Decisions trees

1.1) ID3

Find the feature with the most Information gain:

Feature	$\frac{\lvert V_{a=T}\rvert}{\lvert V\rvert}$	$\frac{\lvert V_{a=F}\rvert}{\lvert V\rvert}$	$H(V_{a=T})$	$H(V_{a=F})$	$IG(V,a)-H(V)$
A	$3/4$	$1/4$	$H(2/3)$	$H(0/1)$	$-\frac{3}{4}H(\frac{2}{3})-\frac{1}{4}H(\frac{0}{1})$
B	$2/4$	$2/4$	$H(1/2)$	$H(1/2)$	$-\frac{2}{4}H(\frac{1}{2})-\frac{2}{4}H(\frac{1}{2})$
C	$2/4$	$2/4$	$H(1/2)$	$H(1/2)$	$-\frac{2}{4}H(\frac{1}{2})-\frac{2}{4}H(\frac{1}{2})$

$$IG(V,a)-H(V) = -\frac{\lvert V_{a=T}\rvert}{\lvert V\rvert}\cdot H(V_{a=T}) - \frac{\lvert V_{a=F}\rvert}{\lvert V\rvert}\cdot H(V_{a=F})$$

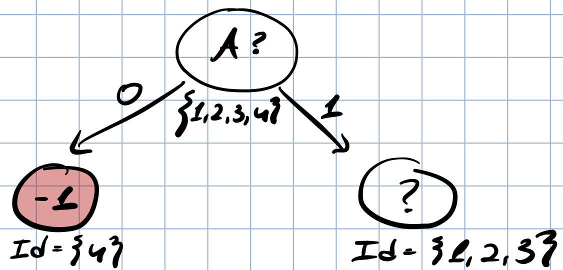
- **A** $-\frac{3}{4}H(\frac{2}{3})-\frac{1}{4}H(\frac{0}{1}) = -\frac{3}{4}\cdot 0.918 - 0 = -0.688$

$$H(\tfrac{2}{3}) = -\tfrac{2}{3}\cdot \log_2(\tfrac{2}{3}) - \tfrac{1}{3}\log_2(\tfrac{1}{3}) = 0.918$$

$$H(0) = -\log(1) = 0$$

- B, C: $-H(\frac{1}{2}) = -1$

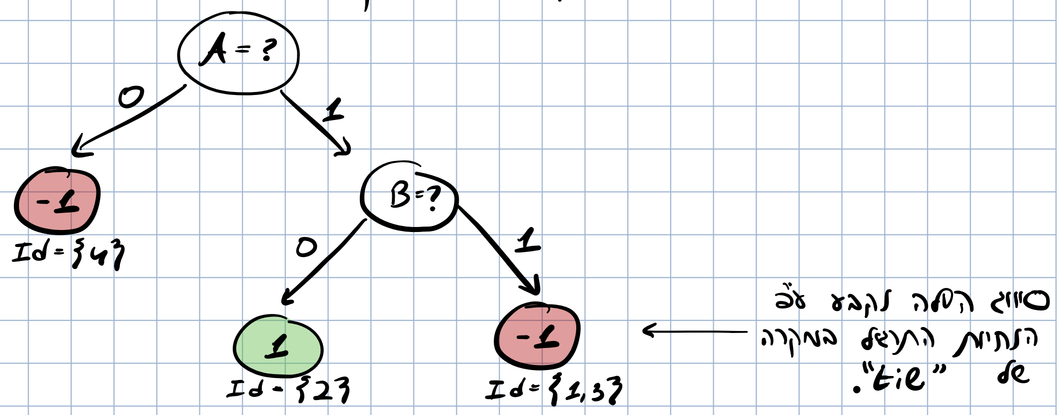
לכן $\underset{(A)}{-0.688} > \underset{(B,C)}{-1}$ נבחר בפיצ'ר A לחלוקה הראשונית



- בעלה A=0 ישנו Sample בודד ולכן לא נחלק אותו יותר.
- נבחר פיצ'ר לחלוקת העלה 1 =A:

