« BAYESIAN LEARNING »

1. BAYESIAN DECISION

Stéphane Derrode, Dpt MI

stephane.derrode@ec-lyon.fr

Contents for *1. Bayesian decision*

1. Notations and reminders
2. Bayes' decision theory with examples
3. Proof of Bayes' decision theory
4. Gaussian case

BAYESIAN DECISION

1. Notations and reminders

Notations for continuous real-valued RV

- Series of observations of length ***N*** are modeled by a stochastic process with as many random variables as there are samples:

$$\boldsymbol{Y} = \boldsymbol{Y}_1^N = \{Y_1, Y_2, \ldots, Y_n, \ldots, Y_N\}$$

- Each random variable Y_n is assumed to be real-valued and characterized by a pdf (mostly Gaussian).

- Notations:

$$p(Y = y) = p(Y \in dy) = p(y)$$

$$p(\boldsymbol{Y} = \boldsymbol{y}) = p(\boldsymbol{Y} \in d\boldsymbol{y}) = p(\boldsymbol{y}) = p(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_N = \boldsymbol{y}_N)$$

Notations for discrete RV

- Series of classes/labels are modeled by a stochastic process with as many random variables as there are samples:

$$\boldsymbol{X} = \boldsymbol{X}_1^N = \{X_1, X_2, \ldots, X_n, \ldots, X_N\}$$

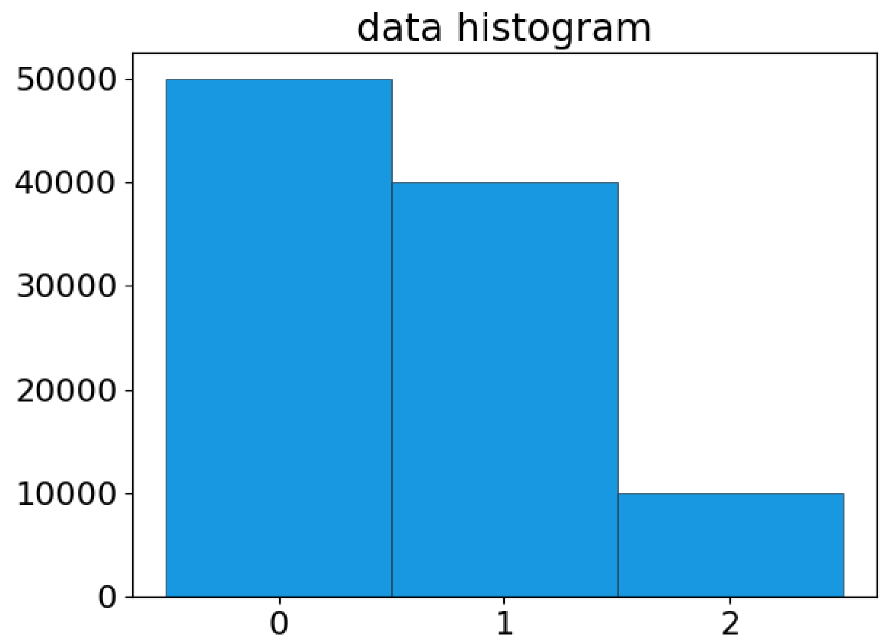
- Each random variable X_n is assumed to be discrete-valued

$$X_n \in \Omega = \{1, \ldots, K\}$$

- Notations:

$$p(X = x) = p(x)$$

$$p(\boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x}) = p(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_N = \boldsymbol{x}_N)$$




```
if __name__ == "__main__":

    nbsample = 100000
    proba = [0.5, 0.4, 0.1]
    sample = np.random.choice(a=[0, 1, 2], size=nbsample, p=proba)

    ax = plt.figure().gca()
    ax.xaxis.set_major_locator(MaxNLocator(integer=True))
    plt.hist(sample, bins=np.arange(len(proba)+1)-0.5, facecolor='#1798E1',\
        edgecolor="#223E4F", linewidth=0.5)
    plt.title('data histogram')
    plt.tight_layout()
    plt.savefig("./fig/SimulSampleDiscrete.png")
    plt.close()
```

