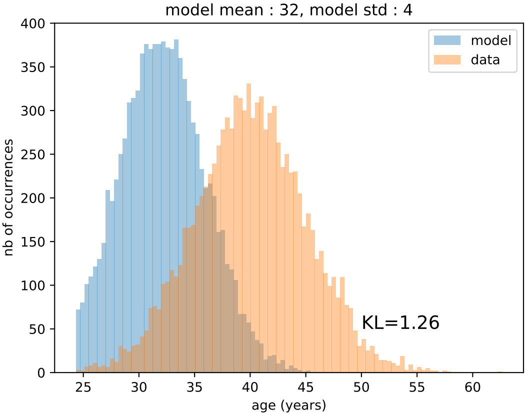
Machine learning II, unsupervised learning and agents: density estimation



Maximum likelihood

KL divergence

Kernel density estimation

Density estimation

Objective : compute a probability distribution that represents the data well.

Maximum Likelihood

The **Maximum Likelihood** method is one example method.
We observe a dataset $D_n = (x_1, ..., x_n)$.
We first need to choose a **model** (which is the distribution) of the dataset, p.
Then, we must optimize the **parameters of this model**, noted θ.

Maximum Likelihood

The **likelihood** (vraisemblance) of the model is

$$L(\theta) = p(x_1, \ldots, x_n|\theta) \tag{1}$$

Here, p denotes the probability of observing the sample $(x_1, \ldots, x_n)$, when the model had the parameter θ, (or the probability density, in the corresponding context).

Maximum Likelihood

The **likelihood** (vraisemblance) of the model is

$$L(\theta) = p(x_1, \dots, x_n|\theta) \tag{2}$$

This is the function that we want to **maximize**, by choosing the best possible θ (optimization problem).
If $(x_1, \dots, x_n)$ are conditionally independant, then it writes :

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta) \tag{3}$$

Remark on max-likelihood

Most of the time it's written this way : "minimise $-logL(\theta)$"
Because the log **transforms the product into a sum**, which is easier to **differentiate**.

$$-logL(\theta) = -\sum_{i=1}^{n} \log(p(x_i|\theta)) \tag{4}$$

Example 1

Exercice 1 : We observe the data $(1, 0)$. We assume that these data come from a random variable that follows a Bernoulli distribution of parameter p. What is the likelihood of these observations as a function of p ?

Example 1

Exercice 1 : We observe the data $(1, 0)$. We assume that these data come from a random variable that follows a Bernoulli distribution of parameter p. What is the likelihood of these observations as a function of p ?

$$L = P(1|p)P(0|p) \tag{5}$$

For which value of p is this likelihood **maximum** ?

Example 2

Exercice 2 : We observe the data $(2.5, 3.5)$. We assume that these data come from a normal law of parameters μ and σ.
What is the likelihood of (μ, σ) ?

Example 2

Exercice 2 : We observe the data $(2.5, 3.5)$. We assume that these data come from a normal law of parameters μ and σ.

$$\begin{aligned} L &= p(2.5|\mu,\sigma)p(3.5|\mu,\sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{2.5-\mu}{\sigma})^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{3.5-\mu}{\sigma})^2} \end{aligned} \tag{6}$$

Example 2

Exercice 2 : We observe the data $(2.5, 3.5)$. We assume that these data come from a normal law of parameters μ and σ.

$$\begin{aligned} L &= p(2.5|\mu,\sigma)p(3.5|\mu,\sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{2.5-\mu}{\sigma})^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{3.5-\mu}{\sigma})^2} \end{aligned} \quad (7)$$

We wan show that the likelihood is maximum for :

- $\hat{\mu} = \frac{2.5+3.5}{2}$
- $\hat{\sigma^2} = \frac{(2.5-\hat{\mu})^2+(3.5-\hat{\mu})^2}{2}$

Kullbach-Leibler Divergence

The KL divergence is a measure of the discrepancy between two **distributions**.

Expected value (espérance)

- For a discrete random variable X that takes the values x_i with probability p_i :

$$E(X) = \sum_{i=1}^{n} p_i x_i \tag{8}$$

- For a continuous random variable X with density p(x) :

$$E(X) = \int x p(x) dx \tag{9}$$

Kullbach-Leibler Divergence

- The samples $(x_1, .., x_n)$ are described by an empirical distribution.
- The **Kullbach-Leibler divergence** is a tool to compare distributions.
- It is not a distance : it is not symmetric, no triangular inequality.