Feature	$\frac{\lvert V_{a=T}\rvert}{\lvert V\rvert}$	$\frac{\lvert V_{a=F}\rvert}{\lvert V\rvert}$	$H(V_{a=T})$	$H(V_{a=F})$	$IG(V,a)-H(V)$
B	$2/3$	$1/3$	$H(\frac{1}{2})$	$H(0)$	$-\frac{2}{3}H(\frac{1}{2})-\frac{1}{3}H(0)$
C	$1/3$	$2/3$	$H(1)$	$H(\frac{1}{2})$	$-\frac{1}{3}H(1)-\frac{2}{3}H(\frac{1}{2})$

הפיצול לפי הפיצ'רים B ו-C שקול עבור העלה הנ"ל. מאחר ועומק העץ המקס' הוא 2, אין עוד חלוקות בהמשך ולכן אין חשיבות לבחירת הפיצ'ר [על פי הזמן שניתן לנו, כמובן]. נבחר אקראית בפיצ'ר B ונקבל את העץ הבא:

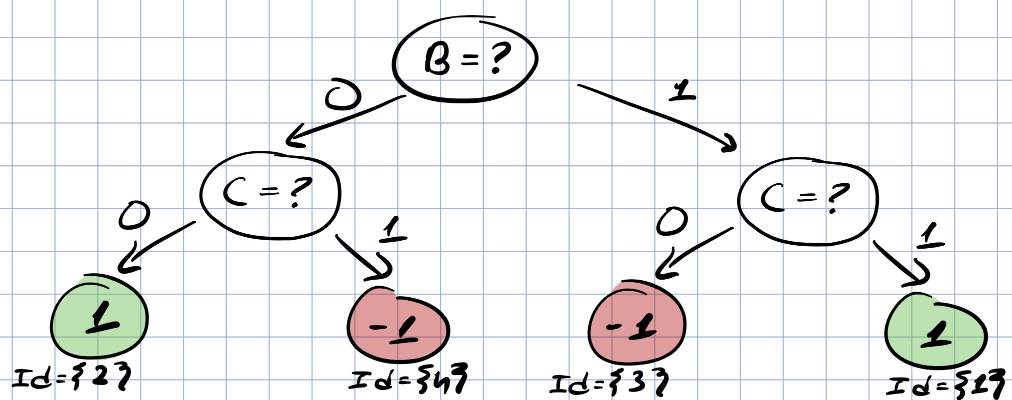


Training Error: $\frac{\#incorrect\_predictions}{\#all\_predictions} = \frac{1}{4}$

MisClassified tuples: (1, 1)

(sample's Id, sample's label)