A few reminders

- The expectation of a discrete RV is given by

$$E[X] = \sum_{k \in \Omega} k \; p(X = k)$$

$$E[g(X)] = \sum_{k \in \Omega} g(k) \; p(X = k)$$

- The expectation of a continuous RV by

$$E[Y] = \int_{-\infty}^{\infty} y \; p(Y = y) \; dy$$

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) \; p(Y = y) \; dy$$

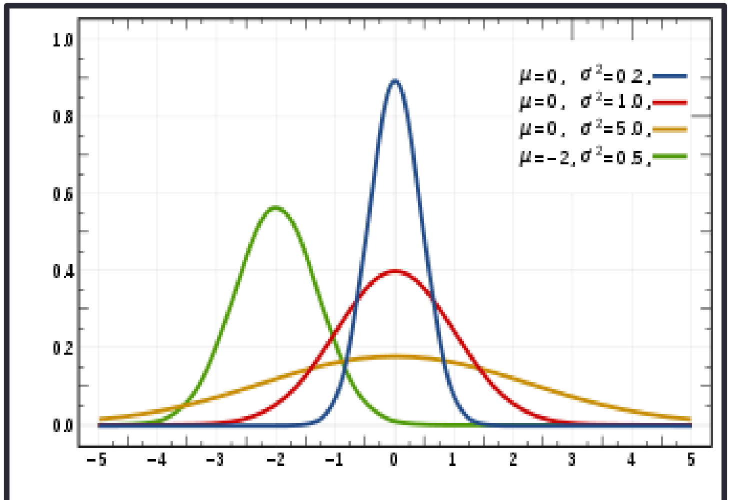
The probability density of the normal distribution is

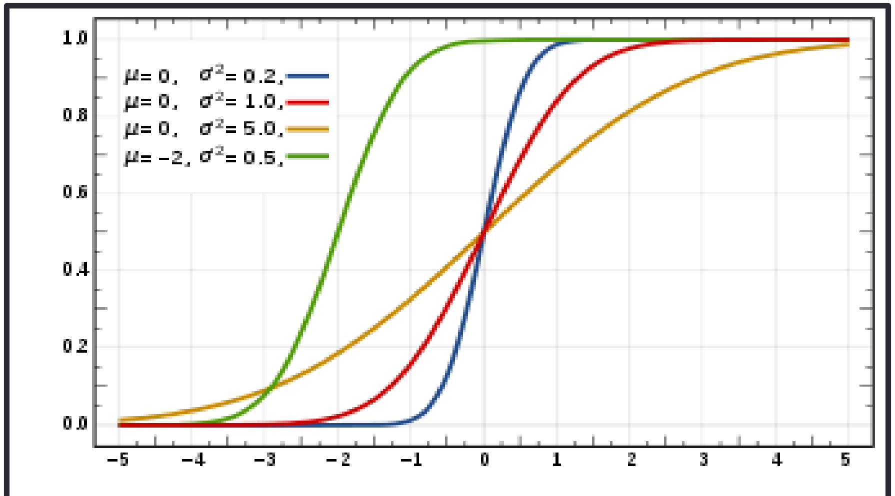
$$\boxed{f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}}}$$

where

- μ is the mean or expectation of the distribution (and also its median and mode),
- σ is the standard deviation, and
- σ^2 is the variance.