Kullbach-Leibler Divergence

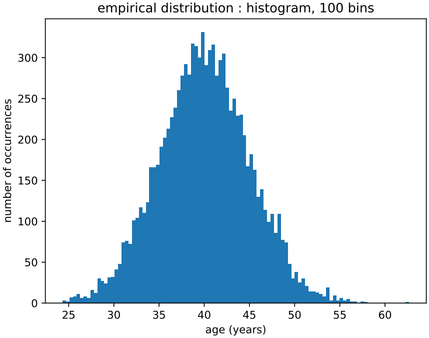
- $$\mathcal{D}[p||q] = \mathbb{E}_{\sim p}[\log(\frac{p}{q})] \tag{10}$$
- For discrete variables
$$\mathcal{D}[p||q] = \sum_i p(i) \log \frac{p(i)}{q(i)} \tag{11}$$
- for continuous variables
$$\mathcal{D}[p||q] = \int_X p(x) \log \frac{p(x)}{q(x)} dx \tag{12}$$

Exercice 2 : **Fitting a distribution**

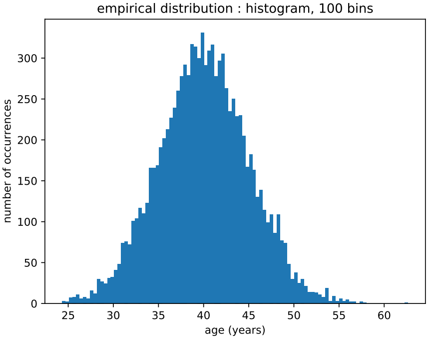
- **cd kl_divergence**
- A two dimensional dataset is contained in **empirical_distribution.csv**. It represents the **age distribution** of some groupe of people. We want to study this age distribution.
- load it in **compute_kl.py**. We will use the functions provided in the file in order to find the best model, meaning here the model M, such that $KL(M||\tilde{P})$ is smallest, with $\tilde{P}$ the empirical distribution of the data.

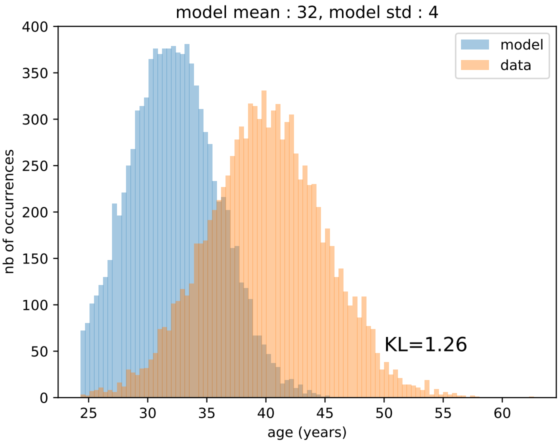
Exercice 2 : **Fitting a distribution**

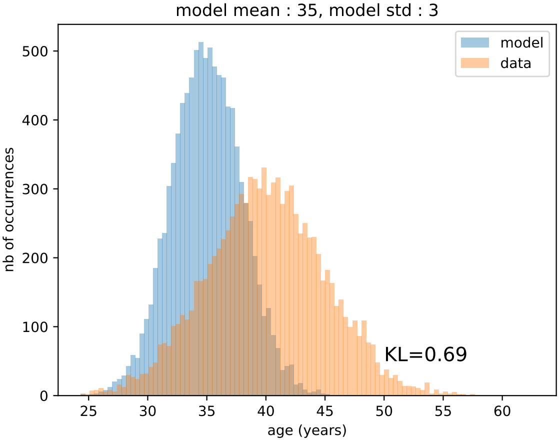
- **First step** : choice of the model
- Plot the histogram of the data : what model seems to be relevant ?



- We will use **normal laws**. We want to fint the normal law that is **the closest to the empirical data**
- We measure the proximity between the model and the empirical data with the KL divergence.