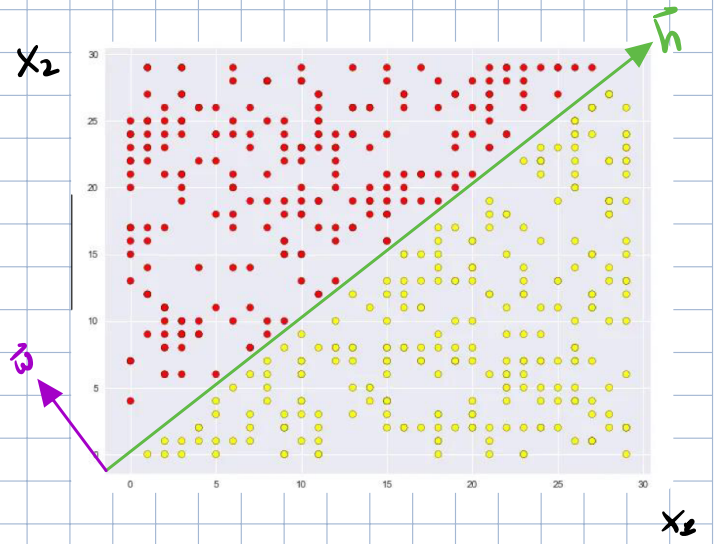
1.2



Decision trees strike again

2.1) $h(X_1, X_2) = \text{Sign}(W^T X + b)$

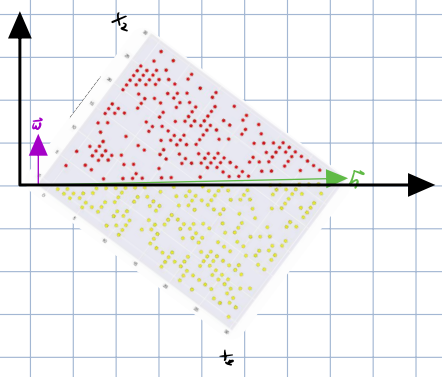
Yes. There are linear classifiers that achieve perfect Accuracy on the dataset. For example we can look at the following $\vec{h}$ defined by $\vec{w}$.



2.2) **No.** Decision trees with depth = 1 only split the dataset once with a single feature. Looking at the given dataset, we can see that there is no $\Theta_1 \in X_1$ or $\Theta_2 \in X_2$ s.t the dataset is seperable solely by this value.

2.3) **No,** for the same logic mentioned above. To simplify the explanation, we could look at $\Theta_1 \in X_1$ as a vertical line in the graph, and $\Theta_2 \in X_2$ as a horizontal line. We can see that there are no 2 values $\in \Theta_1 \cup \Theta_2$ for which the data is seperable. (=NO 2 vertical/horizontal lines)

2.4) **Yes.** Rotating the data 45° clockwise will make it seperable by a horizontal line. Illustration:



Information Gain

3.1) The problem with the chatGPT proof is an incorrect mathematical transition from the assumptions $\{Entropy(D) \geq 0,\ Entropy(D_v) \geq 0\}$ to the conclusion that the substraction $Entropy(D) - \sum_v \frac{|D_v|}{|D|} \cdot Entropy(D_v)$ is also non negative.

3.2) Prove that $IG(V,a) = H(V) - \frac{|V_{a=T}|}{|V|} H(V_{a=T}) - \frac{|V_{a=F}|}{|V|} H(V_{a=F}) \geq 0$

Given that $H(v) = H(P_v) = -P_v \log_2(P_v) - (1-P_v)\log_2(1-P_v)$ holds

$\forall \beta_1, \beta_2, \alpha \in [0,1]$: $-\alpha H(\beta_1) - (1-\alpha) H(\beta_2) \geq -H(\alpha\beta_1 + (1-\alpha)\beta_2)$

$IG(V,a) =$ / definition of IG

$H(V) - \frac{|V_{a=T}|}{|V|} H(V_{a=T}^{\beta_1}) - \frac{|V_{a=F}|}{|V|} H(V_{a=F}^{\beta_2}) \geq$ / Given Property

$H(V) - H\left(\frac{|V_{a=T}|}{|V|} \cdot P_{V_{a=T}} + \frac{|V_{a=F}|}{|V|} \cdot P_{V_{a=F}}\right) =$ / definition of $P_{V_{a=T}}$ and $P_{V_{a=F}}$

$H(V) - H\left(\frac{|V_{a=T}|}{|V|} \cdot \frac{|\{(x,y) \in V_{a=T} \mid y=1\}|}{|V_{a=T}|} + \frac{|V_{a=F}|}{|V|} \cdot \frac{|\{(x,y) \in V_{a=F} \mid y=1\}|}{|V_{a=F}|}\right) =$

$H(V) - H\left(\frac{|\{(x,y) \in V_{a=T} \mid y=1\}| + |\{(x,y) \in V_{a=F} \mid y=1\}|}{|V|}\right) =$ / $V_{a=T} \cup V_{a=F} = V$, $V_{a=T} \cap V_{a=F} = \emptyset$

$H(V) - H\left(\frac{|\{(x,y) \in V \mid y=1\}|}{|V|}\right) =$ / definition of P_v

$H(V) - H(P_v) = 0$

JLV

Linear Classification

- For homogeneous & $h(x) = +1 \iff w^T x = \underbrace{\|w\| \|x\|}_{>0} \cos\sphericalangle(w,x) > 0 \iff \cos\sphericalangle(w,x) > 0$

  invariant to the scale of w.

