**EXERCISE 1**

1. Provide a formal (domain-independent and solution-independent) definition of overfitting.
2. Discuss the problem of overfitting in learning with Decision Trees and illustrate possible solutions to it.

Let $h$ to be an hypothesis, we denote $error(h)$ the TRUE error of $h$.

$h: X \to Y$, $error(h) = \frac{1}{|X|}\sum_{x \in X} \delta(h(x) \neq f(x))$ where $f$: target funct. and $X$ is finite.

If $D$ is the dataset, the sample error is $error_D(h) = \frac{1}{|D|}\sum_{x \in D} \delta(h(x) \neq f(x))$ ↑ LABELS

$h$ is *overfitting* if $\exists\, h'$ such that:

$$error_D(h) < error_D(h') \quad \text{and} \quad error(h) > error(h')$$

The decision trees overfits when we have a leaf for each sample in the dataset, we can overcome this with *pruning*:

```
PRUNING(h:tree){
    until termination condition do {
        n = random node from h
        h' = we cut from h the subtree with root n
        a = accuracy(h)
        a' = accuracy(h')
        if a' > a {
            h = h'
        }
    }
    return h
}
```

A possible term. cond. might be: the accuracy didn't increased in the last $\kappa$ iterations.

**EXERCISE 2**

1. Describe the *Naive Bayes Classifier* and highlight the approximation made with respect to the Bayes Optimal Classifier.
2. Provide design and implementation choices for solving the following problem through *Naive Bayes Classifier*:
   *Classification of scientific papers in categories according to their main subject. The categories to be considered are: ML (Machine Learning), KR (Knowledge Representation), PL (Planning). Data available for each scientific paper are: title, authors, abstract and publication site (name of the journal and/or of the conference).*

Given a dataset $D = \{(x_i, t_i)\}_{i=1}^{N}$ the optimal Bayes classifier is given by considering $\mathbb{P}(c \mid x, D) = \sum_{h \in H} \mathbb{P}(c \mid x, D, h)\mathbb{P}(h \mid x, D) = \sum_{h \in H} \mathbb{P}(c \mid x, h)\mathbb{P}(h \mid D)$

where $H$ is the set of all the hypothesis. The naïve Bayes classifiers assume that the attributes of the samples are independent:

if $x_i = (x_{i1} \cdots x_{id})^T$ $\mathbb{P}(x_{ij} = a \mid x_{ik} = a') = \mathbb{P}(x_{ij} = a)$ $\forall\, j \neq k$.

$$\Rightarrow \mathbb{P}(c \mid x, D) = \mathbb{P}(c \mid x_{i1} \cdots x_{id}, h) = \mathbb{P}(x_{i1} \cdots x_{id} \mid c, D)\mathbb{P}(c \mid D) \cdot \frac{1}{\alpha}$$

$$\Rightarrow \operatorname{argmax} \mathbb{P}(c \mid x, D) = \operatorname{argmax} \mathbb{P}(x_{i1} \cdots x_{id} \mid c, D)\mathbb{P}(c \mid D) =$$

$$\operatorname{argmax} \mathbb{P}(c \mid D) \cdot \prod_{j=1}^{d} \mathbb{P}(x_{ij} \mid c, D)$$

$\mathbb{P}(c \mid D)$: $\frac{\text{number of } c \text{ sample}}{\text{number of sample}}$

$\mathbb{P}(x_{ij} \mid c, D)$: $\frac{\text{number of } c \text{ sample with } x_i(j) = x_{ij}}{\text{number of } c \text{ sample}}$

A row of the dataset is

TITLE AUTHORS ABSTRACT SITE |

$(TI_i, AU_i, AB_i, SI_i) \in D$ and $c_i \in \{ML, PL, KR\}$

We have $\mathbb{P}(ML|D) = \frac{|\{(x_i, c_i) \in D : c_i = ML\}|}{|D|}$ $\mathbb{P}(KR|D) = \frac{|\{(x_i, c_i) \in D : c_i = KR\}|}{|D|}$

$\mathbb{P}(PL|D) = \frac{|\{(x_i, c_i) \in D : c_i = PL\}|}{|D|}$

Then, For a given title $T^* \in \{\text{All possible titles}\}$ we have

$$\mathbb{P}(T^* | C, D) = \frac{|\{(x_i, c_i) \in D : c_i = C \wedge TI_i = T^*\}|}{|\{(x_i, c_i) \in D : c_i = C\}|}$$

The same Follows For the other attributes. A new $x'$ is classified:

$$\ell(x') = \arg\max_C \mathbb{P}(C|D) \cdot \prod_{j=1}^{5} \mathbb{P}(x_{ij} | C, D)$$

---

**EXERCISE 3**

Consider a dataset $\mathcal{D} = \{((a_1, s_1), p_1), \ldots, ((a_N, s_N), p_N)\}$ containing the number of hours $a_i$ a student has attended a course, the number of hours $s_i$ s/he has studied for the course and whether or not s/he has passed the exam $p_i = \{0, 1\}$.

