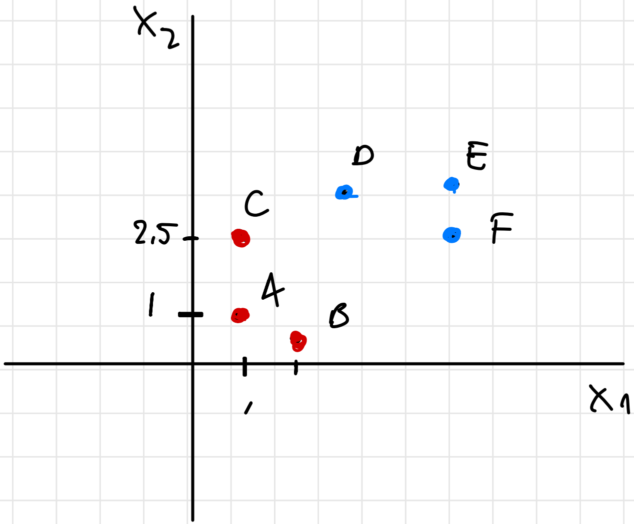
K-NEAREST NEIGHBORS

AND

DECISION TREES

(1)



1—NN CLASS.

+ LEAVE-ONE-OUT CROSSVALIDATION

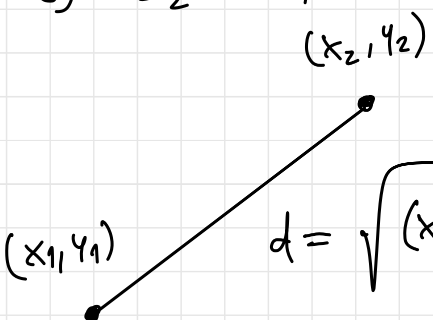
↳ A B C D ~~E~~ F

↑

PREDICT CLASS FOR E

→ REPEAT FOR EVERY POINT

b) L_2 DISTANCE



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$L_2(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

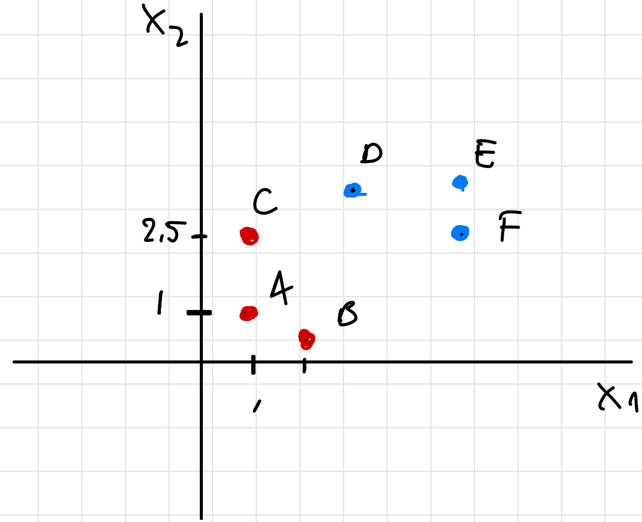
$$L_2(A,B) = \sqrt{(1-2)^2 + (1-0{,}5)^2} = 1.12$$