- The geometric (signed) margin of $x \in \mathbb{R}^d$ is $\frac{w^T x}{\|w\|}$

- For non-homogeneous & $h(x) = +1 \iff w^T x \geq -b$

  b being almost equivilent to the margin.

4.1) In the homogeneous case we had a degree of freedom in scaling w. For the non-homogeneous case we still have a degree of freedom as long as we scale b along with w. We can consider the classifier to be a $w' \in \mathbb{R}^{d+1}$ vector s.t $w' = \begin{pmatrix} w \\ b \end{pmatrix}$ and reframe the classification as such& $h = +1 \iff w'^T \begin{pmatrix} x \\ 1 \end{pmatrix} > 0$.

In that case we get the same classifier, and we have an homogeneous equation which is invariant to the scale of w!

4.2) A is b_2, B is b_4, C is b_2, D is b_3.

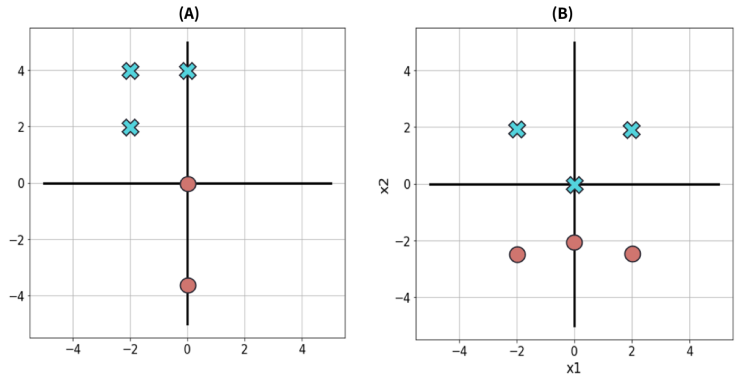
K-Nearest Neighbors

- A training point is not considered a neighbor of itself.

i) $k=1$, $d(u,v) = \|u-v\|_2$

ii) $k=3$, $d(u,v) = \|u-v\|_2$

iii) $k=1$, $d(u,v) = \|u-v\|_1$



5.1)

model \ dataset	A	B
i (k=1, L2)	NO. The red point on the axes intersection is misclassified as blue.	NO. The blue point on the axes intersection is misclassified as the red below it and vice versa.
ii (k=3, L2)	NO. red points will be blue as there could only be 1 red neighbor an 2 blue ones.	Yes
iii (k=1, L1)	Yes	NO. blue point on intersection will be classified by the red point below it and vice versa.

5.2)

- Answers unchanged for models (i) & (ii).

In these models the distance is calculated by the L2 norm which is an euclidian distance. This norm is invariant to rotations. [Im not sure if calculations are considered 'brief' enough, so added below, ignore if not needed). If the distance calculation is unchanged, so is the classifier.

- For model (iii) answers might change.

This model uses the L1 norm as a distance metrics, and it is affected by rotation as the distance is affected by the points relation to the axes.

- For dataset A: rotation could decrease the L2 distance between the (0,0), (2,2) points s.t they'll be nearest neighbors.

- For dataset B: a (45°) rotation could lead to a tie in the distances from the (0,0) point. ($2\sqrt{2}$)

---

some calculations for 5_2)

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad X' = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1\cos\theta - X_2\sin\theta \\ X_1\sin\theta + X_2\cos\theta \end{pmatrix}$$

$$\|X\|_2 = \sqrt{X_1^2 + X_2^2}$$

$$\|X'\|_2 = \sqrt{(X_1\cos\theta - X_2\sin\theta)^2 + (X_1\sin\theta + X_2\cos\theta)^2}$$

$$= \sqrt{X_1^2\cos^2\theta + X_2^2\sin^2\theta - 2X_1X_2\cos\theta\sin\theta + X_1^2\sin^2\theta + X_2^2\cos^2\theta + 2X_1X_2\sin\theta\cos\theta}$$

$$= \sqrt{X_1^2(\cos^2\theta + \sin^2\theta) + X_2^2(\sin^2\theta + \cos^2\theta)} = \sqrt{X_1^2 + X_2^2} = \|X\|_2$$