We will write $Y \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$





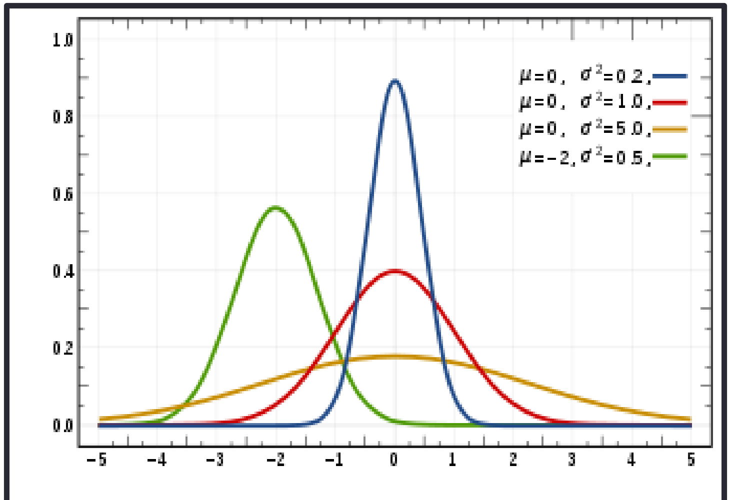
- The mean is the expected value of Y

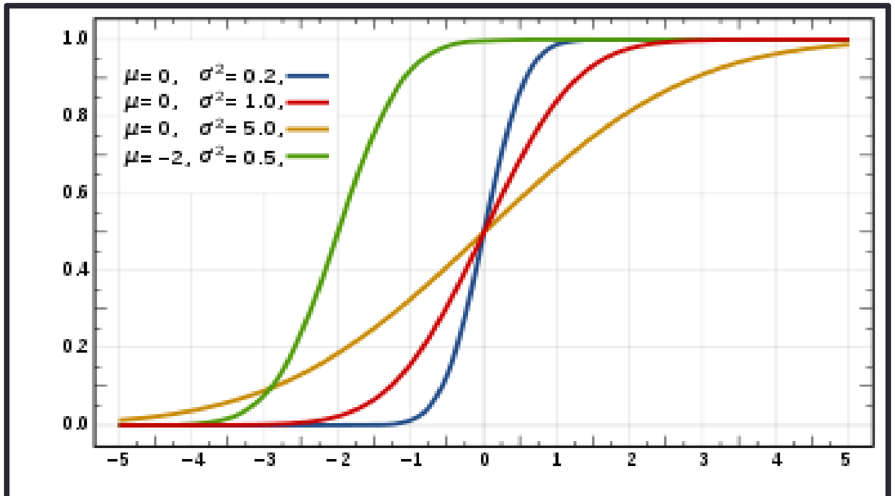
$$\mu = E[Y]$$

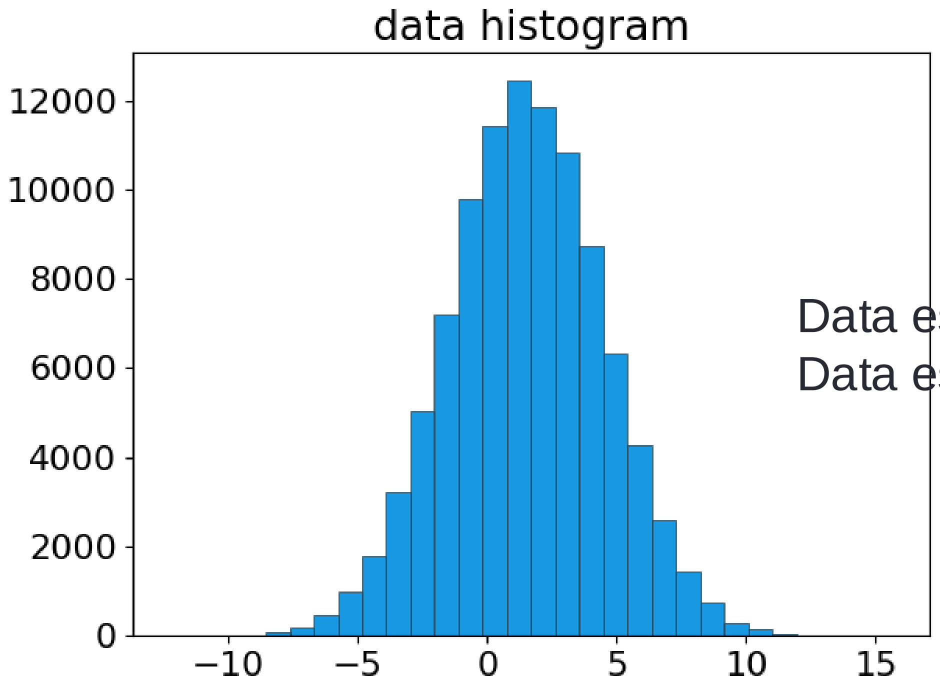
- The variance is the expected squared deviation

$$\sigma^2 = E\left[(Y-\mu)^2\right]$$

We will write $Y \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$







Data estimated mean : 1.51
Data estimated var : 8.98

```
if __name__ == "__main__":
    nbsample = 100000
    mean, std = 1.5, 3.0
    sample = np.random.normal(loc=mean, scale=std, size=nbsample)
    print('estimated mean :', np.mean(sample), '\nestimated var :', np.var(sample))

    ax = plt.figure().gca()
    plt.hist(sample, bins=30, facecolor='#1798E1', edgecolor="#223E4F", linewidth=0.5)
    plt.title('data histogram')
    plt.tight_layout()
    plt.savefig("./fig/SimulSampleGaussian.png")
    plt.close()
```

