EXERCISE 1

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: decribe exactly the target function to learn and the dataset.
2. Describe qualitatively how attributes are chosen when building a Decision Tree.
3. Simulate the execution of ID3 algorithm on the data set above and generate the corresponding output tree.

The target function is $\delta : \overbrace{\underset{\{Yes, No\}}{\underset{\shortparallel}{\text{Furniture}}} \times \underset{\mathbb{N}}{\underset{\cap}{\text{Nr rooms}}} \times \underset{\{Yes, No\}}{\underset{\shortparallel}{\text{New Kitchen}}}}^{X} \longmapsto \{Yes, No\}$

So the Dataset is $D = \{(x_i, t_i)\}_{i=1}^{5}$, $x_i \in X$, $t_i \in \{Yes, No\}$.

We know that the entropy of a binary dataset with $p^{\oplus}$ = ratio of "Yes" samples is : $-p^{\oplus} \ln p^{\oplus} - (1-p^{\oplus}) \ln (1-p^{\oplus})$. Let A to be an attribute and $v \in A$ a value. $D_v = \{(x_i, t_i) \in D : A(x_i) = v\}$. We choose the attributes that maximize the information gain: $\text{Gain}(D, A) = \text{entropy}(D) - \sum_{v \in A} \frac{|D_v|}{|D|} \text{entropy}(D_v)$

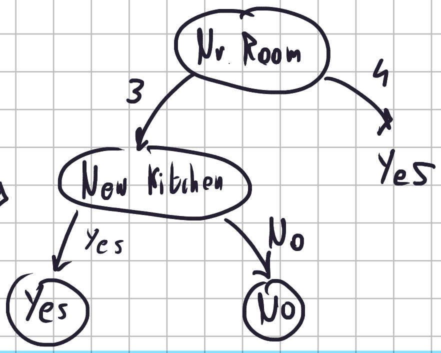
For the given dataset, we consider

$$\text{Gain}(D, \text{Furniture}) = 0.673 - \left(\frac{2}{5} 0.69 + \frac{3}{5} \cdot 0.63\right) = 0.019$$

$$\text{Gain}(D, \text{Nr Room}) = 0.673 - \frac{2}{5} \cdot 0.63 = 0.421$$

$$\text{Gain}(D \text{ New Kit}) = 0.673 - \frac{4}{5} 0.69 \simeq 0.121$$

⇒



EXERCISE 2

1. Provide a formal definition of 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.*

Given a dataset, the maximum at posteriori hypothesis is

$$h_{MAP} = \underset{h}{\text{argmax}}\ \mathbb{P}(h \mid D) = \underset{h}{\text{argmax}}\ \mathbb{P}(D|h)\mathbb{P}(h)$$

If we assume that the hypothesis are uniformly distributed, i.e. $\mathbb{P}(h)$ constant, then we get the maximum likelihood hypothesis : $h_{ML} = \underset{h}{\text{argmax}}\ \mathbb{P}(D|h)$

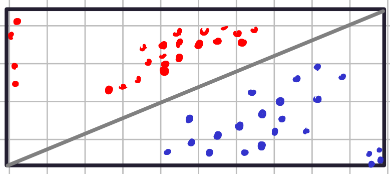
The class returned by h_{ML} is not necessarily the best one, it depends on the dataset, the best classifier is the BOC : $c = \underset{c}{\text{argmax}}\ \mathbb{P}(c | D, x) =$

$$\underset{c}{\text{argmax}} \sum_{h} \mathbb{P}(c | x, h) \mathbb{P}(h | D)$$ *total probabilities.

EXERCISE 3

Briefly describe a linear classification method and discuss its performance in presence of outliers. Use a graphical example to illustrate the concept.

the SVM aims to find a linear discriminant that maximizes the margin between the separator and the closest point to it, and the solution depends on the support vectors on that margins, so an outliers dont affect the solution, hence this method is robust to outliers.



