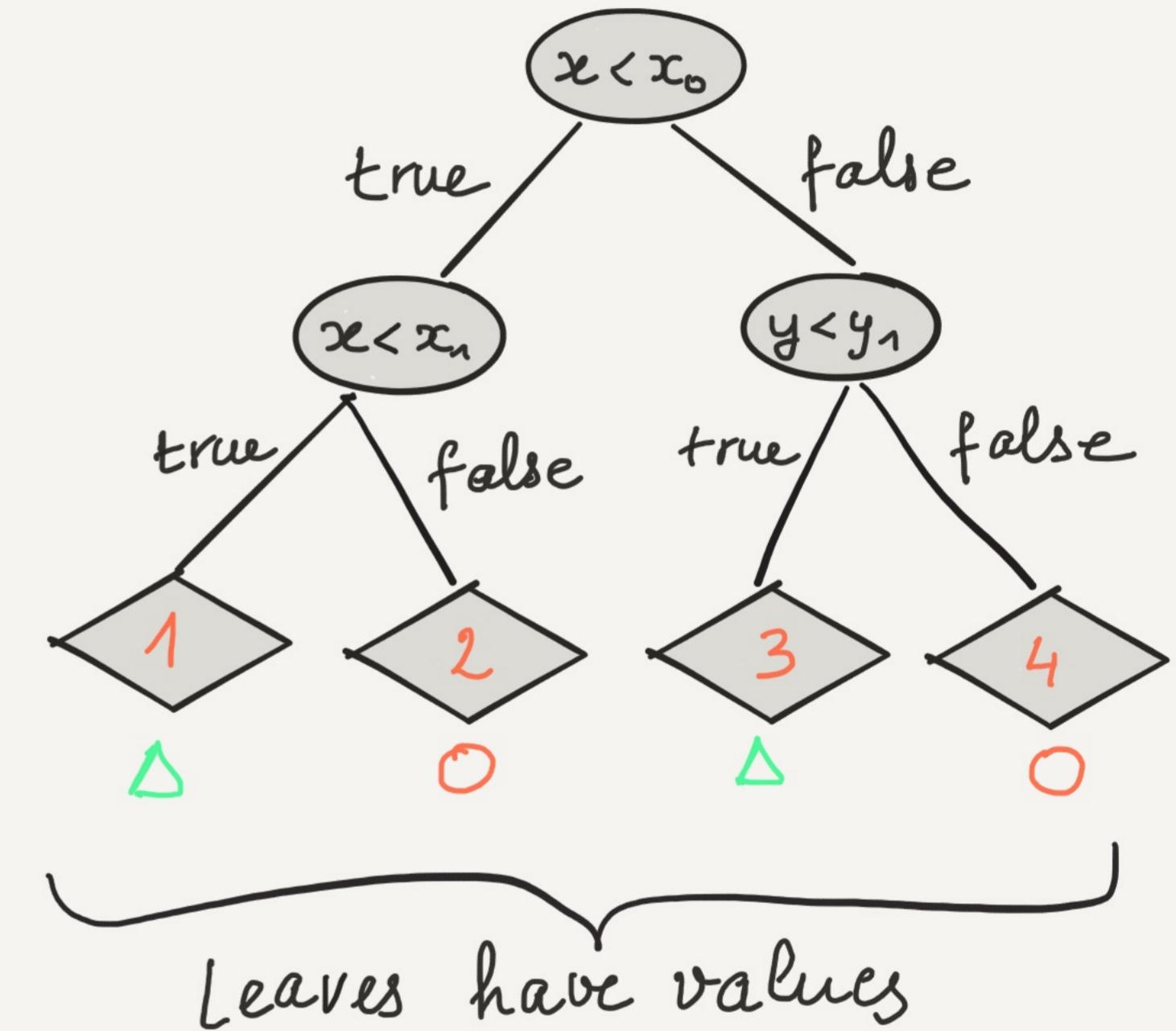
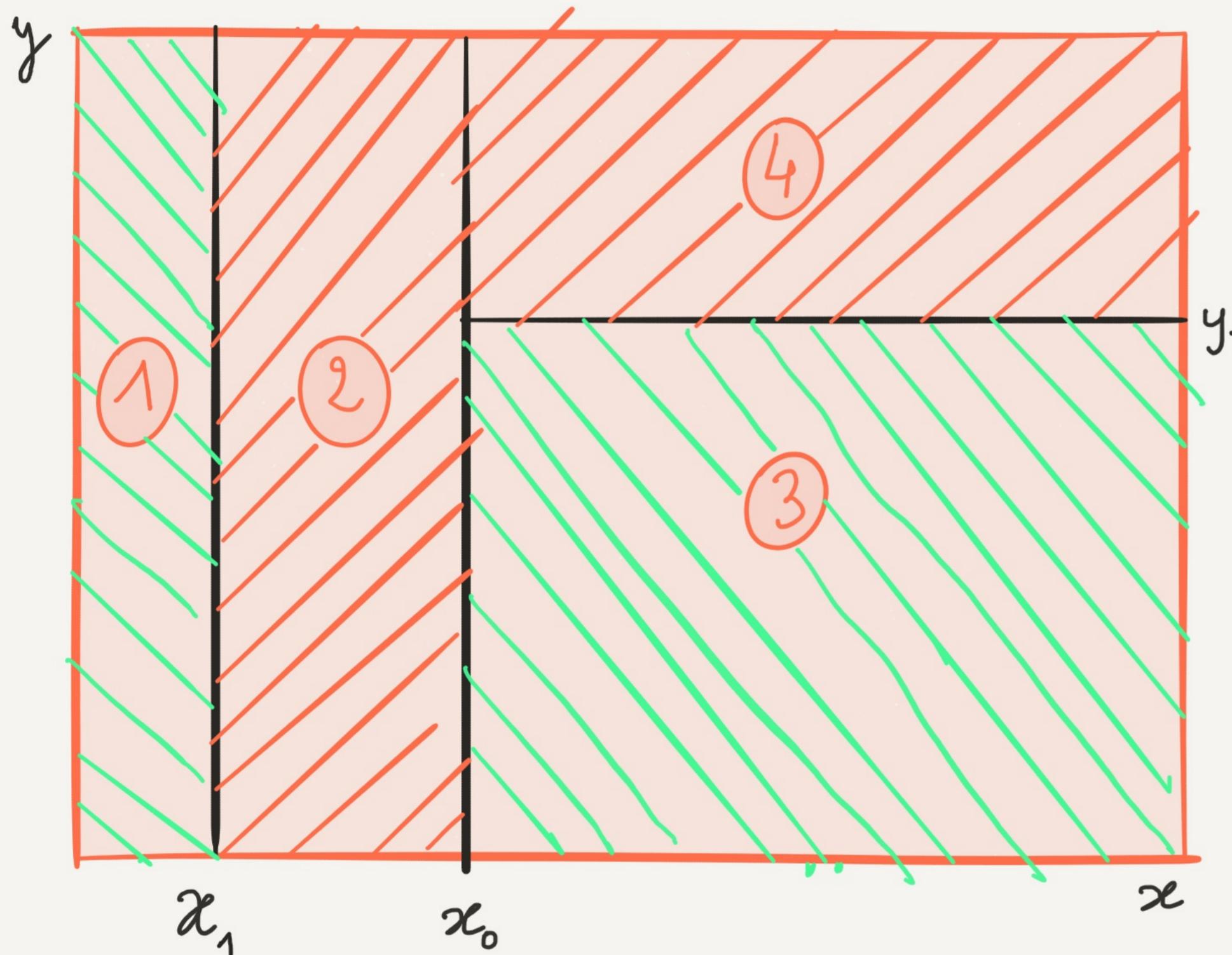
becomes **linearly separable**