1. Define a model based on **logistic regression** that, given the values of $a$ and $s$, estimates whether a student passes the exam or not.
2. Discuss which are the parameters of the model that have to be learned based on the given data.
3. What is a suitable error function for learning the parameters of the model?

p = all the labels in D

Given D, we estimate $\mathbb{P}(P|x, D) = \prod_{i=1}^{N} \pi_i^{p_i} (1-\pi_i)^{(1-p_i)}$ con $\pi_i = \mathbb{P}(p_i = 1 | x_i) = \sigma(w^T x_i)$

$\Rightarrow \sigma(w^T x_i) = (1 - \exp(w^T x_i))^{-1}$. We want to maximize $\mathbb{P}(P|x, D)$ that depends on w. $\ln \mathbb{P}(P|x, D) = \sum_{i=1}^{N} p_i \ln(\pi_i) + (1-p_i) \ln(1-\pi_i)$ recall that $\pi_i$ depends on w.

so: $E(w) = -\sum_{i=1}^{N} p_i \ln(\pi_i) + (1-p_i) \ln(1-\pi_i)$ and we can Find the optimal $w^*$ by using a method based on $\nabla E$ and $H(w) = \nabla\nabla E$ (Hessian).

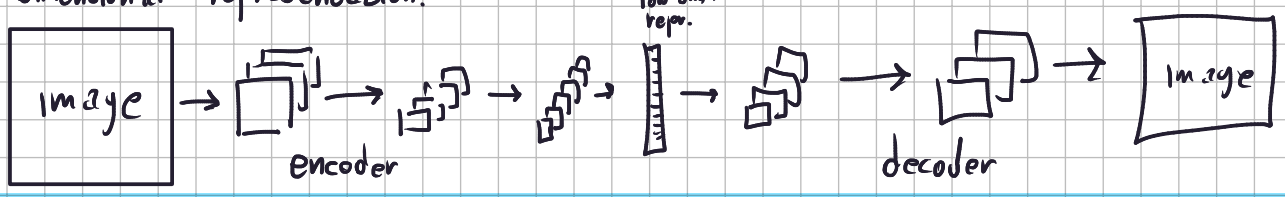
---

**EXERCISE 4**

1. Briefly describe what is the architecture of an autoencoder and its purpose.
2. Draw an example of autoencoder.

An autoencoder is a NN that is composed by two NN:

1) the First one transForm the input (like an image) in a low dimension representation, usually a 1-dim. tensor.
2) the second, From the given low-dimensional encoding, reconstruct the high-dimensional representation.



**EXERCISE 5**

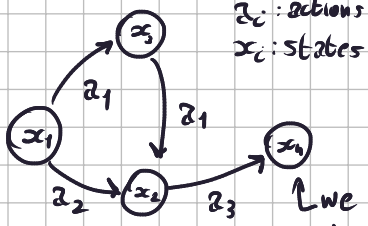
1. Describe the concept of full observability in models representing dynamic systems.
2. Describe the difference between a Markov Decision Process (MDP) and a Partially Observable Markov Decision Process (POMDP), referring to their formal models.
3. Draw and explain the graphical models of MDP and POMDP.

A dynamic system is *Fully observable* if we can read his entire state vector $x_i$.

A POMPD is not Fully observable, From $x_t$ we can get $z_t$ such that: $z_t$ is a restriction of $x_t$

OR there is a distribution $D(z, x_t)$ over $z$ that gives the probability of observing a given $z$: $\mathbb{P}(z_t = z^* \mid x_t) = D(z, x_t)$

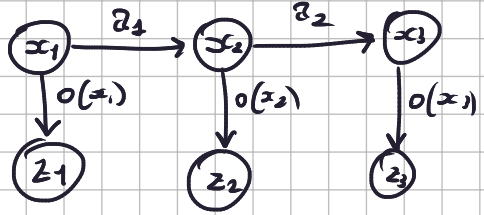
MDP




$a_i$ : actions
$x_i$ : states

we can observe the state $x_i$

POMPD




we can observe only $z_i$ from $x_i$.

**EXERCISE 6**

Consider a two-layers ANN which receives in input vectors $\mathbf{x}$ of dimension 128 and produces output vectors $\mathbf{y}$ of dimension 10. The hidden layer of the ANN is composed of 50 units which use the ReLU activation function. The output units use a linear activation function. The weight matrices of the hidden and output layers are denoted $W_1$ and $W_2$, respectively.

1. Provide the dimensions of the weight matrices $W_1$ and $W_2$
2. Provide the formula explicitly stating how the values of $\mathbf{y}$ are computed given an input vector $\mathbf{x}$ in terms of the weight matrices and the activation functions (you can ignore the bias terms).

Col : in
Row : out

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & & \\ 3 & & \\ 4 & & \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

$W_1$ is a map From $\mathbb{R}^{128}$ to $\mathbb{R}^{50}$ so $W_1$ have 128 columns and 50 rows.

$W_2$ is a map From $\mathbb{R}^{50}$ to $\mathbb{R}^{10}$ so $W_2$ have 50 columns and 10 rows.

$h_1(x) = \text{ReLU}(W_1 x)$, $h_2 = W_2 h_1 \Rightarrow y(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x)$