Question 1. (5 points) The following data have been collected and we want to learn the general concept *Acceptable*, by using Decision Tree Learning.

House	Furniture	Nr rooms	New kitchen	Acceptable
1	No	3	Yes	Yes
2	Yes	3	No	No
3	No	4	No	Yes
4	No	3	No	No
5	Yes	4	No	Yes

1. Formalize the learning problem.
2. Describe how variables are chosen when building a Decision Tree.
3. Execute the ID3 algorithm on the data set above and generate the corresponding output tree.

Note: point 3 can be answered even if point 2 is not properly addressed, by using any invented method (or invented numbers) for the selection of the variables.

The target function is $f = \overset{\text{Furniture}}{\{0,1\}} \times \overset{\text{rooms}}{\mathbb{N}} \times \overset{\text{new Kitchen}}{\{0,1\}} \to \overset{\text{acceptable}}{\{0,1\}}$

The entropy of a dataset is $-p_\oplus \ln p_\oplus - (1-p_\oplus)\ln(1-p_\oplus)$ where $p_\oplus$ = ratio of samples labeled as 1.

IF D is the set and A an attribute of the input, we have:

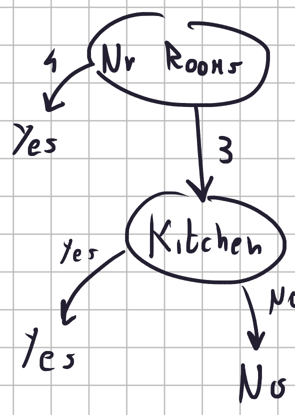
$$GAIN(D,A) = ent(D) - \sum_{v \in A} \frac{|D_v|}{|D|} ent(D_v) \quad \text{where } D_v = \{\text{samples with } A = v\}$$

We choose the attribute that maximize the gain. For this problem:

$p_\oplus$ of D is $3/5$ so $ent(D) = 0.673$

Att : Furniture. $\begin{cases} |D_0| = 3 & en(D_0) = 0.636 \\ |D_1| = 2 & en(D_1) = 0.693 \end{cases} \Rightarrow GAIN = 0.0142$

Att : Nr Rooms $\begin{cases} |D_3| = 3 & en(D_3) = 0.636 \\ |D_4| = 2 & en(D_4) = 0 \end{cases} \Rightarrow GAIN = 0.2914$ $\Rightarrow$



Att : Kitchen $\begin{cases} |D_0| = 4 & en(D_0) = 0.693 \\ |D_1| = 1 & en(D_1) = 0 \end{cases} \Rightarrow GAIN : 0.1186$

Question 2. (5 points)

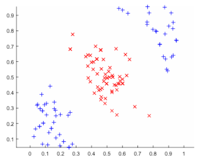
1. What is a maximum likelihood (ML) hypothesis?
2. Comment the following statement: in a classification problem, the class returned by the ML hypothesis on a new instance x is always the most probable class.

Let $h^* = \underset{h}{\arg\max}\, \mathbb{P}(h|D)$, we can rewrite $h^* = \underset{h}{\arg\max}\, \mathbb{P}(D|h)\mathbb{P}(h)$. IF we assume that the hypothesis have the same probability, we get the ML hypothesis : $h_{ML} = \underset{h}{\arg\max}\, \mathbb{P}(D|h)$

The ML doesn't always return the most probable class, since the hypothesis usually have different probabilities $\mathbb{P}(h)$. The most probable is given by the Bayes Optimal Classifier : $c^* = \underset{c}{\arg\max} \sum_h \mathbb{P}(c|x,h)\mathbb{P}(h|D)$

Question 3. (5 points) Consider classification of the data shown in the figure below by using SVMs.

1. Explain which kernel function you would choose to obtain perfect separability.
2. How SVMs classification can be applied if the data are note perfectly separable?



We notice how the red samples are the one closer to $(1/2 \;\; 1/2)^T$, so i consider a kernel that compares the distances from that point.

$$K(x, x') = \big|\, \|x - c\| - \|x' - c\| \,\big| \quad \text{where} \quad c = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

If the dataset is not perfectly separable, i add some slack variables ξ_i

$$\begin{aligned} &\operatorname{argmin} \|w\|^2 \\ &\text{s.t. } t_i y(x_i) \geq 1 \;\; \forall i \end{aligned} \quad \text{becomes} \quad \begin{aligned} &\operatorname{argmin} \|w\|^2 + \gamma \sum_i \xi_i \quad \leftarrow \gamma \in \mathbb{R} \\ &\text{s.t. } t_i y(x_i) \geq 1 - \xi_i \;\; \forall i \\ &\xi_i > 0 \;\; \forall i \end{aligned}$$

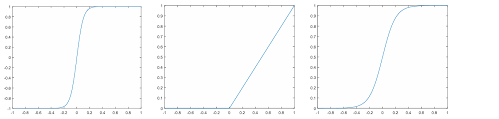
Question 4. (5 points) Let $x \in \mathcal{X} = \{(1.8, 2.1)^T, (3.2, 1.4)^T, (2.5, 4.1)^T, (1.6, 7.7)^T, (3.1, 9.1)^T\}$ and $t \in T = \{0, 1, 1, 0, 1\}$. Consider designing an artificial neural network for learning the function $t = f(x, \theta)$.