$$(-1, 1) \circ D (1, 1)$$

$$(-1, -1) D \circ (1, -1)$$

DECISION TREES

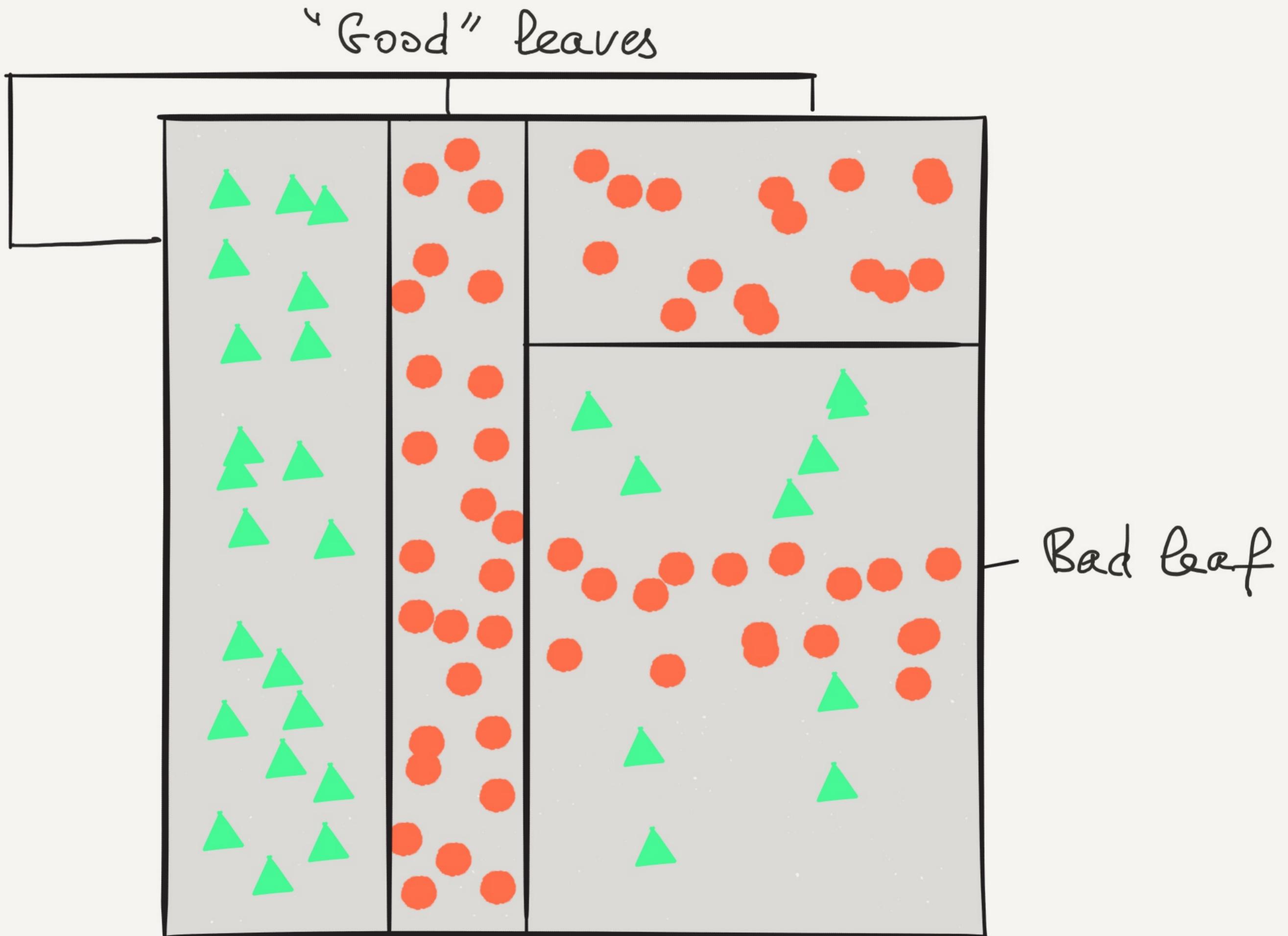
1. Introduction



Tree cells have borders that are parallel to the axis

2. Mesures de qualité

"Good" leaves have homogeneous data, "bad" leaves don't.



We can quantify how "good" a cell is using homogeneity measures. Let us mention 3 options.

A. Empirical Risk

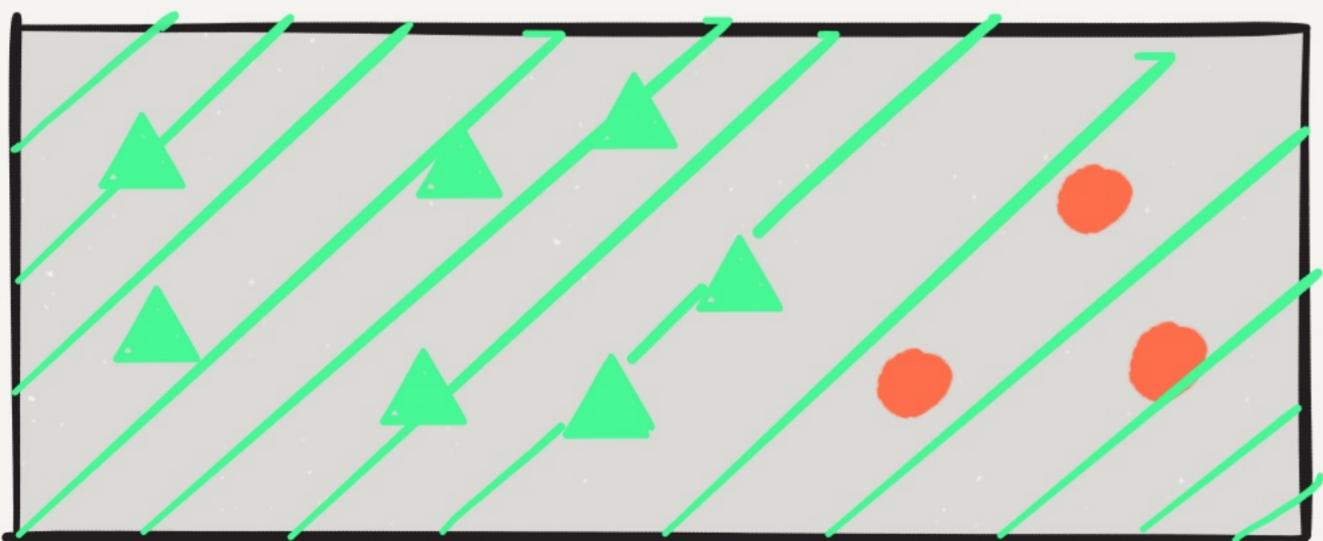
It starts with a "loss" function:

$$l : \{\Delta, 0\} \times \{\Delta, 0\} \rightarrow \mathbb{R}_+$$

For instance, the symmetric loss function:

$$\begin{aligned} l(\Delta, \Delta) &= l(0, 0) = 0 \\ l(\Delta, 0) &= l(0, \Delta) = 1 \end{aligned}$$

The associated empirical risk is then defined by:



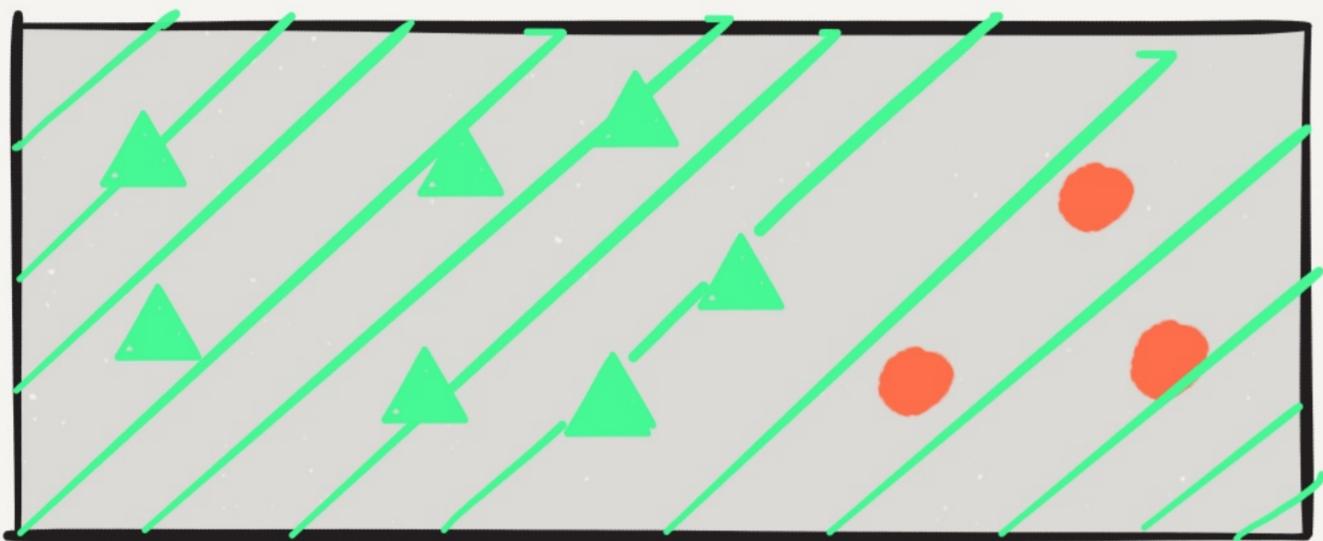
$$\min_{\bar{y} \in \{\Delta, 0\}} \sum_{(x, y) \in \text{cell}} l(y, \bar{y})$$

In this example it is 3.

B Gini measure

Define: $P_{\Delta} = \frac{\# \Delta}{N}$ and $P_0 = \frac{\# O}{N}$

For example



$$N = 10$$

$$P_{\Delta} = \frac{7}{10}$$

$$P_0 = \frac{3}{10}$$

GINI MEASURE then writes:

$$2 P_{\Delta} P_0$$

It reaches a maximum when $P_{\Delta} = P_0 = 1/2$ and a minimum when

$P_{\Delta} = 0$ or $P_0 = 0$.