POINTS	CLOSEST POINT	PREDICTED CLASS	TRUE CLASS
A	B	●	●
B	A	●	●
C	A	●	●
D	C	●	●
E	F	●	●
F	E	●	●

a) L_1 DISTANCE

$$L_1(x,y)=\sum_i |x_i-y_i|$$

$$d=|x_1-y_1|+|x_2-y_2|$$

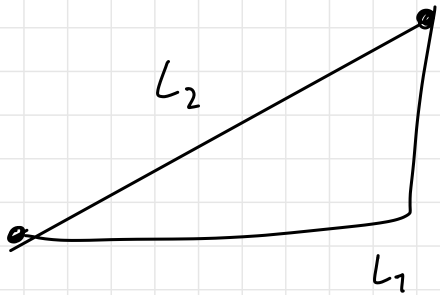


$$L_1(A,B)=|\underbrace{1-2}_{1}|+|\underbrace{1-0,5}_{0,5}|=1,5$$

$$L_1(A,C)=|\underbrace{1-1}_{0}|+|\underbrace{1-2,5}_{1,5}|=1,5$$

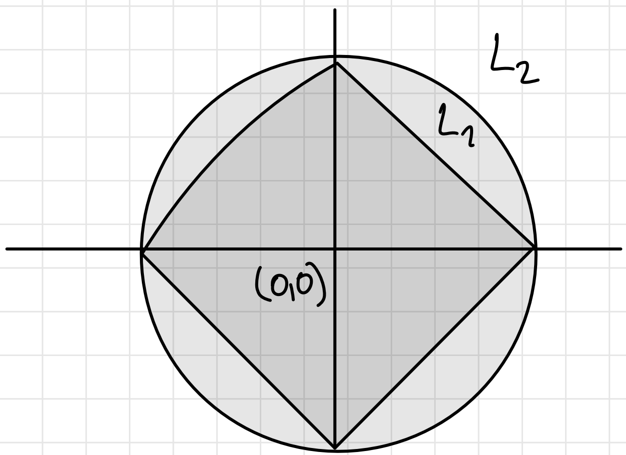
POINT	CLOSEST POINT	PREDICTED CLASS	TRUE
A	B, C	●	●
B	A	●	●
C	A	●	● ✓
D	E	●	●
E	F	●	●
F	E	●	● ✓

L_1 vs. L_2



$L_2 \leq L_1$

Fix $(0,0)$ $\rightarrow$ A SET OF POINTS THAT HAVE DISTANCE 1?



$x^2+y^2=1$

$|x|+|y|=1$

(2)

CLASSES: A, B, C

$N_A = 16$

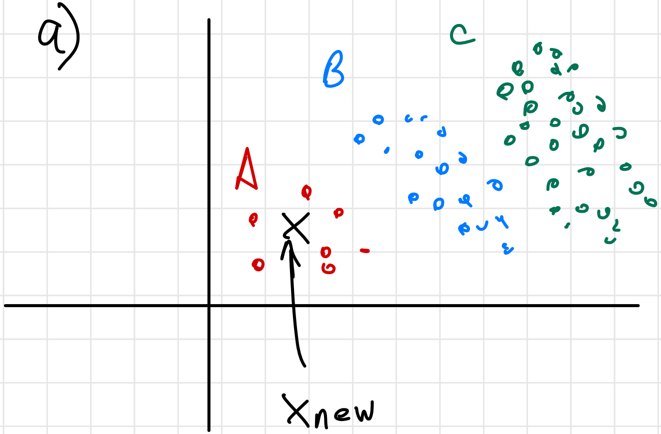
$N_B = 32$

$N_C = 64$

k-NN CLASSIFIER (UNWEIGHTED)

$k = N_A + N_B + N_C$

a)



$X_{new} \longrightarrow$ CLASS C

b) WHAT ABOUT WEIGHTED k-NN?

DON'T KNOW!

↳ DEPENDS ON DISTANCES

③

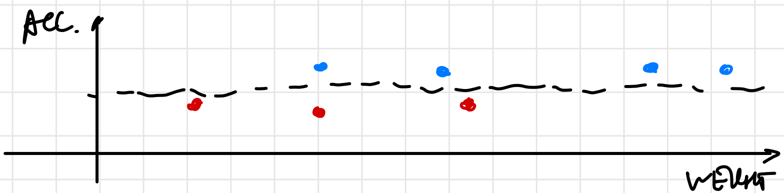
> 100 SAMPLES

Acceleration	max. velocity [km/h]	PS	cylinder capacity [cm^3]	weight [kg]	class
3.6	250	600	3996	2150	car
12.5	178	150	1968	2001	van
3.5	200	113	937	227	motorcycle
...	...	...	...	...	...

You observe that the obtained model performs bad on the test set. What might be the problem? Name

WHY? | HOW TO SOLVE? | WOULD DECISION TREE HAVE THE SAME PROB?

i) DIFFERENT RANGES OF FEATURES → STANDARDIZE | NO



$$x_i \leftarrow \frac{x_i - \mu_i}{\sigma_i}$$

SPLITS ARE BASED ON MISSCLASS. RATE, GINI ...
→ DEPENDS ON LABELS

ii) BAD k - HYPERPARAMETER → OPTIMIZE k (GRID SEARCH) | NO

iii) SHIFT BETWEEN TRAIN AND TEST DATA → CHOOSE TRAIN / TEST SET FROM SAME DISTRIBUTION → YES!

④ 1-KNN WITH i) L_1, ii) L_2 NORM.