BAYESIAN DECISION

2. Bayes' decision theory with examples

Bayesian Decision Theory

Bayes' Theorem with Two Random Variables

$$p(X = x|Y = y) = \frac{p(Y = y|X = x) \cdot p(X = x)}{p(Y = y)}$$

Where:

- $p(X = x|Y = y)$ is the **posterior probability** of $X = x$ given that $Y = y$.
- $p(Y = y|X = x)$ is the **likelihood**, i.e., the probability of observing $Y = y$ given $X = x$.
- $p(X = x)$ is the **prior probability** of $X = x$.
- $p(Y = y)$ is the **marginal probability** of observing $Y = y$, which is computed by summing over all possible values of X:

$$p(Y = y) = \sum_x p(Y = y|X = x) \cdot p(X = x)$$

Explanation

- **Prior Probability** $p(X = x)$: The initial belief about the probability of $X = x$, before any evidence $Y = y$ is observed.
- **Likelihood** $p(Y = y|X = x)$: The probability of observing $Y = y$, given that $X = x$.
- **Posterior Probability** $p(X = x|Y = y)$: The updated probability of $X = x$ after observing $Y = y$.

Bayesian Decision Theory

Example (with Y discrete):

Let X represent the weather condition (Rainy or Sunny), and let Y represent whether a person carries an umbrella (Umbrella or No Umbrella).

1. **Prior**:
$$p(X = \text{Rainy}) = 0.3, \quad p(X = \text{Sunny}) = 0.7$$

2. **Likelihood**:
$$p(Y = \text{Umbrella}|X = \text{Rainy}) = 0.9, \quad p(Y = \text{Umbrella}|X = \text{Sunny}) = 0.1$$

3. **Marginal Probability** of $Y = \text{Umbrella}$:
$$p(Y = \text{Umbrella}) = p(Y = \text{Umbrella}|X = \text{Rainy}) \cdot p(X = \text{Rainy}) + p(Y = \text{Umbrella}|X = \text{Sunny}) \cdot p(X = \text{Sunny})$$
Substituting the values:
$$p(Y = \text{Umbrella}) = (0.9 \cdot 0.3) + (0.1 \cdot 0.7) = 0.27 + 0.07 = 0.34$$

4. **Posterior**: To find the updated probability of rain given that the person is carrying an umbrella:
$$p(X = \text{Rainy}|Y = \text{Umbrella}) = \frac{p(Y = \text{Umbrella}|X = \text{Rainy}) \cdot p(X = \text{Rainy})}{p(Y = \text{Umbrella})}$$
Substituting the values:
$$p(X = \text{Rainy}|Y = \text{Umbrella}) = \frac{(0.9 \cdot 0.3)}{0.34} = \frac{0.27}{0.34} \approx 0.79$$

Thus, after observing the umbrella, the updated probability that the weather is rainy is approximately 79%.

Conclusion This process demonstrates how **Bayes' Theorem** updates our belief about a hypothesis (such as the weather) based on new evidence (such as the presence of an umbrella), using conditional probabilities.

Bayesian Decision Theory

$$p(X = x|Y = y) = \frac{p(Y = y|X = x) \cdot p(X = x)}{p(Y = y)}$$

Bayes decision (2 classes, cas particulier)

We try to predict the sex (X=*1* or X=*2*) of a person from its height (Y)

