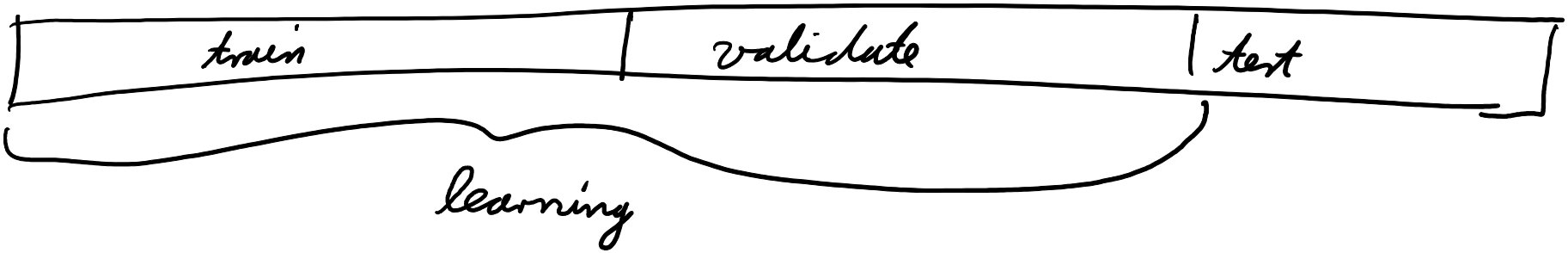
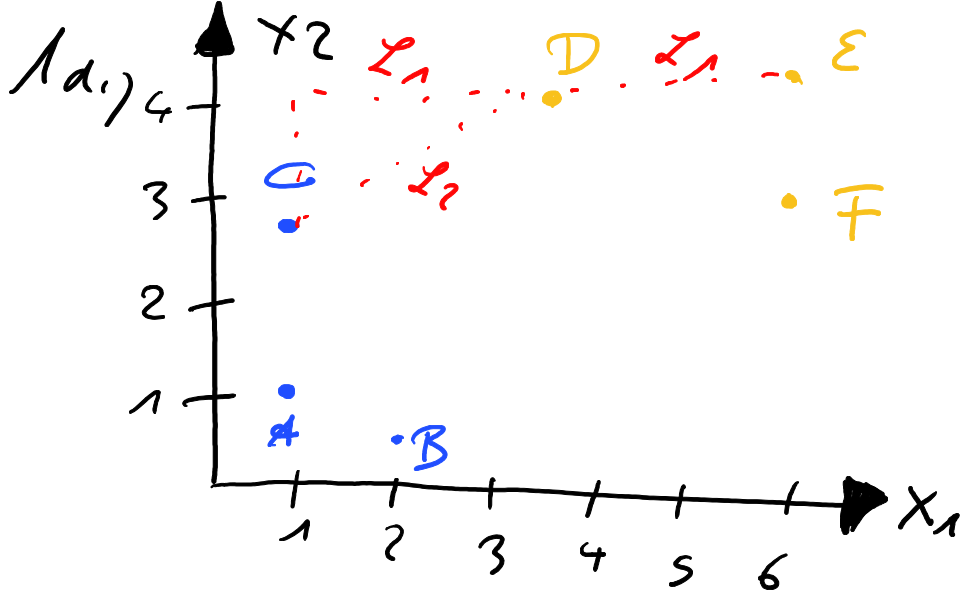
Tutorial 2: kNN & Decision Trees

1NN

- given: training set $D = \{(x_i, y_i)\}_{i=1}^{n}$
- want: classify x_{new}
  - → distance measure: $L_1, L_2, \ldots$
  - → compute nn
  - → label $\hat{=}$ label of nn
- kNN: k neighbours





$$d_1(A,B) = \sum_i^n |A_i - B_i|$$

$$= |1-2| + |1-0.5|$$

$$= \underline{\underline{1.5}}$$

$\mathcal{L}_1$	A	B	C	D	E	F	nn
A		1.5	1.5				B/C ✓
B	1.5						A ✓
C	1.5						A ✓
D					2.5		E ✓
E						1.0	F ✓
F					1.0		E ✓

1.b.) $d_2(A,B) = \left(\sum_i^n (A_i - B_i)^2\right)^{\frac{1}{2}}$

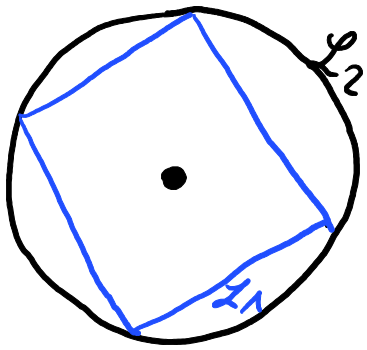
$$= \sqrt{(1-2)^2 + (1-0.5)^2}$$

$$\approx \underline{\underline{1.12}}$$

$\mathcal{L}_2$	A	B	C	D	E	F	nn
A		1.12					B ✓
B	1.12						A ✓
C	1.5						A ✓
D			2.24				C ✗
E						1.0	F ✓
F					1.0		E ✓