L_2 DISTANCE IS ALWAYS SMALLER OR EQUAL TO L_1.

$$d_2(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

$$(z^2 = |z|^2 = |z^2|)$$

PROOF

$$d_2(x,y)^2 = \sum_i (x_i - y_i)^2$$

$$= \sum_i |x_i - y_i|^2$$

$$\leq \sum_i |x_i - y_i|^2 + 2\sum_i \sum_{j \neq i} |x_i - y_i||x_j - y_j|$$

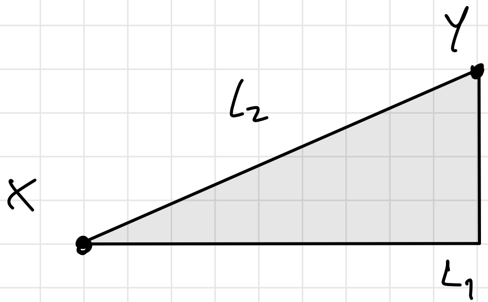
$$= d_1(x,y)^2$$

$$d_2(x,y)^2 \leq d_1(x,y)^2, \quad d_2, d_1 \geq 0 \Rightarrow d_2(x,y) \leq d_1(x,y)$$

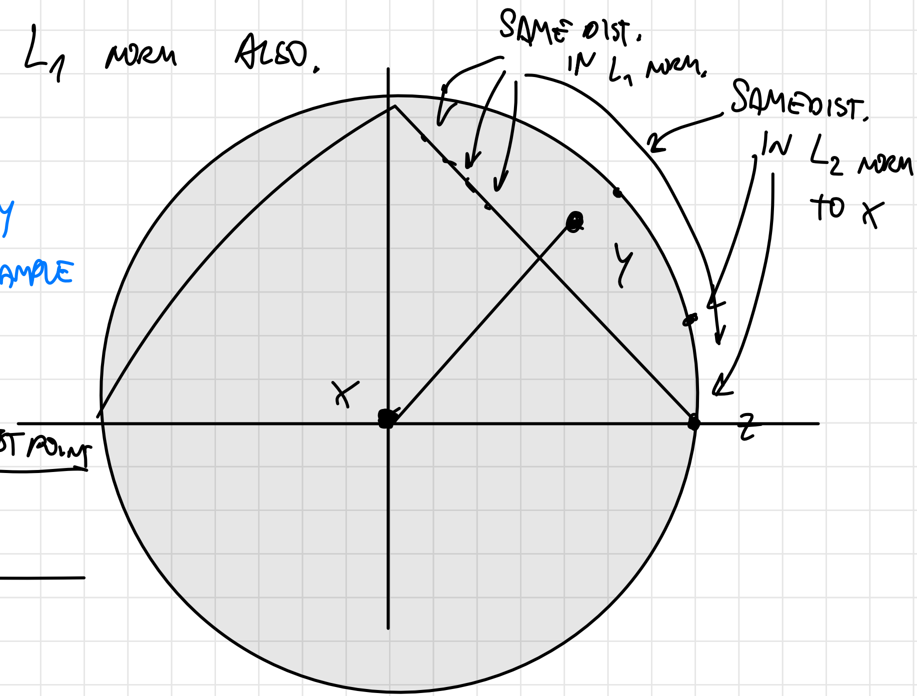
$$d_1(x,y)^2 = \left(\sum_i |x_i - y_i|\right)^2 = \sum_i |x_i - y_i|^2 + 2 \cdot \sum_i \sum_{j \neq i} |x_i - y_i||x_j - y_j|$$

$$(a+b+c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc = \sum \ldots^2 + 2\sum\sum \ldots$$

⑤ GIVEN $x, y \in \mathbb{R}^2$, IF x IS THE NEAREST TO y IN L_2 NORM THEN IT IS NEAREST IN L_1 NORM ALSO.