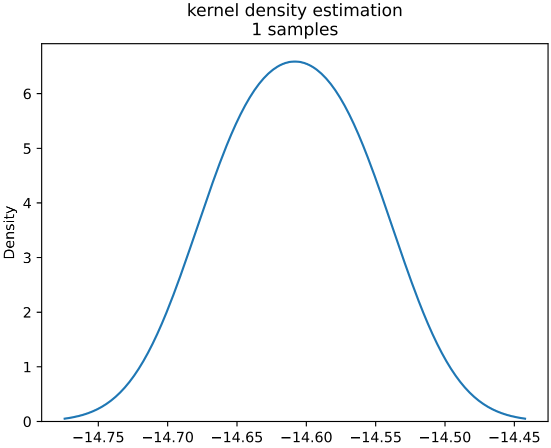

model mean : 32, model std : 4
model
data
KL=1.26
nb of occurrences
age (years)
0
50
100
150
200
250
300
350
400
25
30
35
40
45
50
55
60

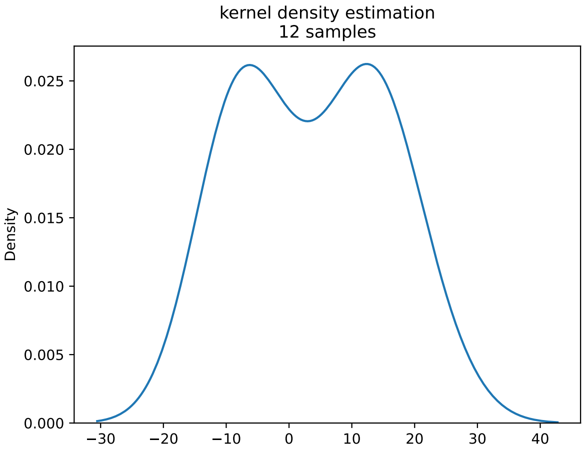

model mean : 35, model std : 3
model
data
KL=0.69
nb of occurrences
age (years)
0
100
200
300
400
500
25
30
35
40
45
50
55
60

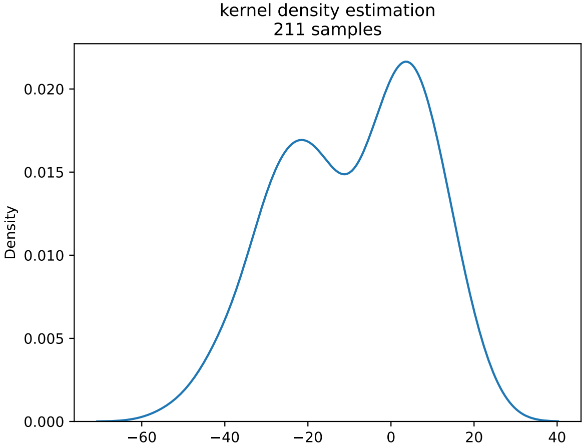
Kernel density estimation (non-parametric model)

`https://seaborn.pydata.org/generated/seaborn.kdeplot.html`

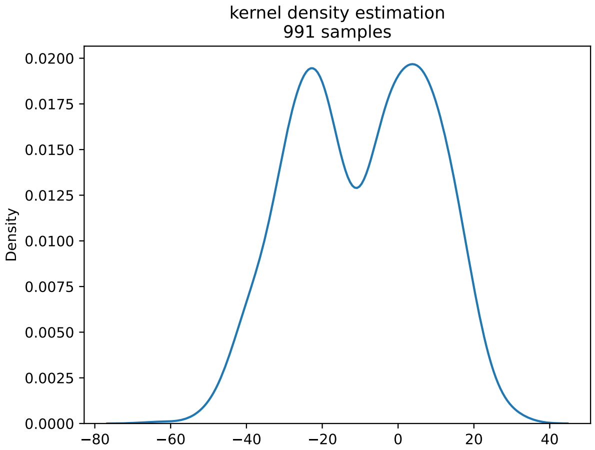
Example in **kde/**


kernel density estimation
1 samples
Density
0
1
2
3
4
5
6
−14.75
−14.70
−14.65
−14.60
−14.55
−14.50
−14.45


kernel density estimation
12 samples
Density
0.000
0.005
0.010
0.015
0.020
0.025
−30
−20
−10
0
10
20
30
40



kernel density estimation
211 samples
Density
0.020
0.015
0.010
0.005
0.000
−60
−40
−20
0
20
40


kernel density estimation
991 samples
Density
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−80
−60
−40
−20
0
20
40