1c.) • different distance measure → different nn, label

• $\mathcal{L}_2 \leq \mathcal{L}_1$



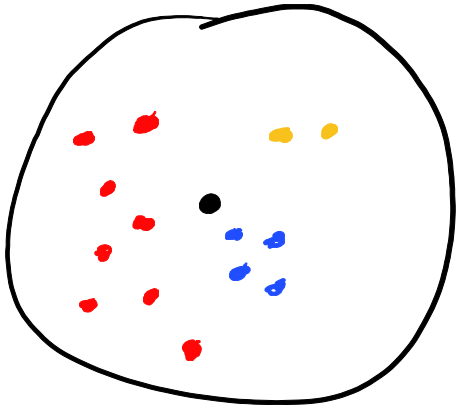
2.) A, B, C

$N_A = 16$

$N_B = 32$

$N_C = 64$

$k = N_A + N_B + N_C$

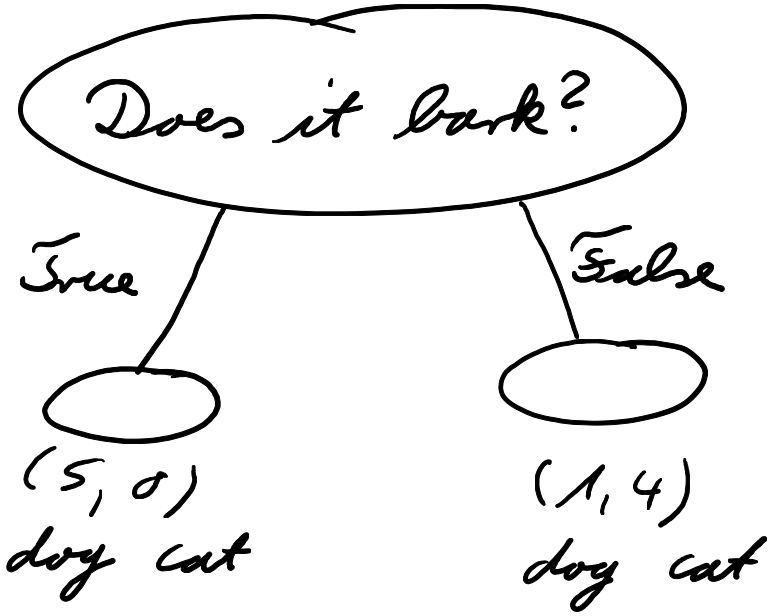


→ label: C

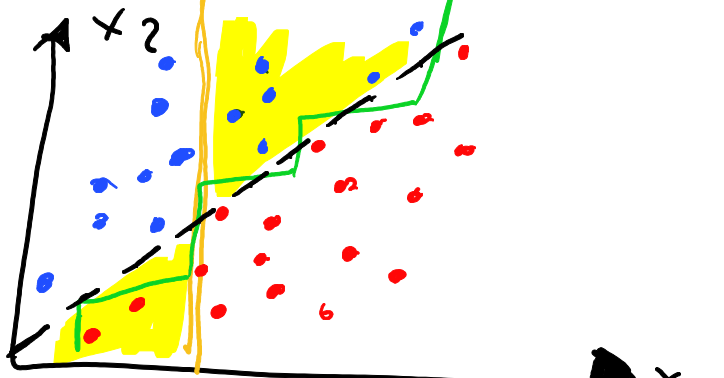
2b.) not enough information

Decision trees

- node $\hat{=}$ feature test
  - → leads to decision boundary
- branch $\hat{=}$ outcome
- leaf $\hat{=}$ region in input space and distribution of samples in region
  - → partitions space in cuboid regions
- pure class distribution
  - → misclassification rate
  - → entropy
  - → Gini index



3.)