C. Entropy

Define Entropy of a leaf as: (log is \log_2)

$$-P_{\Delta} \log P_{\Delta} - P_O \log P_O$$

It reaches a maximum when $P_{\Delta} = P_O = 1/2$ and a minimum when $P_{\Delta} = 0$ or $P_O = 0$.

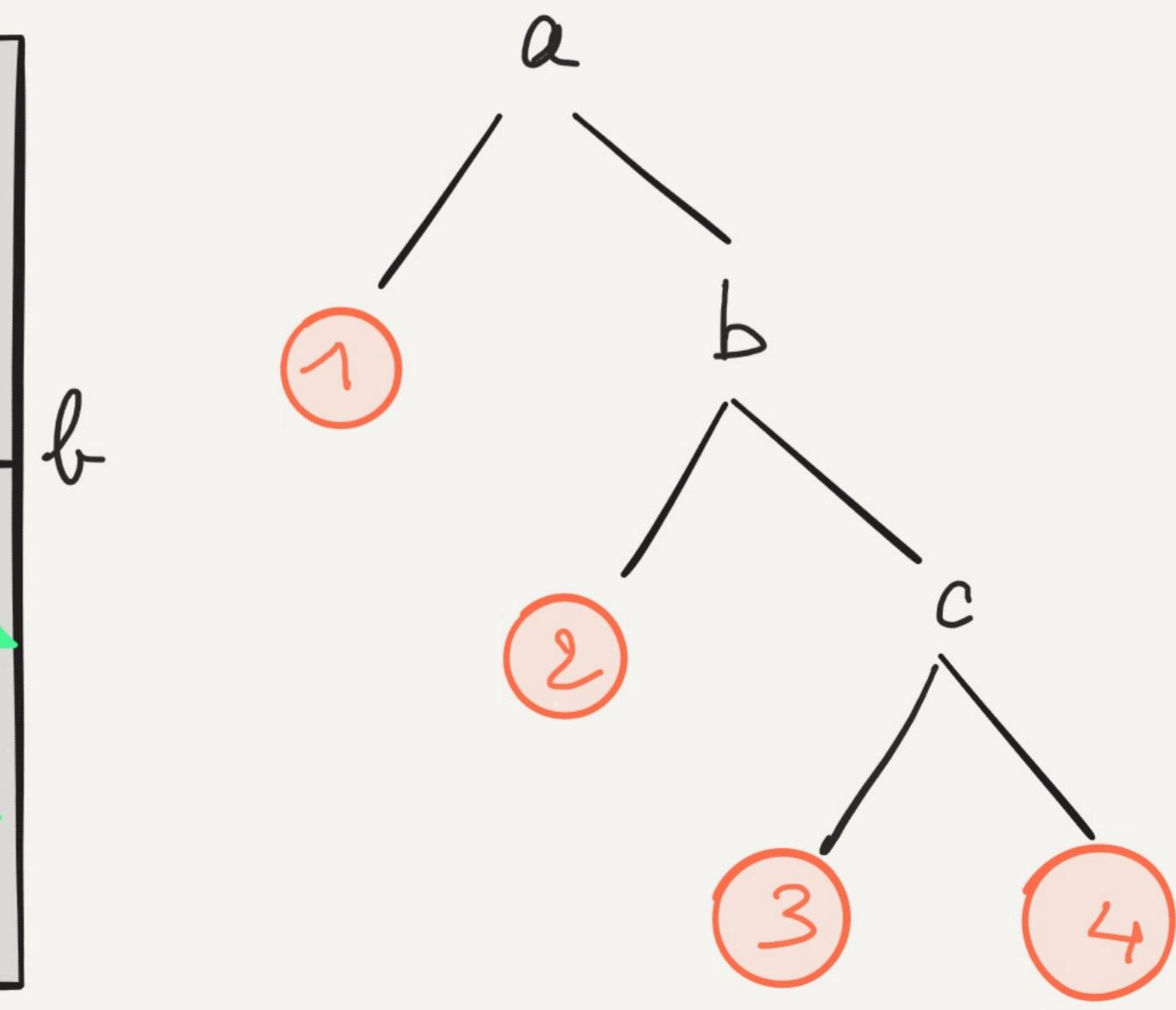
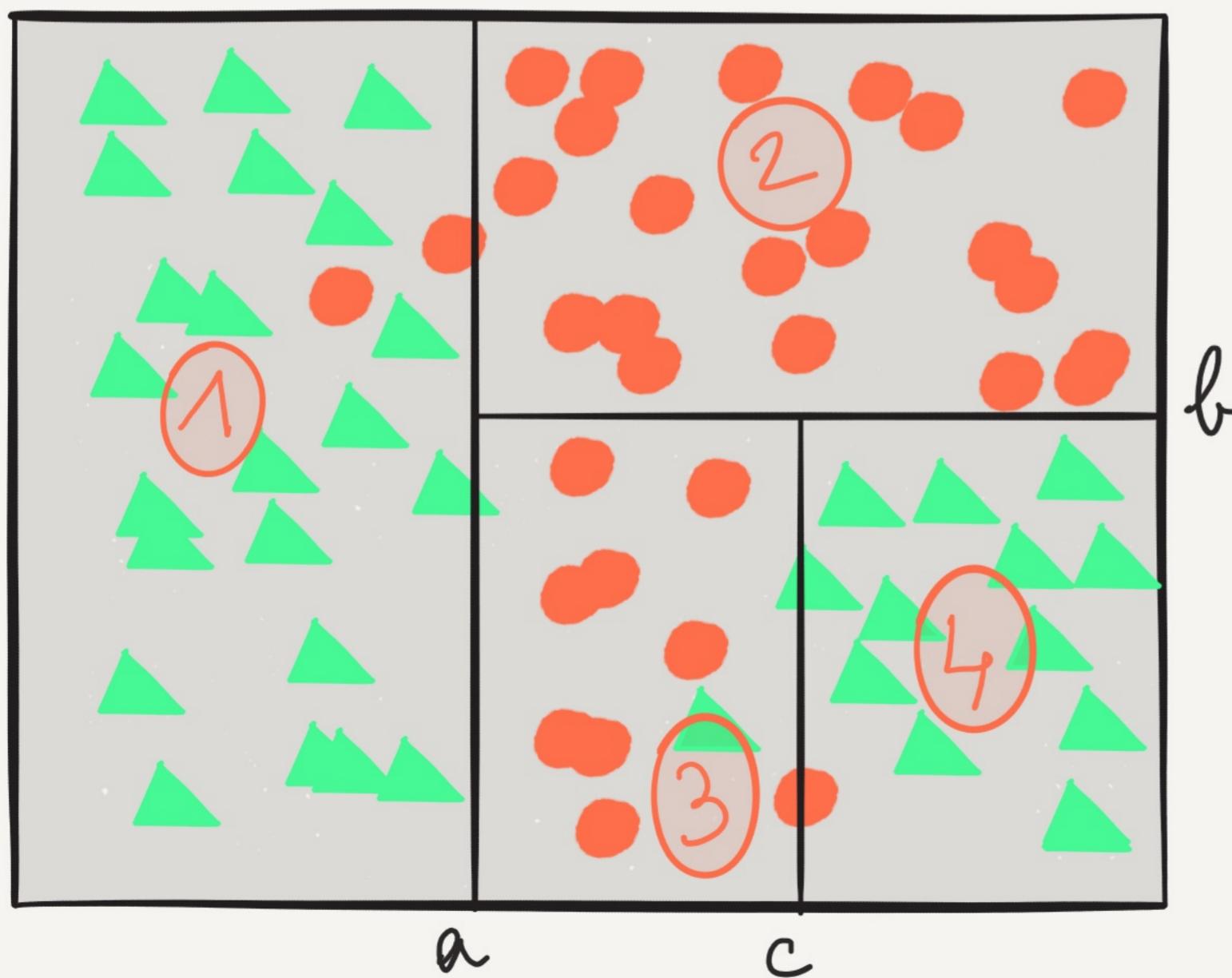
In the previous example Entropy is equal to:

$$-\frac{7}{10} \log \frac{7}{10} - \frac{3}{10} \log \frac{3}{10}$$

3. Greedy Algorithm

Minimizing the empirical risk yields an NP hard problem.