$$p(X = 1|Y = y) \propto p(X = 1)p(Y = y|X = 1)$$

Posterior　　　Prior　　　Likelihood

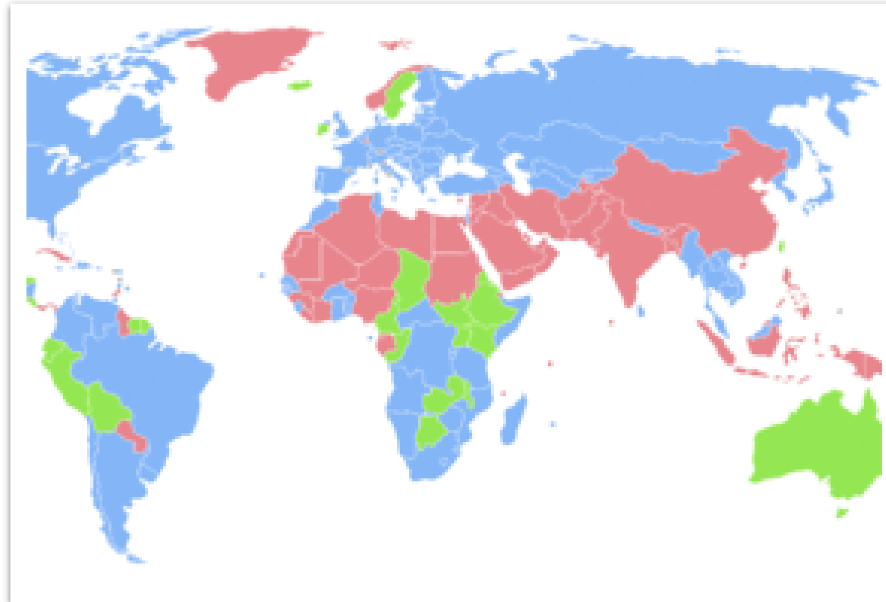
$$p(X = 2|Y = y) \propto p(X = 2)p(Y = y|X = 2)$$

under the following condition $P(X = 2|Y = y) + P(X = 1|Y = y) = 1$

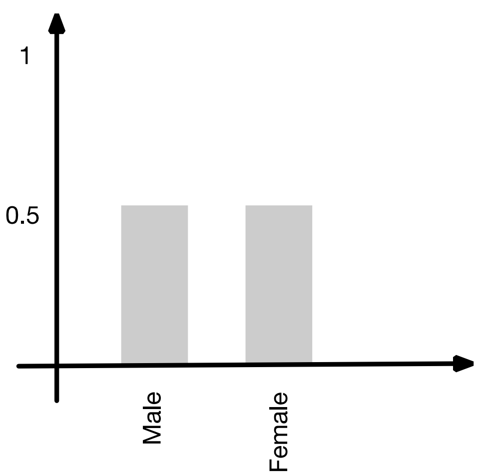
A priori probabilities

Map indicating the human sex ratio by country.[1]

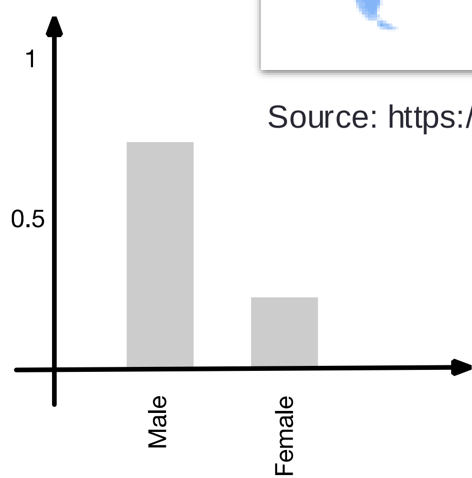
- Countries with more **females** than males.
- Countries with the **same** number of males and females (accounting that the ratio has 3 significant figures, i.e., 1.00 males to 1.00 females).
- Countries with more **males** than females.
- No data



