University of Southern Denmark
IMADA

DM566: Data Mining and Machine Learning
Spring term 2022

Exercise 10

Exercise 10-1 Information Gain
In this exercise, we want to look more closely at the information gain measure.

Let T be a set of n training objects with the attributes $A_1, \dots, A_a$ and the k classes $c_1, \dots, c_k$.

Let $\{T_i^A | i \in \{1, \dots, m_A\}\}$ be the disjoint, complete partitioning of T produced by a split on attribute A (where m_A is the number of disjoint values of A).

(a) *Uniform distribution*
Compute $entropy(T)$, $entropy(T_i^A)$ for $i \in \{1, \dots, m_A\}$ as well as $gain(T, A)$ given the assumption that the class membership of T is uniformly distributed and independent of the values of A. Interpret your result.
Suggested solution:
Independent uniform distribution

$$
\begin{aligned}
p_i &= \frac{1}{k} \forall 1 \leq i \leq k \\
|T_i^A| &= \frac{1}{m_A} \cdot |T| \\
entropy(T) &= -\sum_{i=1}^{k} p_i \log p_i \\
&= -\log \frac{1}{k} \\
&= \log k \\
entropy(T_i^A) &= \log k \\
information\text{-}gain(T, A) &= entropy(T) - \sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} \cdot entropy(T_i^A) \\
&= \log k - m_A \cdot \frac{1}{m_A} \cdot \log k \\
&= 0
\end{aligned}
$$

Interpretation: expected - a split on this attribute should not help.

(b) *Additional uniform distribution*
We want to analyze how the number of different values influences the information gain. For this, we want to compare two attributes, attribute A with m_A values and attribute M' with $m_{A'} = m_A + 1$ values, where the relative frequencies in A' in values 1 to m_A are identical to that of A and in the additional value $m_{A'}$ there is a uniform distribution of the classes. How does $gain(T, A)$ differ from $gain(T, A')$? Interpret your result.
Suggested solution:

$$information\text{-}gain(T, A) = entropy(T) - \sum_{i=1}^{m_A} \frac{|T_i^A|}{|T|} \cdot entropy(T_i^A)$$

$$\begin{aligned} & information\text{-}gain(T, A') \\ &= entropy(T) - \sum_{i=1}^{m_A+1} \frac{|T_i^{A'}|}{|T|} \cdot entropy(T_i^{A'}) \\ &= entropy(T) - \frac{1}{|T|} \left[\sum_{i=1}^{m_A} |T_i^{A'}| \cdot entropy(T_i^{A'}) + |T_{m_A+1}^{A'}| \cdot entropy(T_{m_A+1}^{A'}) \right] \end{aligned}$$

$$\begin{aligned} entropy(T_i^A) &\leq \log k \\ entropy(T_{m_A+1}^{A'}) &= \log k \quad \text{(uniformly distributed, maximal entropy)} \end{aligned}$$

Interpretation:

- In comparison to T^A, for each data object in $T_{m+1}^{A'}$, we add once $\log k$ and subtract a value $\leq \log k$, i.e., altogether we add some value ≤ 0.
- Therefore the information gain must be smaller for A' compared to A.
- Thus, A would be preferable over A' for the split, which makes also sense intuitively.

(c) *Attributes with many values*
Let A be an attribute with random values, not correlated to the class of the objects. Furthermore, let A have enough values, s.t. not any two instances of the training set share the same value of A. What happens in this situation when building the decision tree? What is problematic with this situation?
Suggested solution:
Split on A: Entropy in each branch is 0, as we have pure class sets (in each case, some $p_i = 1$, all others $p_{j(j \neq i)} = 0$).

$$information\text{-}gain(T, A) = entropy(T) - 0 \quad \text{(maximal!)}$$

Hence we choose A as root and the tree is done.

Exercise 10-2 Neurons

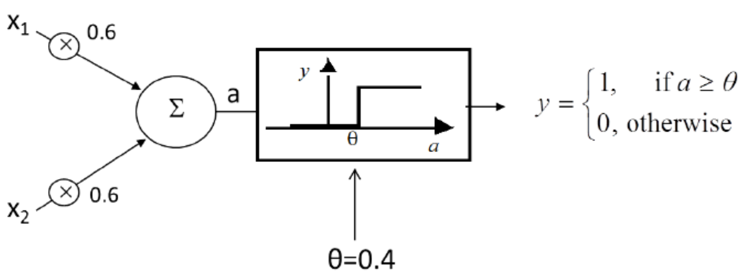
Sketch two trained threshold logic units (that is, individual TLUs, no hidden layer) that can represent for two Boolean variables $x_1, x_2 \in \{0, 1\}$ and the AND and the OR function, respectively.