Therefore a greedy heuristic is used

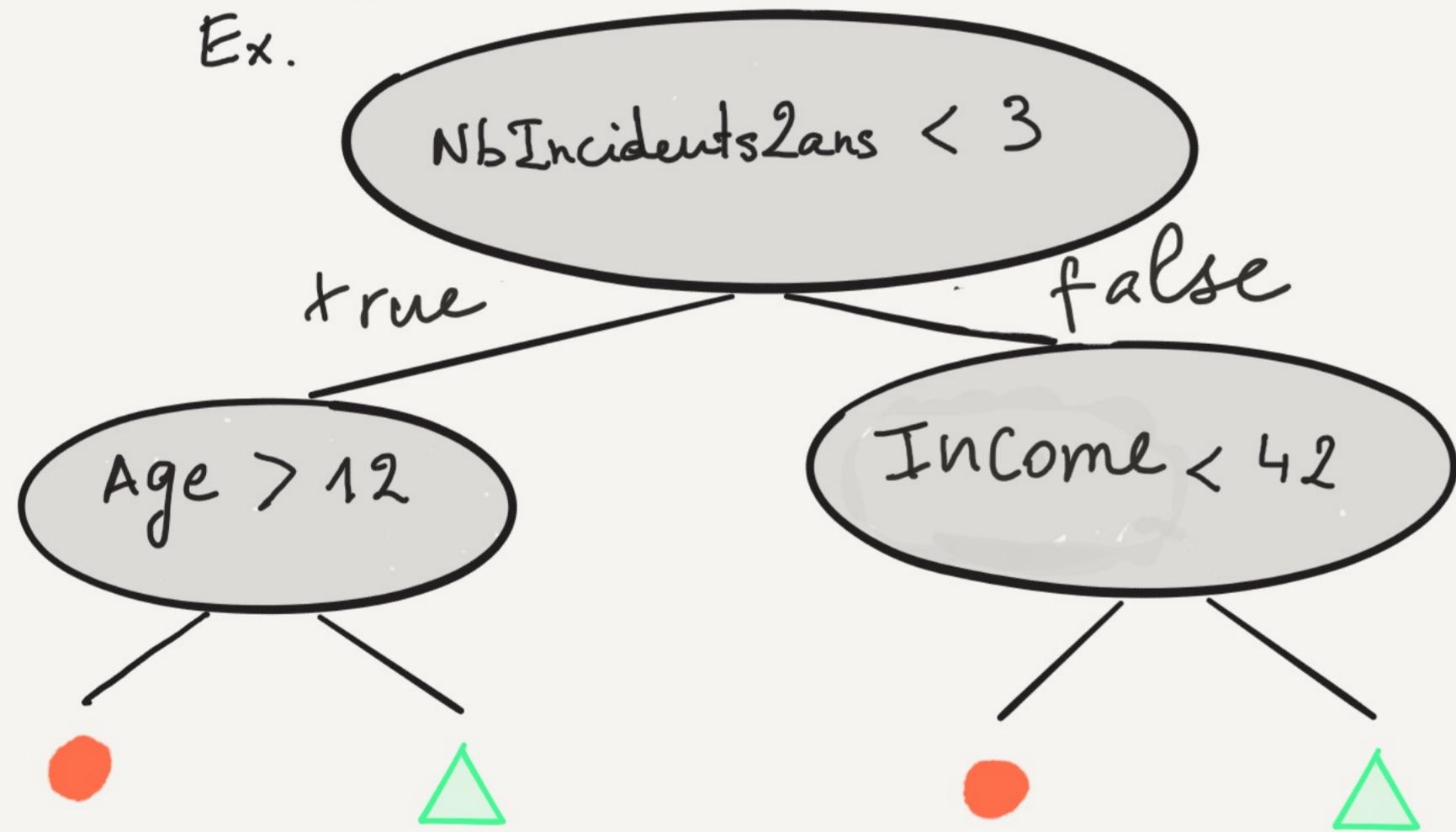


Strengths & Weaknesses

Strengths

- Simple
- Easy to interpret

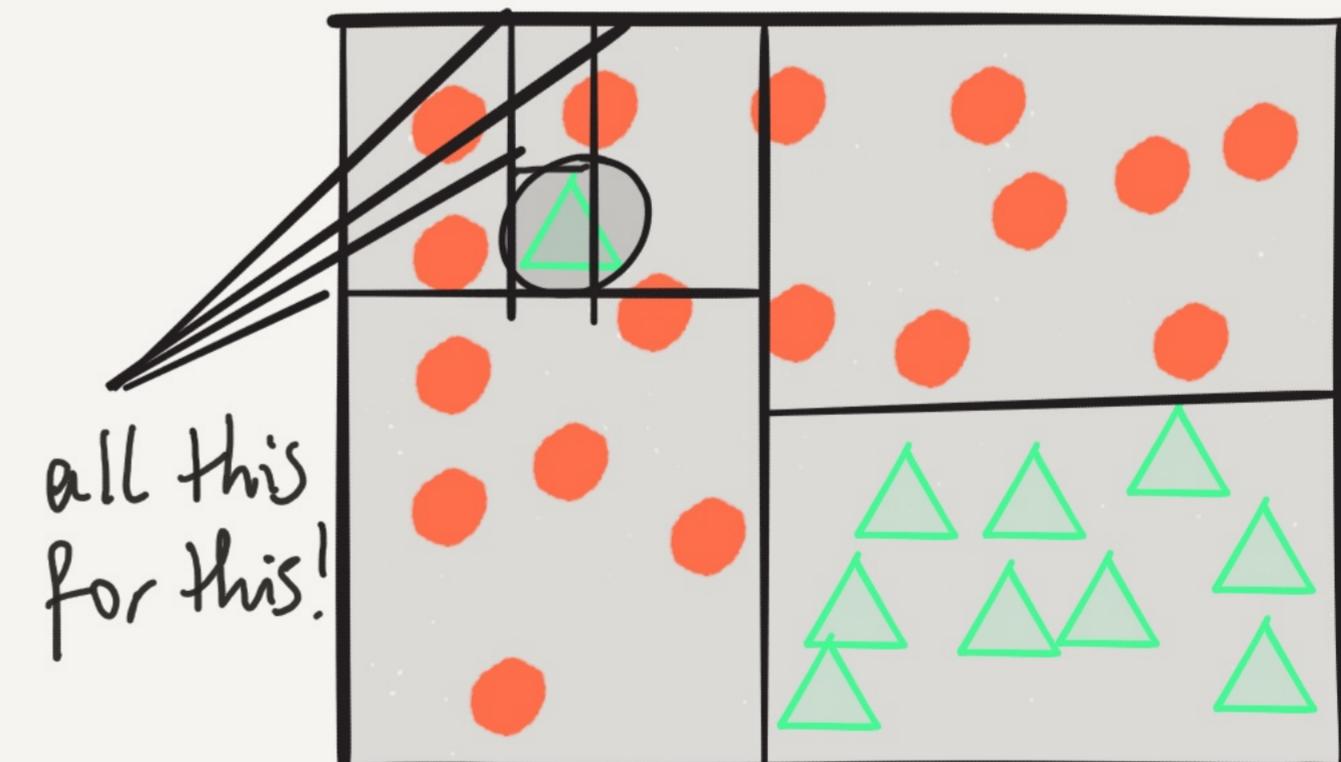
Ex.



- Non linear

Weaknesses

- Heuristic, no guarantees
- Tree can change a lot with new data



- Linear complexity over the dimension

RANDOM FORESTS

1. Main ideas

Random Forests aim at diminishing the weak points of decision while trying to maintain most of their benefits.

The 2 main tools are:

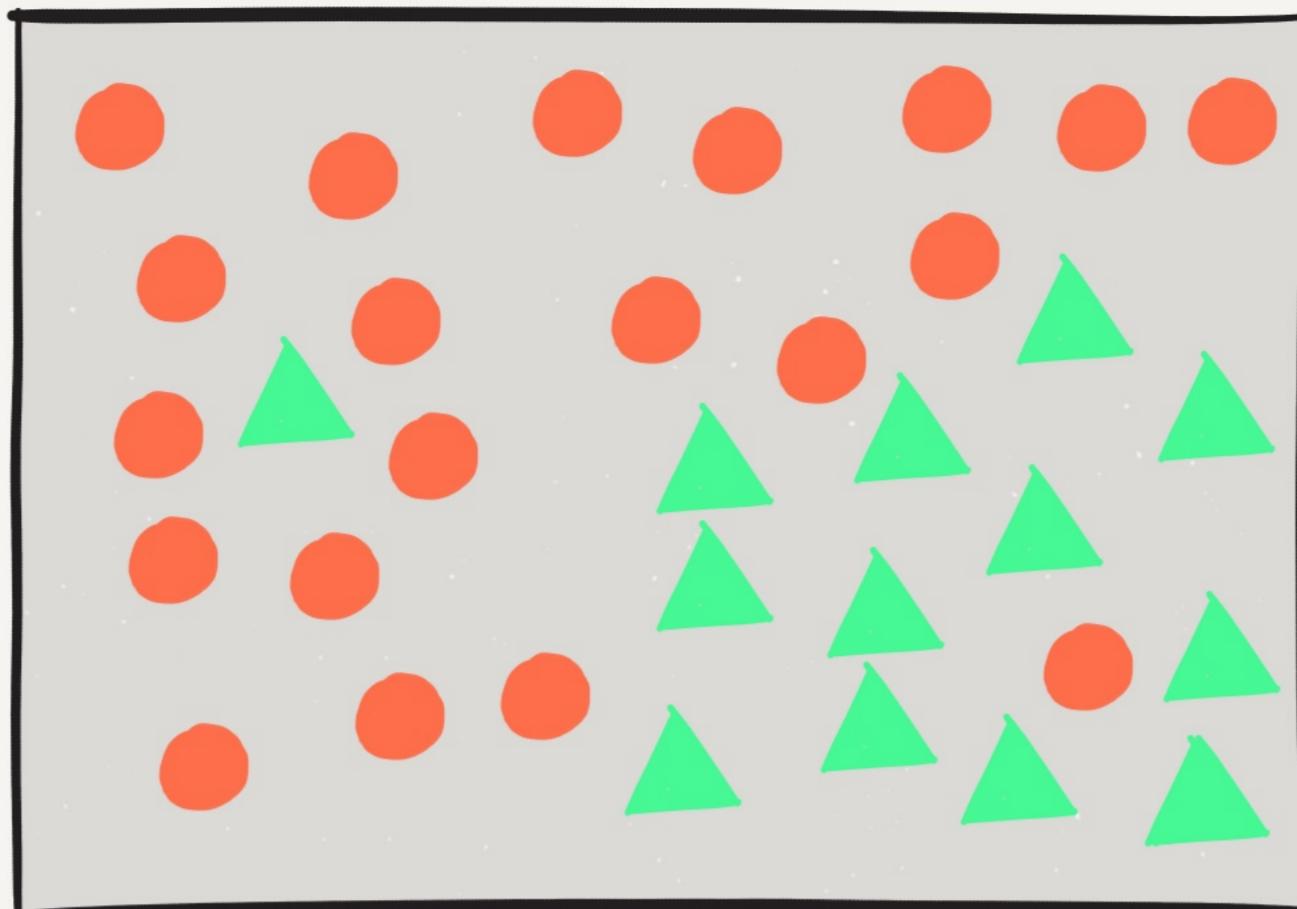
-

Randomisation

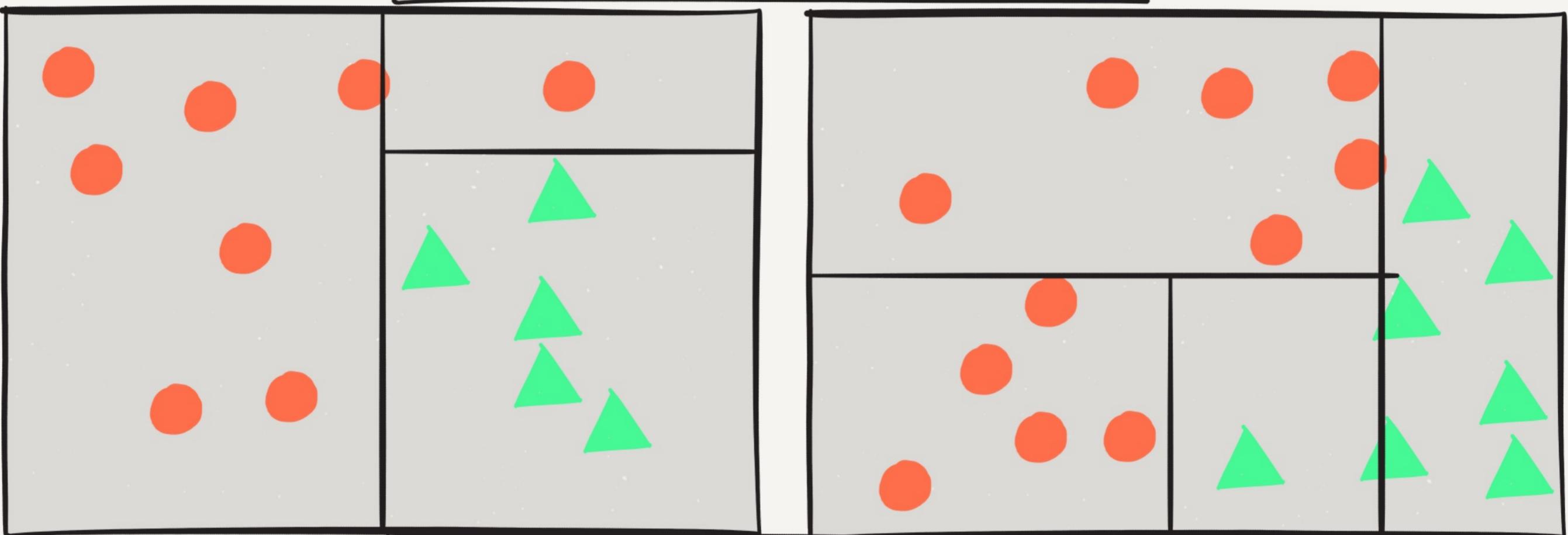
-

Averaging

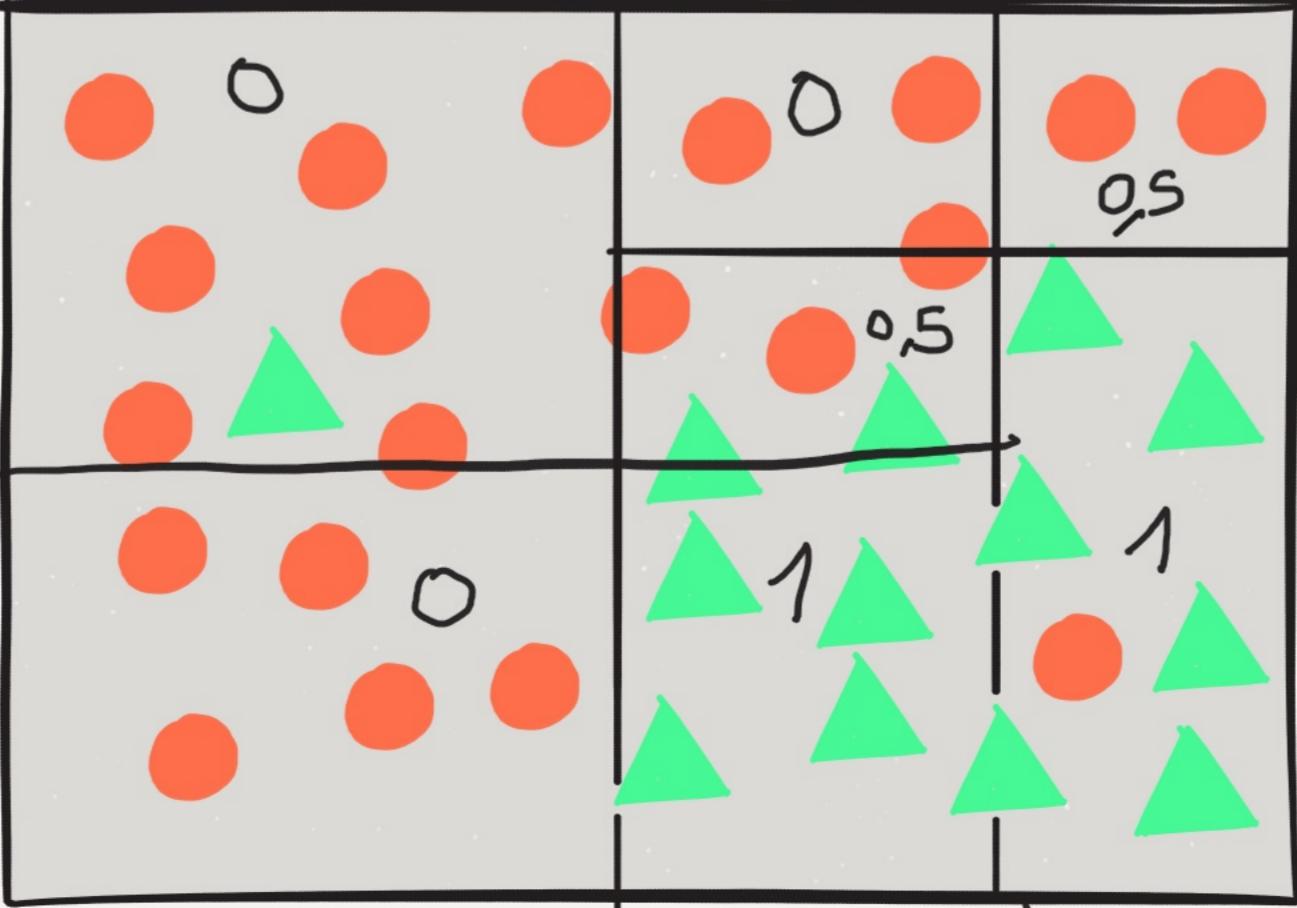
Illustration



Randomisation



Averaging

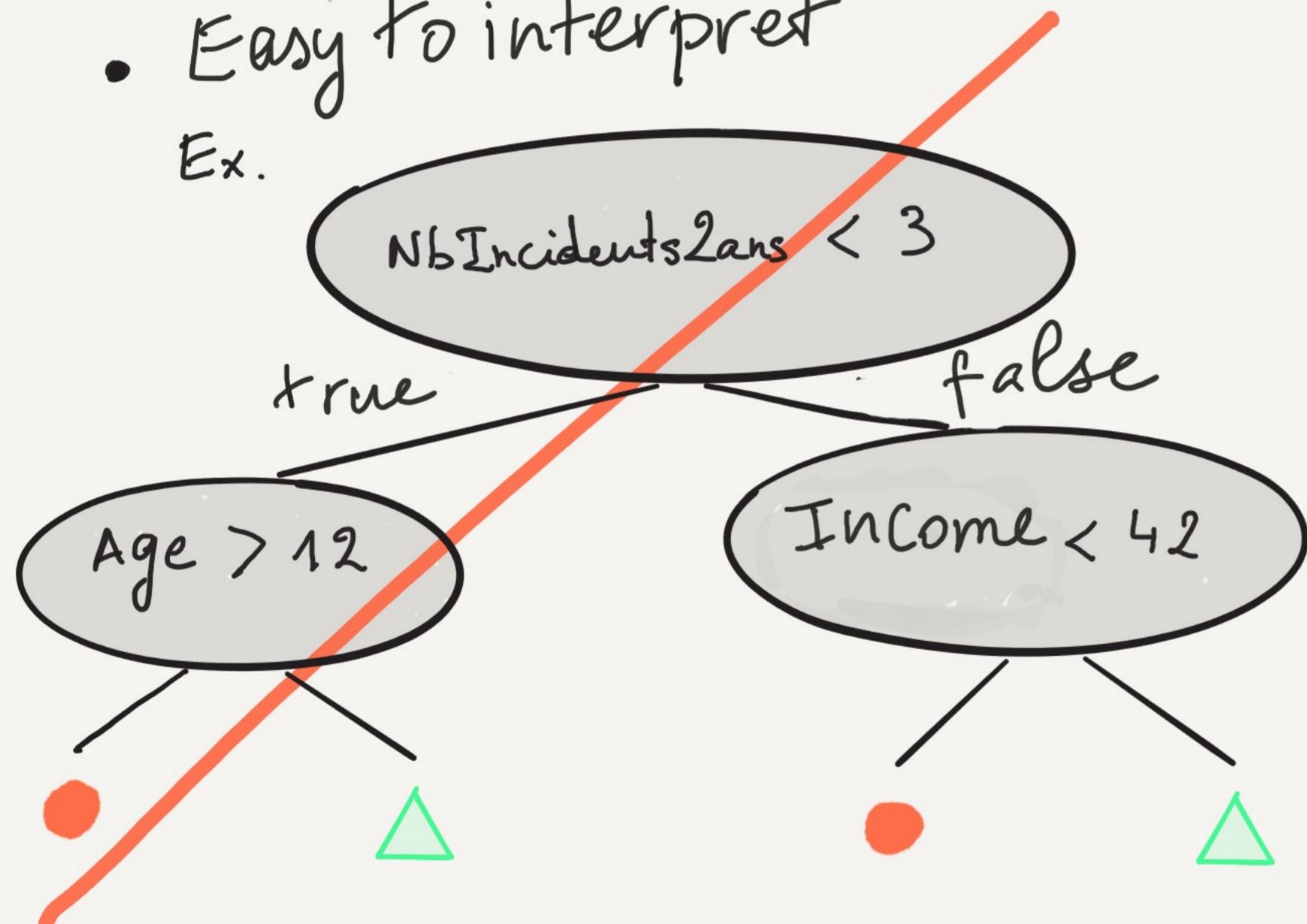


Strengths & Weaknesses

Strengths

- Simple
- Easy to interpret

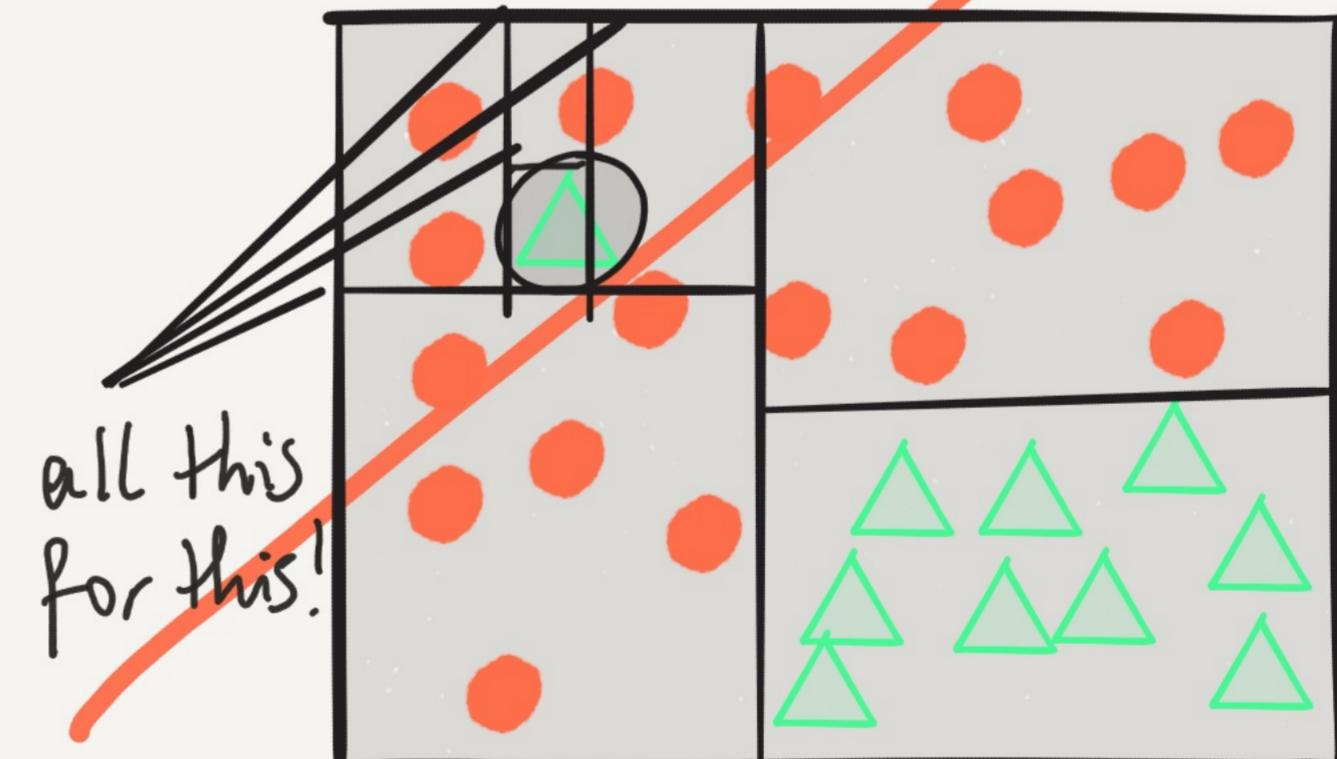
Ex.



- Non linear

Weaknesses

- Heuristic, no guarantees
- Tree can change a lot with new data



- Linear complexity over the dimension

BOOSTING

1. Framework

Combine **weak-learners** in a sequential and additive way:

$$h = \underbrace{\alpha_1 h_1 + \dots + \alpha_L h_L}_{\text{weak learner}} + \underbrace{\text{scalars}}_{\text{weak learner}}$$

"Strong" learner
Weak-learners example: shallow trees (**stumps**, i.e. trees with depth 1)

2. Déroulement.

Sequential means:

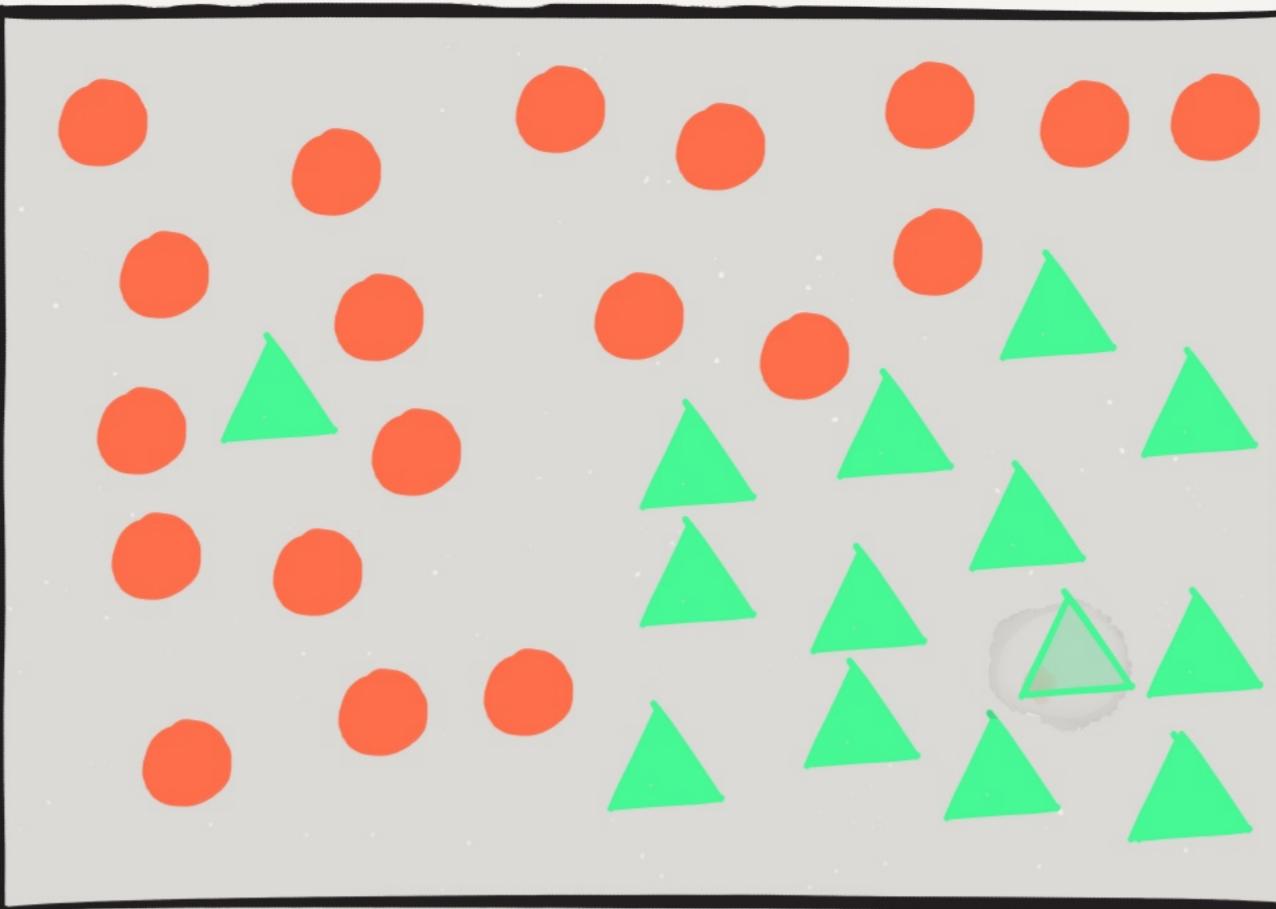
First find: $\alpha_1 h_1$

Then find: $\alpha_2 h_2$ (using results from $\alpha_1 h_1$)
etc.

Averaging means:

Then form $h = \alpha_1 h_1 + \alpha_2 h_2 + \dots$

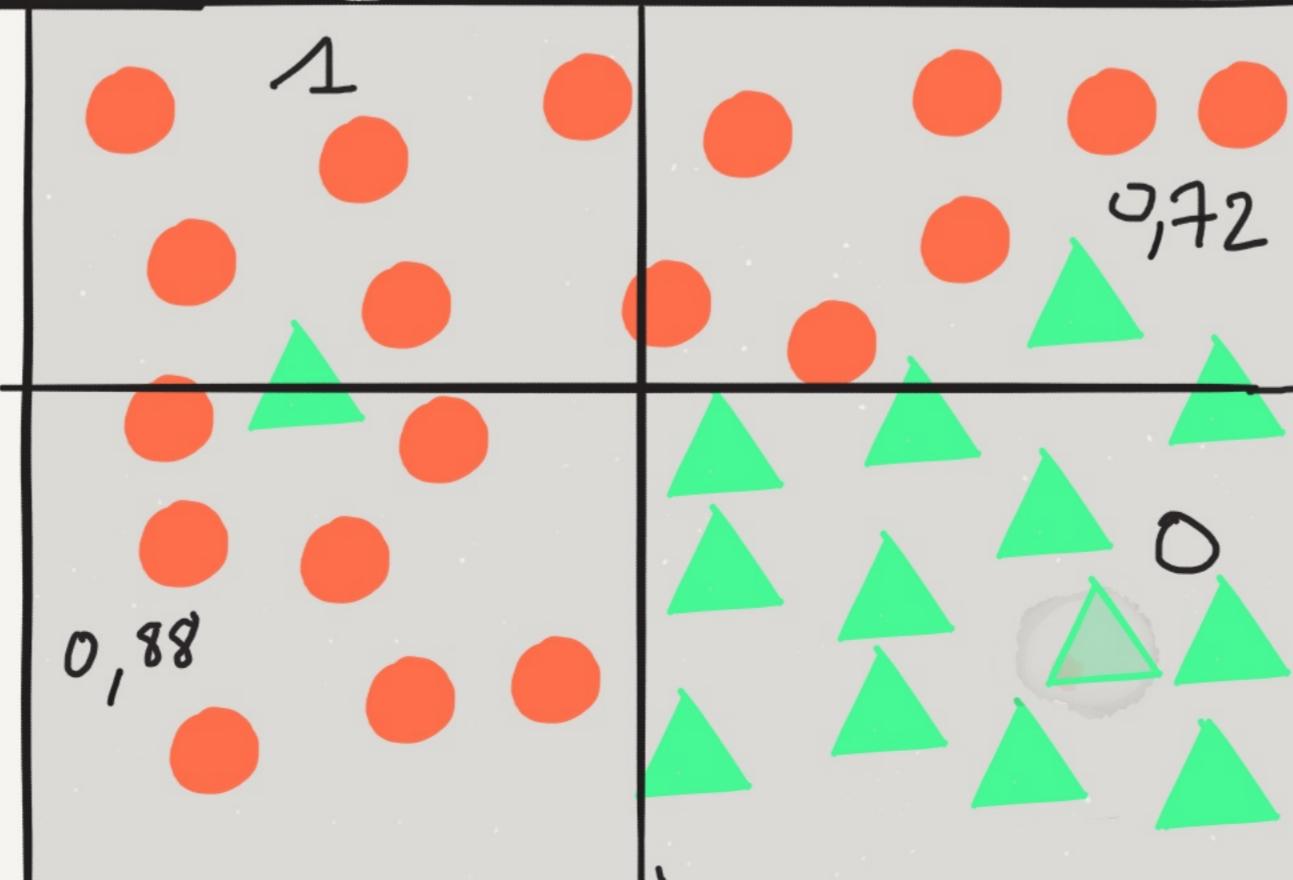
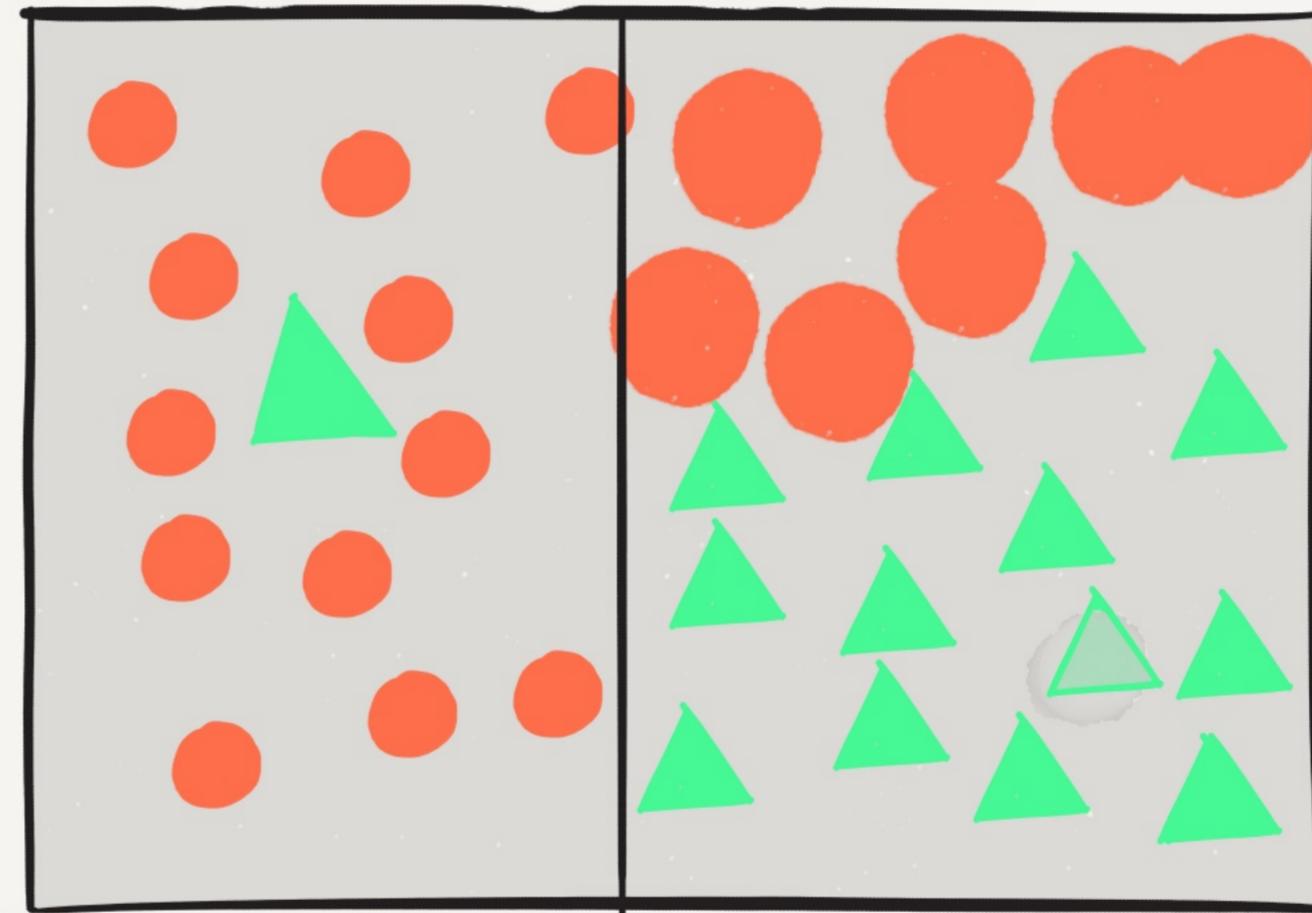
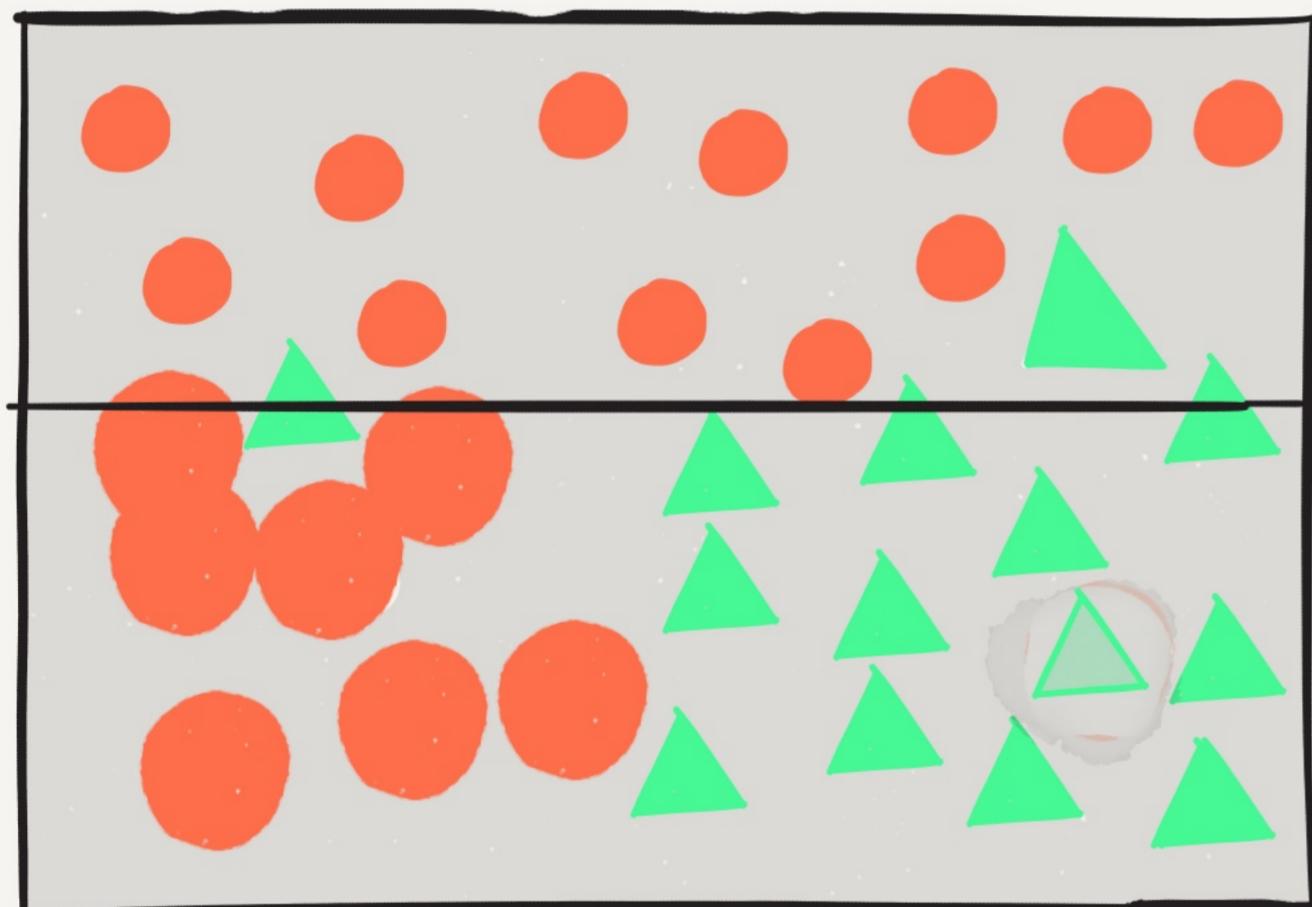
3. Boosting : Illustration