1. Explain what is a suitable activation function for the output layer of the network.
2. For the selected activation function explain if the output units saturate and how learning the parameters of the network is affected if this is the case.

Since it's binary class. For the output we consider the sigmoid function (in this way the model is differentiable) and we predict by considering $y(x) = \text{Step}(f(x, \theta))$. In this way the output unit saturate only when the correct answer is given.

Question 5. (5 points)

1. Consider an image of $w_{im} \times h_{im}$ elements (pixels) and a convolution kernel of dimension $3 \times 3 \times 16$. What are the possible values of the stride and padding in order to convolve without skipping any pixels, while obtaining a feature map with the same dimensions with the input image.
2. Associate the correct name of the activation function to the plots above and provide the corresponding mathematical definition. The list of names is {*ReLU, Sigmoid, Hyperbolic Tangent*}.

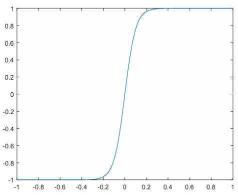


the following relations holds:

$$W_{out} = \frac{W_{in} - W_K + 2p}{s} + 1 \qquad \text{in our case} \begin{cases} W_{im} = \dfrac{W_{im} - 3 + 2p}{s} + 1 \\ h_{im} = \dfrac{h_{im} - 3 + 2p}{s} + 1 \end{cases} \begin{cases} p = \frac{1}{2}\left(W_{im}(s-1) - s + 3\right) \\ s = \dfrac{h_{im} - 3 + 2p}{h_{im} - 1} \end{cases}$$

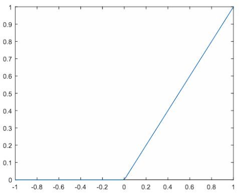
$$h_{out} = \frac{h_{in} - h_K + 2p}{s} + 1$$

$$\begin{cases} p = \frac{1}{2}\left(W_{im}(s-1) - s + 3\right) \\ s = \dfrac{1}{h_{im} - 1}\left(h_{im} - 3 + W_{im}(s-1) - s + 3\right) \end{cases} \Rightarrow \begin{cases} p = 1 \\ s = 1 \end{cases}$$



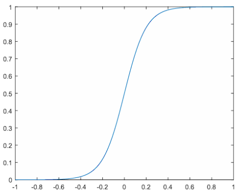
tanh

$2\sigma(2x) - 1$



ReLU

$\max(0, x)$

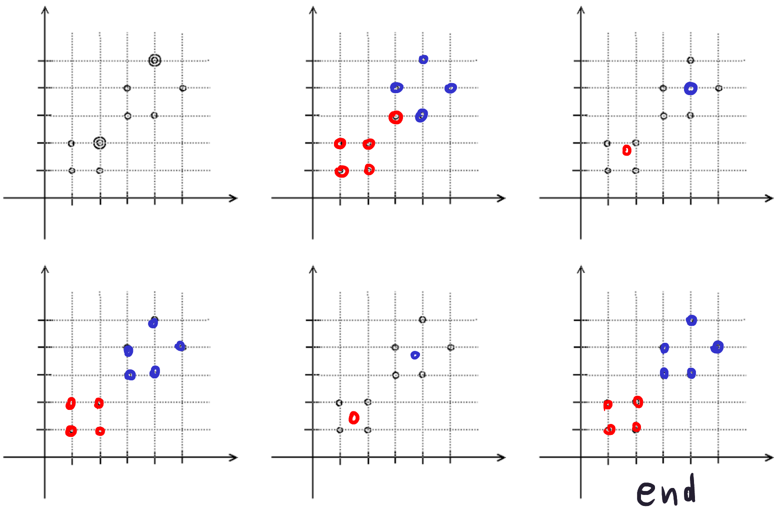


Sigmoid

$\sigma(x) = \dfrac{1}{1 + \exp(-x)}$

Question 6. (5 points)

Simulate the execution of K-means in this 2-D data set with k=2 and initial centroids indicated by double circles: use one diagram for each step of the algorithm. Describe explicitly how each step is obtained and what is the termination condition of the algorithm. Drawing only the steps is not sufficient.



at each steps we divide the samples in two groups (group 1: colser to centroid 1, group 2: the others) and then recompute the centroids by considering the center of mass of the groups.

we stop when the new centroids computed are equals to the one of the previous step.