Sketch the related linear separations in the boolean space.

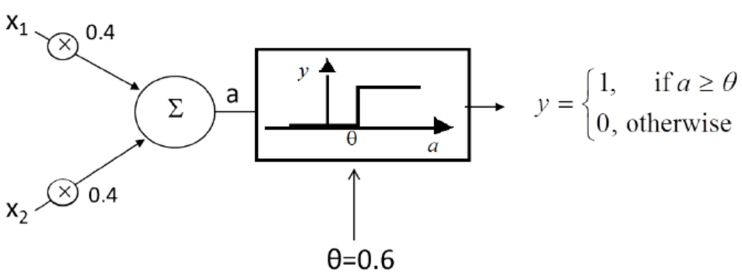
Suggested solution:

Diverse solutions are possible, examples are:

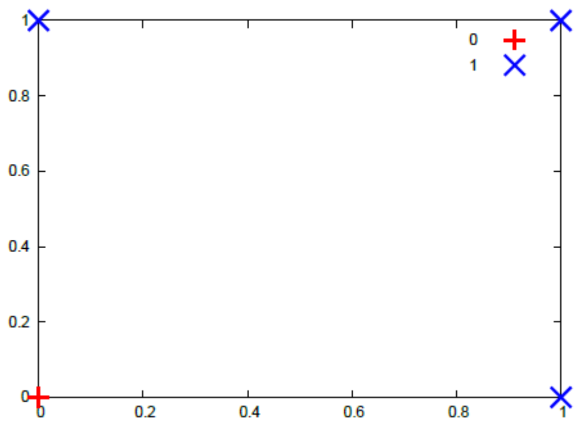
$x_1 \vee x_2$:



$x_1 \wedge x_2$:



Sketch the related linear separations in the boolean space



(example for OR)

for the equations

$$\langle (w_1, w_2), (x_1, x_2) \rangle - \theta = 0$$

We have for $w_1 = w_2 = 0.6$, $\theta = 0.4$:

$$x_2 = \frac{2}{3} - x_1$$

and for $w_1 = w_2 = 0.4$, $\theta = 0.6$:

$$x_2 = 1.5 - x - 1$$

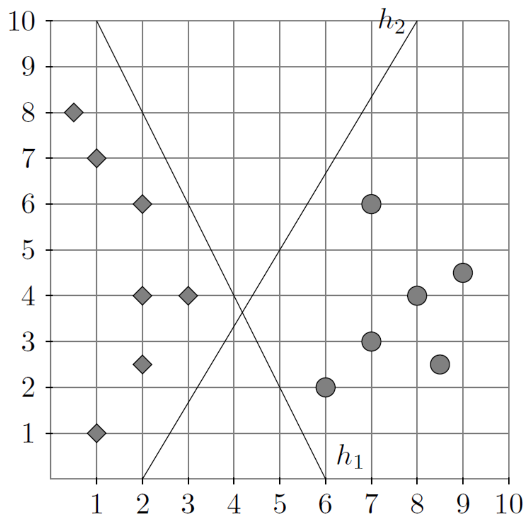
Or in general:

$$x_2 = \frac{\theta}{w_2} - \frac{w_1}{w_2} x_1$$

We have infinitely many other possibilities for separating lines, relating to other weights and thresholds.

Exercise 10-3 Support vectors and margin

Consider the following dataset with points from two classes c_1 (diamonds) and c_2 (circles).



(a) Give the equations for hyperplanes h_1 and h_2.

Suggested solution:

We can start with two points that define a line (i.e., a hyperplane in the two dimensional space). For h_1, we can use (e.g.) (6,0) and (1,10). Thus the slope is

$$m_1 = \frac{10}{1-6} = -2$$

Now using (6,0) as a point of the line, we get the equation:

$$\frac{x_2 - 0}{x_1 - 6} = -2 \Longrightarrow 2x_1 + x_2 - 12 = 0$$

For h_2 we can take $(2, 0)$ and $(8, 10)$, thus the slope is

$$m_2 = \frac{10}{8-2} = \frac{5}{3}$$

With $(2, 0)$ as a point of the line, we get the equation

$$\frac{x_2 - 0}{x_1 - 2} = \frac{5}{3} \Longrightarrow 5x_1 - 3x_2 - 10 = 0$$

(b) Name all the support vectors for h_1 and h_2.
Suggested solution:
The support vectors for h_1 are $(2, 6), (3, 4)$, and $(6, 2)$.
The support vectors for h_2 are $(3, 4)$ and $(7, 6)$.