Data

$\alpha_1 h_1$

$\alpha_2 h_2$



Theoretical viewpoint: Anyboost

Framework

- data

$$(\underbrace{x_i}_{\mathbb{R}^D}, \underbrace{y_i}_{\mathbb{R}})_{1 \leq i \leq n}$$

- weak-learners $\mathcal{H} \subset \{\mathbb{R}^D \rightarrow \mathbb{R}\}$
- loss function $l: \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$

Let us start with this expression for a given h

$$\sum_{i=1}^n l(h(x_i), y_i)$$

Slightly modify h adding a small δh^* : $h + \delta h^*$

$$\sum_{i=1}^n l((h + \delta h)(x_i), y_i) = \sum_{i=1}^n l(h(x_i), y_i) + \sum_{i=1}^n \partial_l(h(x_i), y_i) \cdot \delta h(x_i)$$

Derivative is then

$$\delta h \rightarrow \sum_{i=1}^n \delta h(x_i) \partial_1 l(h(x_i), y_i)$$

Simply choose Δh minimizing

$$\sum_{i=1}^n \Delta h(x_i) \partial_1 l(h(x_i), y_i)$$

Find α such that $\sum_{i=1}^n l(h(x_i) + \alpha \Delta h(x_i), y_i)$
(line search)

is minimum.

5. Adaboost

$$l(y, y') = e^{-y y'}$$

$\mathcal{H} = \{ \text{shallow trees} \}$ ex stumps

$$\delta_h \rightarrow \sum_{i=1}^n \alpha_i l(h(x_i), y_i) \delta_h(x_i) = \sum_{i=1}^n -y_i e^{-y_i h(x_i)} \delta_h(x_i)$$

$$w_i = e^{-y_i h(x_i)} \quad (h = \alpha_1 h_1 + \dots + \alpha_t h_t)$$

$w_i \leq 1$ when x_i is well classified, otherwise $w_i > 1$

Find h_{t+1} so that the following quantity is minimum.

$$\sum_{i=1}^n w_i (-y_i h_{t+1}(x_i))$$

Want ± 1

It corresponds to
the # of errors of h_{t+1}
weighted by w_i

Then look for $\alpha_{t+1} = \operatorname{argmin}_{\alpha} \sum_{i=1}^n w_i e^{-y_i \alpha h_{t+1}(x_i)}$

$$E_{t+1} = \{(x_i, y_i) : y_i h_{t+1}(x_i) = -1\}$$

Then:

$$\sum_{i=1}^n w_i e^{-y_i \alpha h_{t+1}(x_i)} = \sum_{i \in E_{t+1}} w_i e^\alpha + \sum_{i \notin E_{t+1}} w_i e^{-\alpha}$$

We look for α_{t+1} such that:

$$\left(\sum_{i \in E_{t+1}} w_i \right) e^{\alpha_{t+1}} - \left(\sum_{i \notin E_{t+1}} w_i \right) e^{-\alpha_{t+1}} = 0$$

Equivalently:

$$\alpha_{t+1} = \frac{1}{2} \ln \left(\frac{\sum_{i \notin E_{t+1}} w_i}{\sum_{i \in E_{t+1}} w_i} \right)$$

PERFORMANCE METRICS

Confusion Matrix

		ground truth	
		True	False
prediction	True	True Positive	False Alarm
	False	Miss	True Negative

Metric #1: TPR

$$\frac{\text{True Positive}}{\text{Total \# positive}}$$

Metric #2: FPR

$$\frac{\text{False Alarms}}{\text{Total \# Negatives}}$$

ROC curve (Soft classifiers)

Features	Predictions	Ground Truth	Points
x_1	$h(x_1) 0.92$	$y_1 1$	$(0, 0)$
x_2	$\geq h(x_2) 0.86$	$y_2 1$	$(0, \frac{1}{\#P})$
\vdots	\vdots	$\vdots 0$	$(0, 1 - \frac{1}{\#P})$
x_n	$h(x_n) 0.03$	$y_n 0$	$(\frac{1}{\#N}, 0)$