PROOF BY COUNTEREXAMPLE



	CLOSEST POINT
L_1 NORM	Z
L_2 NORM	Y

$x = (0,0)$

$z = (1,0)$

$y = (0{,}5,\ 0{,}65)$

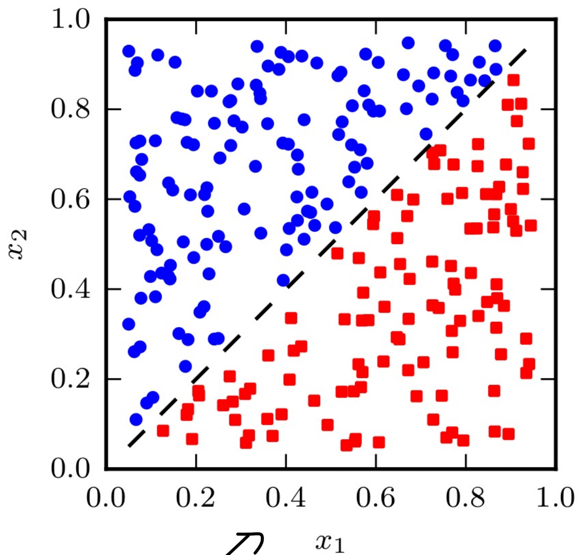
$L_1(x,z) = |1-0| + |0-0| = 1$

$L_2(x,z) = \sqrt{(1-0)^2 + (0-0)^2} = 1$

$L_1(x,y) = |0{,}5 - 0| + |0{,}65 - 0| = 1{,}15$

$L_2(x,y) = \sqrt{(0{,}5-0)^2 + (0{,}65-0)^2} = 0{,}64$

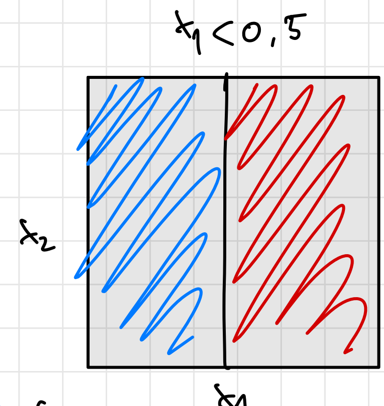
6



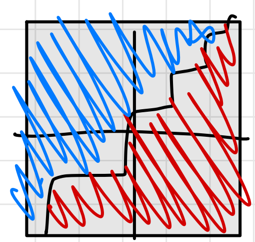
FEATURES

IS THERE A TREE WITH DEPTH = 1 THAT HAS 100% ACCURACY?

NO!



EVEN WITH A DEEPER TREE :



7

No.	x_1 (Team or Individual)	x_2 (Mental or Physical)	x_3 (Skill or Chance)	y (Win or Lose)
1	T	M	S	W
2	I	M	S	W
3	T	P	S	W
4	I	P	C	W
5	T	P	C	L
6	I	M	C	L
7	T	M	S	L
8	I	P	S	L
9	T	P	C	L
10	I	P	C	L

4

6

10

$p(y=W) = \frac{4}{10}$

$p(y=L) = \frac{6}{10}$

a) ENTROPY $i_H(y)$ OF CLASS LABELS y

$\downarrow$

$$-\sum_i p(x_i) \log p(x_i)$$

$$i_H(y) = -p(y=W)\log p(y=W) - p(y=L)\log p(y=L)$$

$$= -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$

No.	x_1 (Team or Individual)	x_2 (Mental or Physical)	x_3 (Skill or Chance)	y (Win or Lose)
1	T	M	S	W
2	I	M	S	W
3	T	P	S	W
4	I	P	C	W
5	T	P	C	L
6	I	M	C	L
7	T	M	S	L
8	I	P	S	L
9	T	P	C	L
10	I	P	C	L

4

6

10

b) BUILD OPTIMAL TREE (DEPTH = 1) USING ENTROPY MEASURE :

SPLIT ON X_1 : $N_T = 5$, $N_I = 5$ $\Rightarrow$ $p(x_1 = T) = p(x_1 = I) = \frac{1}{2}$

$p(y = W \mid x_1 = T) = \frac{2}{5}$ $\quad$ $p(y = L \mid x_1 = T) = \frac{3}{5}$

$p(y = W \mid x_1 = I) = \frac{2}{5}$ $\quad$ $p(y = L \mid x_1 = I) = \frac{3}{5}$

$$\left.\begin{aligned} i_H(x_1 = T) &= -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} \\ i_H(x_1 = I) &= -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} \end{aligned}\right\} 0{,}97$$

$$\Delta(x_1) = i_H(y) - p(x_1 = T)\, i_H(x_1 = T) - p(x_1 = I)\, i_H(x_1 = I) = 0$$

SPLIT ON x_2 :

$$p(x_2 = M) = \frac{4}{10} \qquad p(x_2 = P) = \frac{6}{10}$$

$$p(y = W \mid x_2 = M) = \frac{2}{4} \qquad p(y = L \mid x_2 = M) = \frac{2}{4}$$

$$p(y = W \mid x_2 = P) = \frac{2}{6} \qquad p(y = L \mid x_2 = P) = \frac{4}{6}$$

$$i_H(x_2 = T) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$$

$$i_H(x_2 = P) = -\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6} \simeq 0.92$$

$$\Delta(x_2) = i_H(y) - p(x_2 = M)\, i_H(x_2 = M) - p(x_2 = P)\, i_H(x_2 = P)$$

$$= 0,018$$

SPLIT ON x_3 !