(c) Which of the two hyperplanes is better at separating the two classes based on the margin?
Suggested solution:
We compute the margins for the two classifiers by computing the distance from the support vectors to the hyperplanes.
For h_1: $2x_1 + x_2 - 12$ we have $w = (2, 1)$ and $b = -12$, such that $H_1 = \langle w, x\rangle + b$.
Recall that the distance of a point $x = (x_1, x_2)$ to the hyperplane is $\frac{\langle w,x\rangle + b}{||w||}$.
The distance of support vector $(3, 4)$ is thus:

$$\frac{6 + 4 - 12}{\sqrt{2^2 + 1^2}} = \frac{-2}{\sqrt{5}}$$

The distance of support vector $(6, 2)$ is:

$$\frac{12 + 2 - 12}{\sqrt{2^2 + 1^2}} = \frac{2}{\sqrt{5}}$$

The total margin is therefore $2\frac{2}{\sqrt{5}} \approx 1.79$

For h_2: $5x_1 - 3x_2 - 10$ we have $w = (5, -3)$ and $b = -10$.
The distance of support vector $(3, 4)$ is:

$$\frac{15 - 12 - 10}{\sqrt{5^2 + 3^2}} \approx \frac{-7}{5.83}$$

The distance of support vector $(7, 6)$ is:

$$\frac{35 - 18 - 10}{\sqrt{5^2 + 3^2}} \approx \frac{7}{5.83}$$

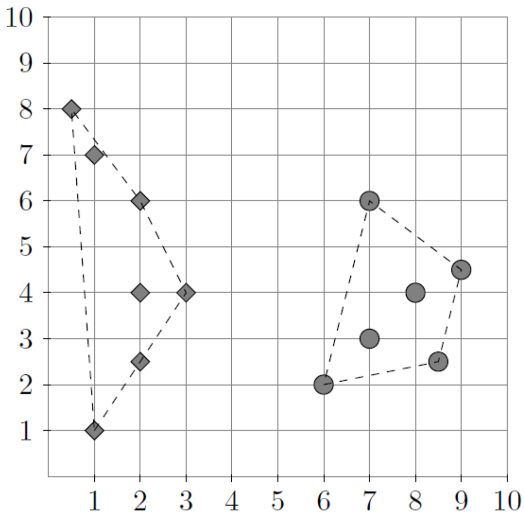
The total margin is therefore $\frac{14}{\sqrt{34}} \approx 2.4$.

In conclusion, h_2 is better than h_1.

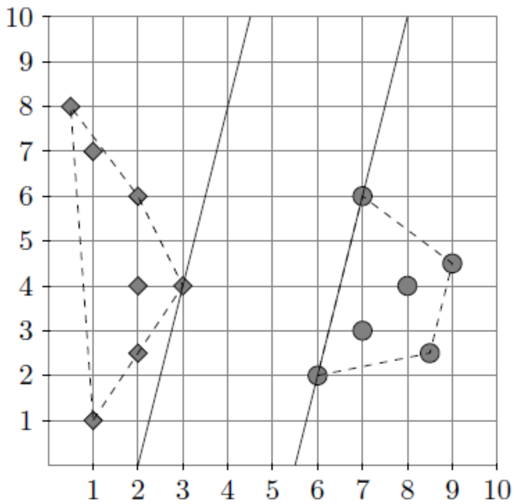
(d) Find the best separating hyperplane for this dataset, give its equation, and show the corresponding support vectors.

Suggested solution:

Sketch the convex hull:



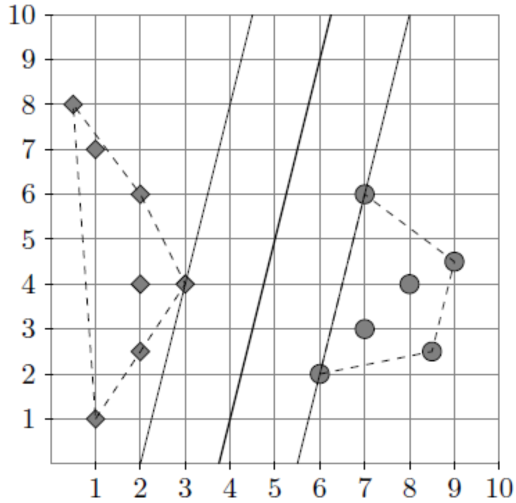
Sketch the maximum possible margin, obvious by the convex hull.



The support vectors are thus $(6,2)$ and $(7,6)$ for circles and $(3,4)$ for diamonds. The optimal hyperplane is therefore

$$h : 4x_1 - x_2 - 15$$

which is exactly half-way between the lines passing through the support vectors (which are $4x_1 - x_2 - 22$ and $4x_1 - x_2 - 8$):



The margin is $\frac{14}{\sqrt{17}} \approx 3.395$.