Source: https://en.wikipedia.org/wiki/Human_sex_ratio



Global scale

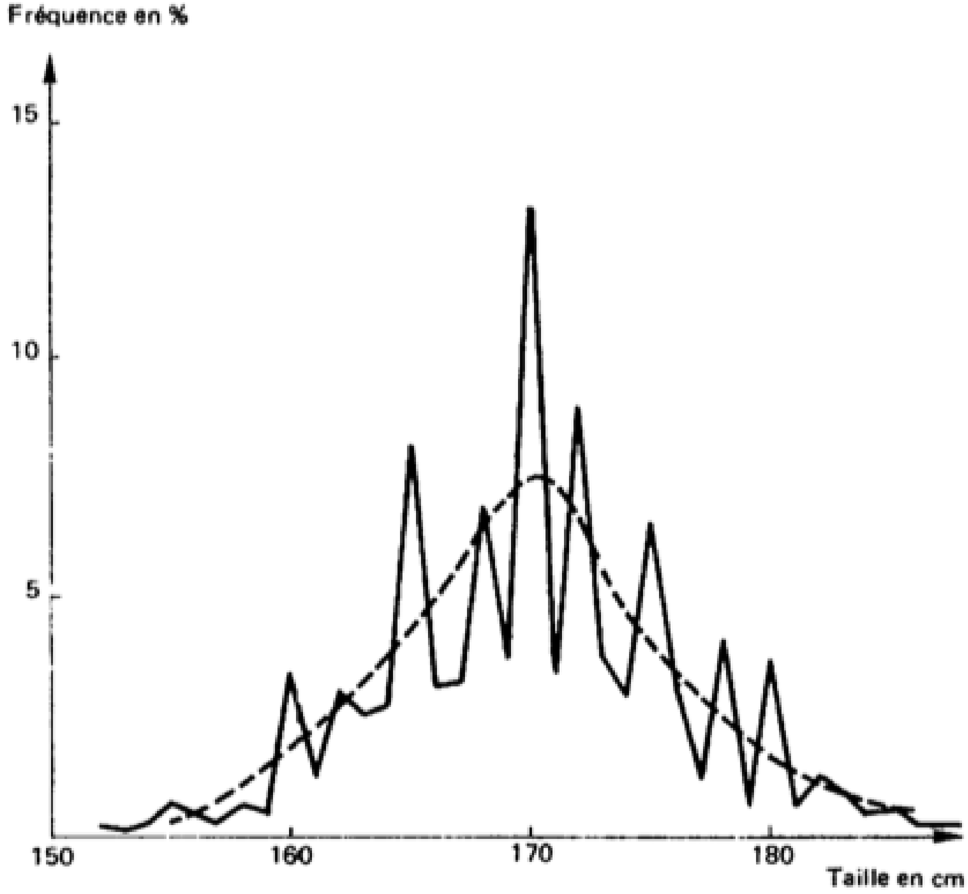


Classroom scale

Conditional probabilities

Normalized histogram of men' size in France in the seventies and estimated Gaussian density.

$$p(Y = y|X = 1)$$



Nicolas Bayes, 1702-1761,
English statistician

BAYESIAN DECISION

3. Proof of Bayes' decision theory

2. Bayesian Decision Theory

- Bayesian Decision Theory is a fundamental statistical approach to the problem of pattern classification.
- Quantifies the trade-off between various classifications using probability and the costs that accompany such classifications.

Fingerprint classification

- **Assumptions:**
  - Decision problem is posed in probabilistic terms.
  - All relevant probability values are known.
- The classification is to estimate a realization of the hidden *X* from the observable *Y*.

Bayesian strategy for classification

- **A priori law** (prior): $p(X = k) = p(k) = \pi_k, \text{on } \Omega = \{1, \ldots, K\}$
- **Conditional laws** (likelihood) :

$$p(Y = y | X = k) = f_k(y), \text{on } \mathbb{R}$$

- **Joint law**: $p(Y = y, X = k), \text{on } \mathbb{R} \times \Omega$
- **Mixture**: $$p(Y = y) = \sum_{k=1}^{K} p(Y = y, X = k) = \sum_{k=1}^{K} \pi_k \; f_k(y)$$
- **A posteriori laws** (posterior):

$$p(X = k | Y = y) = \frac{p(Y = y, X = k)}{p(y)} = \frac{\pi_k \; f_k(y)}{\sum_{l=1}^{K} \pi_l \; f_l(y)}$$

Assume y to be an observation and x its (true) class or label.

- **Classification strategy**

$$\begin{aligned} \hat{s} : \mathbb{R} &\longrightarrow \Omega \\ y &\longrightarrow \hat{x} \end{aligned}$$

$$\hat{s}(y) = \hat{x} \begin{cases} = x & \text{true} \\ \neq x & \text{wrong} \end{cases}$$

- **Loss function**

$$L : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

$$L(i,j) = \begin{cases} 0 & \text{if } i = j \\ \lambda_{i,j} > 0 & \text{else} \end{cases}$$

$$L(i,j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{sinon} \end{cases}$$

L is called the "0-1 loss" function

Assume $\hat{s}$ and L given, how can we measure the quality of $\hat{s}$?

Suppose that we have ***N*** independent observations and we know the true labels of the sample.

$$\boldsymbol{y} = \{y_1, \ldots, y_N\}$$
$$\boldsymbol{x} = \{x_1, \ldots, x_N\}$$

The total loss for the sample is

$$L(\hat{s}(y_1), x_1) + \ldots + L(\hat{s}(y_N), x_N)$$

We try to minimize this loss.

According to the law of large numbers

$$\frac{L(\hat{s}(y_1), x_1) + \ldots + L(\hat{s}(y_N), x_N)}{N} \xrightarrow{N \to \infty} E[L(\hat{s}(Y), X)]$$

The quality of the strategy $\hat{s}$ is measured by (when *N* is large)

$$E[L(\hat{s}(Y),X)]$$

which is called the « mean loss ».