$$p(x_3 = S) = \frac{5}{10} \qquad p(x_3 = S) = \frac{5}{10}$$

$$p(y = W \mid x_3 = S) \qquad p(y = L \mid x_3 = S) \quad \ldots$$

$$p(y = W \mid x_3 = C) \qquad p(y = L \mid x_3 = C) \quad \ldots$$

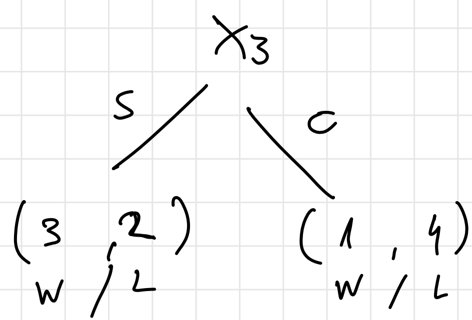
$$i_H(x_3 = C) = -\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} \approx 0.97$$

$$i_H(x_3 = C) = -\frac{1}{5} \log \frac{1}{5} - \frac{4}{5} \log \frac{4}{5} \approx 0.72$$

$$\Delta(x_3) = i_H(y) - p(x_3 = S)\, i_H(x_3 = S) - p(x_3 = C)\, i_H(x_3 = C) = 0.125$$

$$\Delta(x_2) = 0.018$$

$$\Delta(x_1) = 0$$



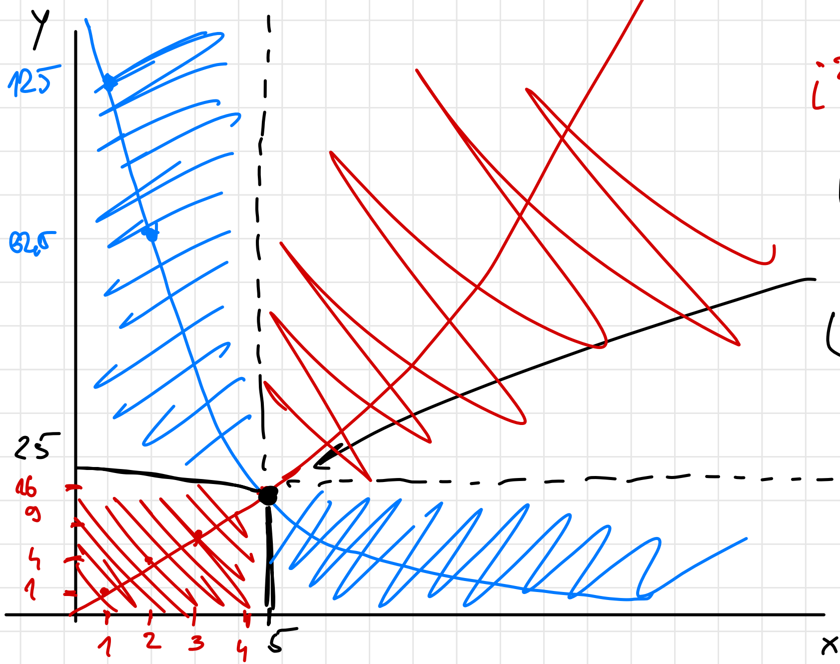
(8) 2 CLASSES C_1, C_2, POINTS IN $\mathbb{R}^2$

FIND: MINIMAL DEPTH TREE THAT ASSIGNS AS MANY POINTS CORRECTLY

→ HOW MANY ARE MISSCLASSIFIED ?

C_1 POINTS $\left\{ (i, i^2) \mid i \in \{1, \dots, 100\} \right\} \subseteq \mathbb{R}^2$

C_2 POINTS $\left\{ \left(i, \frac{125}{i}\right) \mid i \in \{1, \dots, 100\} \right\} \subseteq \mathbb{R}^2$



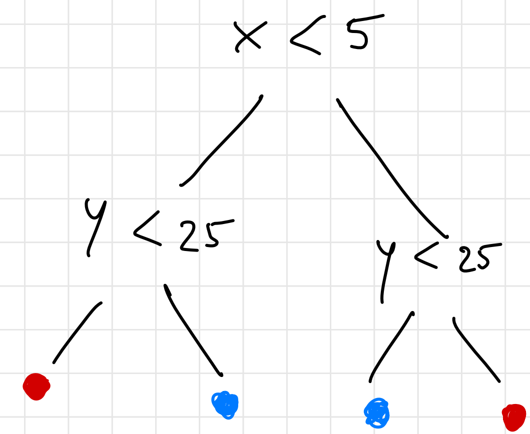
$(i, i^2) = \left(i, \frac{125}{i}\right)$

$i^2 = \frac{125}{i}$

$i^3 = 125$

$i = 5$

$(5, 25)$



1 MISSCLASSIFIED : $(5, 25)$