EXERCISE 4

Given input values $\mathbf{x}_i$ and the corresponding target values t_i with $i = 1, \ldots, N$, the solution of regularized linear regression can be written as:

$$y(\mathbf{x}) = \sum_{i}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x},$$

with $\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\mathbf{t}$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and λ the regularization weight.

Considering a kernel function $k(\mathbf{x}, \mathbf{x}')$:

1. Provide a definition of the Gram matrix.
2. Explain how a kernelized version for regression can be obtained based on the equations provided above.

The Gram matrix is the square matrix $K = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_N) \\ \vdots & & & \\ k(x_N,x_1) & k(x_N,x_2) & \cdots & k(x_N,x_N) \end{bmatrix}$

We can substitute the dot product with the kernel measure: $y(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$

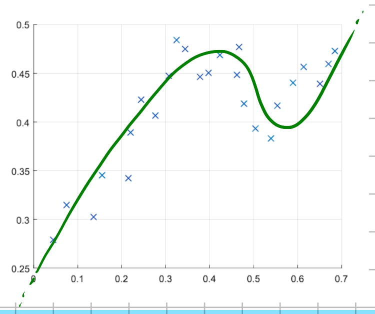
EXERCISE 5

Consider the learning problem of estimating the function $f : \Re \mapsto \Re$ with dataset $D = \{(x_i, y_i)\}$ plotted in the figure below:

1. Describe how to perform regression based on these data using a method of your choice. Specifically, provide a mathematical formulation of the model, highlighting the model parameters.
2. Considering the method you have chosen describe a way to reduce overfitting.
3. Draw a plausible plot of the learned model based on your choices.

We consider a parametrized model $y(x) = (w_0 \ w_1 \ w_2 \ w_3) \cdot \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \end{pmatrix}$, then we define the error function: $E(w) = \sum_{i=1}^{N} (t_i - y(x_i))^2$ and we minimize E by applying the gradient descent, until term. condition: $w \leftarrow w - \eta \nabla E$ where $\eta \in \mathbb{R}^+$.

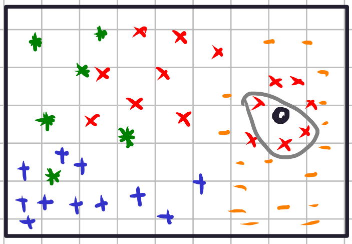
To reduce overfitting, we modify the error function: $E(w) = \sum_{i=1}^{N} (t_i - y(x_i))^2 + \lambda \|w\|^2$ where $\lambda \in \mathbb{R}^+$.



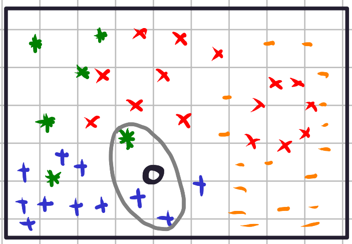
EXERCISE 6

1. Provide the main steps of classification based on K-nearest neighbors (K-NN).
2. Draw an example for a 4-classes classification problem in 2D. Use symbols (*,x,+,-) for the four classes. Graphically show the application of the K-NN algorithm with $K = 3$ for the classification of 3 different query points.

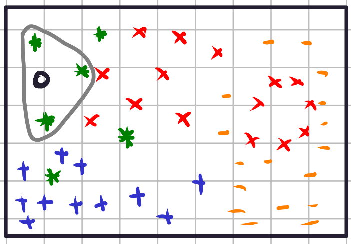
The KNN classify a point with the most popular class between his K closest point in the dataset.



class: x



class: +



class: *

The is the Following algorithms :

```
($D = \{x_i\}_{i=1}^{N}$ , K){
let $m_1 \cdots, m_K$ to be K random points
While (!termination condition){
    For ($i = 1 \ldots, K$){
        Let $T = \{$points in D closer to $m_i\}$
        $m_i = \frac{1}{|T|} \sum_{x \in T} x$
    }
}
```

In this method we assume that the dataset is generated by K Normal distribution :

$\mathbb{P}(x \mid c_i) = \mathcal{N}(x, \mu_i, \Sigma_i)$ the method aims to estimate the means μ_i.