The Bayesian strategy, denoted by $\hat{s}_B$, is the one that minimizes the mean loss

$$E[L(\hat{s}_B(Y),X)] = \min_{\hat{s}} E[L(\hat{s}(Y),X)]$$

Be carefull : this is true for a large number of samples, and we can’t say something for only a few samples.

Exercise : show that the Bayesian strategy $\hat{s}_B$ with the loss function

$$L(i,j) = \begin{cases} 0 & \text{if } i = j \\ \lambda_{i,j} > 0 & \text{else} \end{cases}$$

can be written

$$\hat{s}_B(y) = k = \arg\min_{j \in \Omega} \sum_{i=1}^{K} \lambda_{j,i} \, p(X = i | Y = y)$$

The minimal mean loss is then given by

$$\xi = E[L(\hat{s}_B(Y), X)] = \int_{\mathbb{R}} \phi(y) p(Y = y) \, dy = \int_{\mathbb{R}} \sum_{i=1}^{K} \pi(i) f_i(y) L(\hat{s}_B(y), i) \, dy$$

Specific case: $\Omega = \{1, 2\}$ $L(i,j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{else} \end{cases}$

Express the Bayesian strategy $\hat{s}_B$ and the minimal mean loss ξ of the classifier.

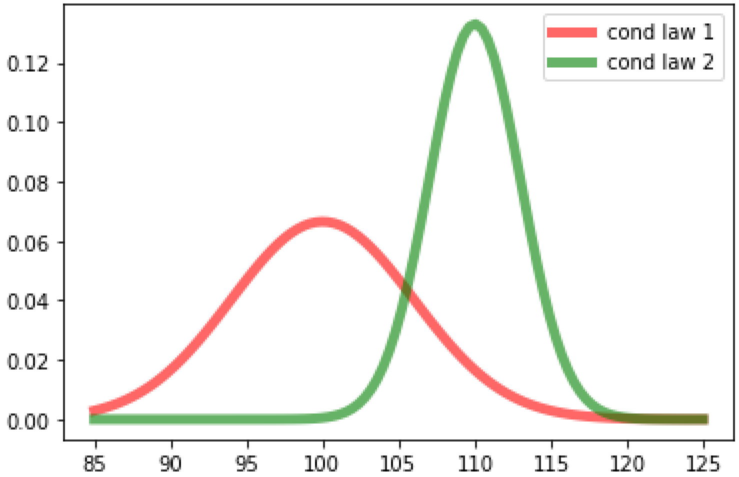
BAYESIAN DECISION

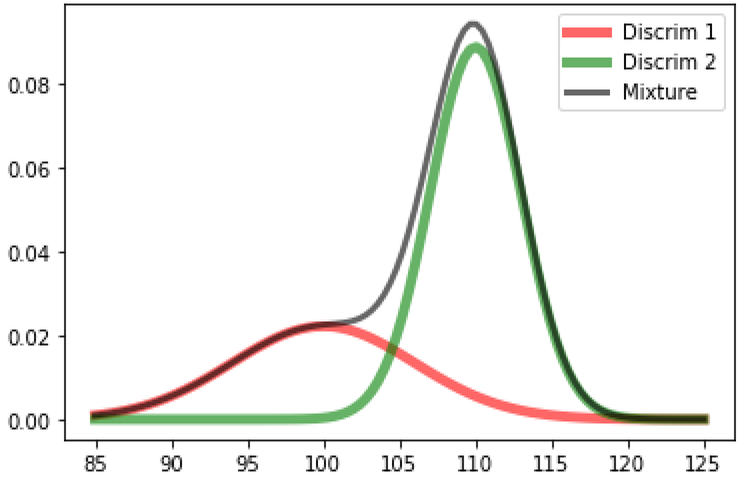
4. Gaussian case

Example

$$\mathcal{N}\left(\mu_1 = 100, \sigma_1 = 6\right)$$
$$\mathcal{N}\left(\mu_2 = 110, \sigma_2 = 3\right)$$

$$\pi_1 = \frac{1}{3}, \pi_2 = \frac{2}{3}$$





Example continued

1. Calculate the Bayesian decision thresholds, *i.e.* when the decision switches from class 1 to 2, and from class 2 to 1. For calculations, you can set