→ No

4.) x_1 : T, I

x_2 : M, P

x_3 : S, C

→ W, L

$N = 10$

W = 4

L = 6

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

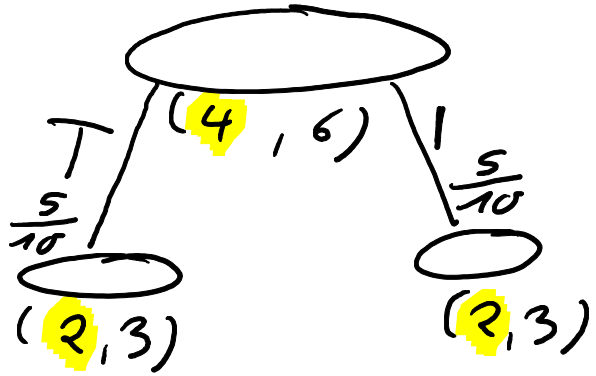
$P(y = L) = \frac{6}{10}$

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

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

≈ 0.97

4b.) Split, test x_1



$P(x_1 = T) = \frac{5}{10}$

$P(x_1 = F) = \frac{5}{10}$

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

$P(y = W \mid x_1 = F) = \frac{2}{5}$, $P(y = L \mid x_1 = F) = \frac{3}{5}$

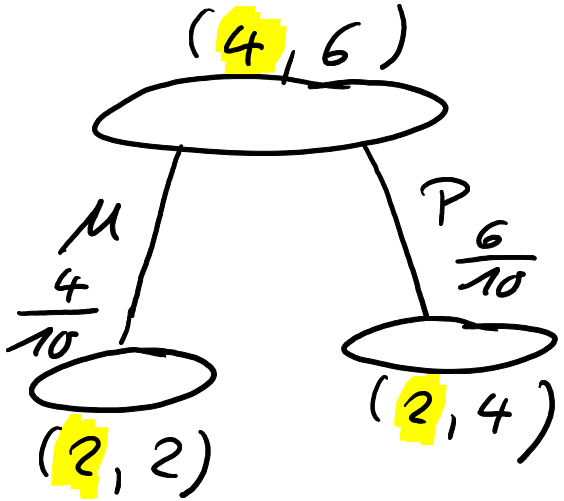
$i_H(x_1 = T) = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5}$

≈ 0.97

$i_H(x_1 = F) = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} \approx 0.97$

$\Delta(x_1) = 0.97 - \frac{1}{2} \cdot 0.97 - \frac{1}{2} \cdot 0.97 = \underline{\underline{0}}$

Split : test x_2 :



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

$P(x_2 = P) = \frac{6}{10}$

$P(y = W \mid x_2 = M) = \frac{2}{4}$

$P(y = L \mid x_2 = M) = \frac{2}{4}$

$P(y = W \mid x_2 = P) = \frac{2}{6}$

$P(y = L \mid x_2 = P) = \frac{4}{6}$

$i_H(x_2 = M) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1.0$

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

$\Delta(x_2) = 0.97 - \frac{4}{10}1.0 - \frac{6}{10}0.92$

≈ 0.016

Split: test x_3:

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

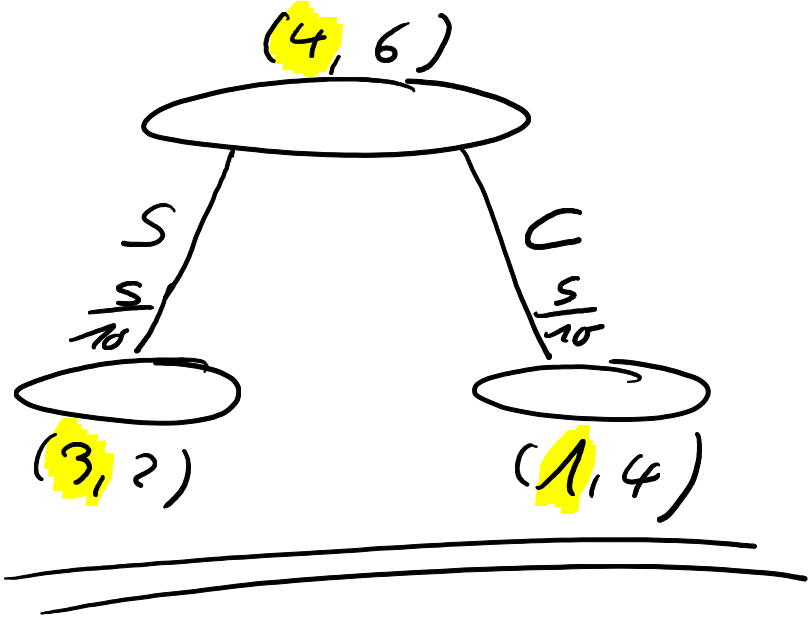
$P(x_3 = C) = \frac{5}{10}$

$P(y = W \mid x_3 = S) = \frac{3}{5}$

$P(y = L \mid x_3 = S) = \frac{2}{5}$

$P(y = W \mid x_3 = C) = \frac{1}{5}$

$P(y = L \mid x_3 = C) = \frac{4}{5}$



$i_H(x_3 = S) \approx 0.97$

$i_H(x_3 = C) \approx 0.72$

$\Delta(x_3) = 0.97 - \frac{5}{10} 0.97 - \frac{5}{10} 0.72$

$= \underline{0.125}$

→ Split on x_3