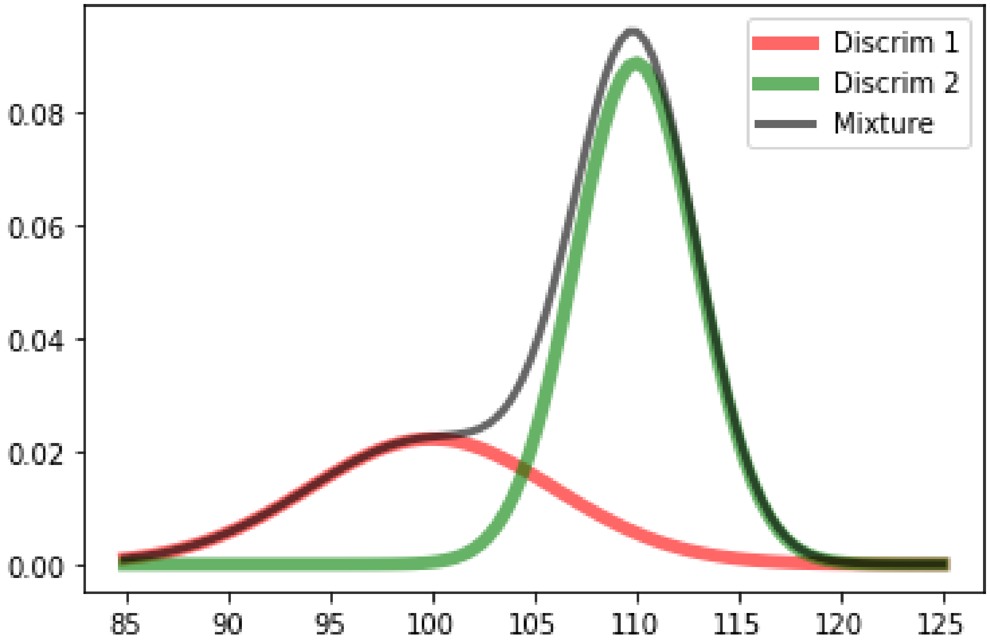
$$\mu_1 = a, \mu_2 = a + 10, \quad \sigma_1 = s, \sigma_2 = s/2 \qquad \pi_1 = \frac{1}{3}, \pi_2 = \frac{2}{3}$$

2. Assuming a $L_{0\text{-}1}$ loss function, calculate the mean loss.

TIP : Error function (special function)

$$erf(x) = \int_0^x e^{-z^2} dz, \quad \text{with } lim_{x\to\infty} erf(x) = 1$$

1. $\tau_1 = 104.5, \tau_2 = 122.1$



$$\pi_1 = \frac{1}{3}, \pi_2 = \frac{2}{3}$$

$$\mathcal{N}\left(\mu_1 = 100, \sigma_1 = 6\right)$$
$$\mathcal{N}\left(\mu_2 = 110, \sigma_2 = 3\right)$$

2. $\xi = 0.098$

$$\pi_1 = \frac{1}{3}, \pi_2 = \frac{2}{3}$$

$$\mathcal{N}(\mu_1 = 100, \sigma_1 = 6)$$

$$\mathcal{N}(\mu_2 = 110, \sigma_2 = 3)$$

$$\xi = \underbrace{\int_{-\infty}^{\tau_1} \pi(2)\ f_2(y)\ dy}_{A} + \underbrace{\int_{\tau_1}^{\tau_2} \pi(1)\ f_1(y)\ dy}_{B} + \underbrace{\int_{\tau_2}^{+\infty} \pi(2)\ f_2(y)\ dy}_{C}.$$

$$A = \frac{1}{3}\left(1 + \operatorname{erf}\left(\frac{\sqrt{2}}{\sigma}\left(\tau_1 - a\right)\right)\right) = 0.023.$$

$$B = \frac{1}{6}\left(\operatorname{erf}\left(\frac{\tau_2}{\sigma\ \sqrt{2}}\right) - \operatorname{erf}\left(\frac{\tau_1}{\sigma\ \sqrt{2}}\right)\right) = 0.075.$$

$$C = \frac{1}{3}\left(1 - \operatorname{erf}\left(\frac{\sqrt{2}}{\sigma}\left(\tau_2 - a\right)\right)\right) = 1.71\ 10^{-5}.$$