université PARIS-SACLAY

SYSTÈME DE CLASSIFICATION BAYESIEN NAÏF

Vincent Guigue
vincent.guigue@agroparistech.fr

MIA PARIS-SACLAY
EKINOCS

AGROPARISTECH
Institut des Sciences et Industries du Vivant et de l'Environnement

Lois de probabilités

Loi de Bernoulli

Définition

Épreuve de Bernoulli = expérience aléatoire qui ne peut prendre que deux résultats (*succès* et *échec*)

p = proba de succès, et $q = 1 - p$ = proba d'échec.

Loi de Bernoulli

Définition

Épreuve de Bernoulli = expérience aléatoire qui ne peut prendre que deux résultats (*succès* et *échec*)

p = proba de succès, et $q = 1 - p$ = proba d'échec.

Loi de Bernoulli

Variable X à support $\mathcal{X} = \{0, 1\}$ telle que :

$P(X = 1) = p$ et $P(X = 0) = 1 - p$

$E(X) = p \quad V(X) = p(1 - p)$

$\Longrightarrow X =$ le nombre de succès de l'épreuve de Bernoulli

Loi binomiale

Définition

Épreuve binomiale = expérience aléatoire telle que :

1. on répète n fois la même épreuve de Bernoulli,
2. les probas p et q restent inchangées pour chaque épreuve de Bernoulli,
3. les épreuves de Bernoulli sont toutes réalisées indépendamment les unes des autres.

Loi binomiale

Définition

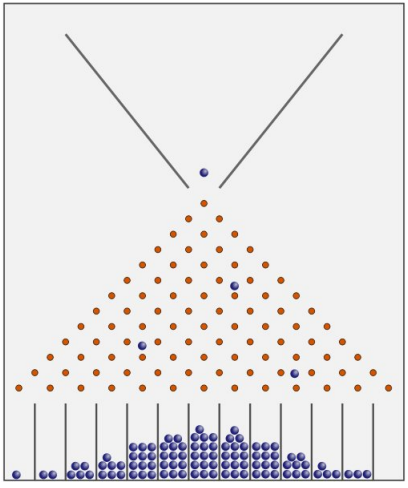
Épreuve binomiale = expérience aléatoire telle que :

1. on répète n fois la même épreuve de Bernoulli,
2. les probas p et q restent inchangées pour chaque épreuve de Bernoulli,
3. les épreuves de Bernoulli sont toutes réalisées indépendamment les unes des autres.

Loi binomiale de paramètres n et p

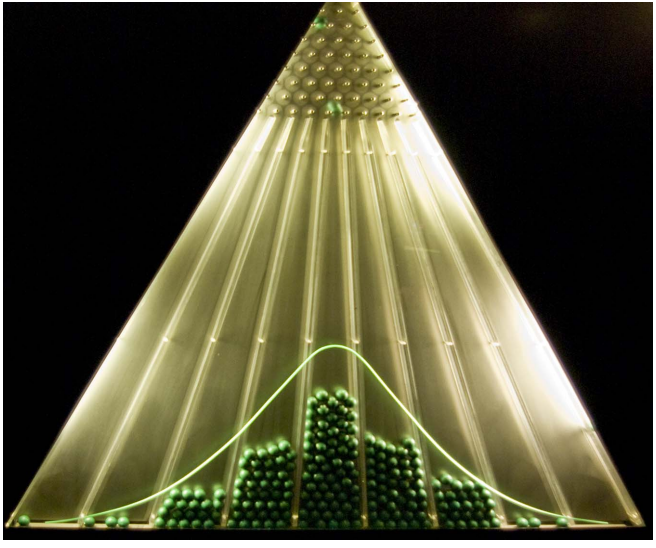
- $X =$ nombre de succès de l'épreuve binomiale
- $X \sim \mathcal{B}(n, p)$
- $P(X = k) = C_n^k \ p^k \ (1-p)^{n-k}, \forall k = 0, \ldots, n$
- $E(X) = np \quad V(X) = np(1-p)$

La planche de Galton



- chaque niveau $\Longrightarrow$ expérience de Bernoulli
- $\Longrightarrow X \sim$ loi binomiale

La planche de Galton



Loi normale

⚠ Loi extrêmement importante : souvent une très bonne approximation de la loi réelle

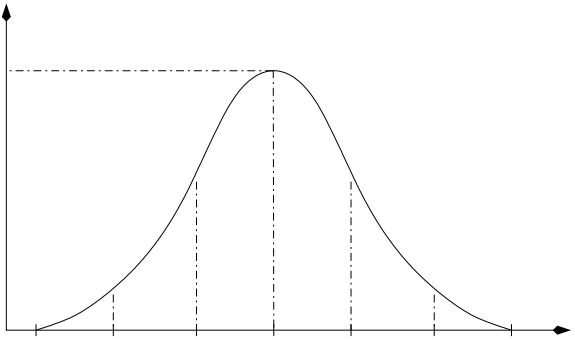
Définition : loi normale de paramètres μ et σ^2

- notée $\mathcal{N}(\mu, \sigma^2)$
- s'applique pour des variables aléatoires continues
- densité positive sur tout $\mathbb{R}$:

$$f(x) = \frac{1}{\sqrt{2\pi}.\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

- $E(X) = \mu \quad V(X) = \sigma^2$

Fonction de densité de la loi normale



Quelques reflexes :

- $2/3$ de la masse entre $+\sigma$ et $-\sigma$
- Support infini...

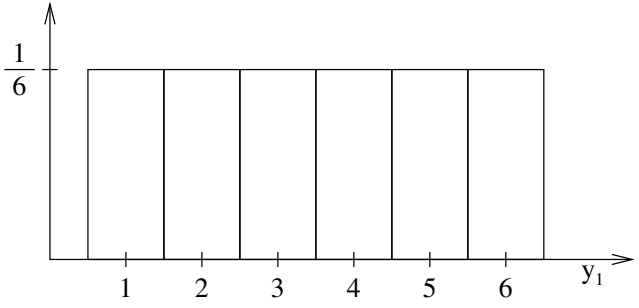
  Mais empiriquement $\sim$ toutes les observations entre $+3\sigma$ et -3σ
- Facile à dériver, à tronquer, ...

Loi normale = limite d'autres lois (1/4)

Lancés de dés à 6 faces

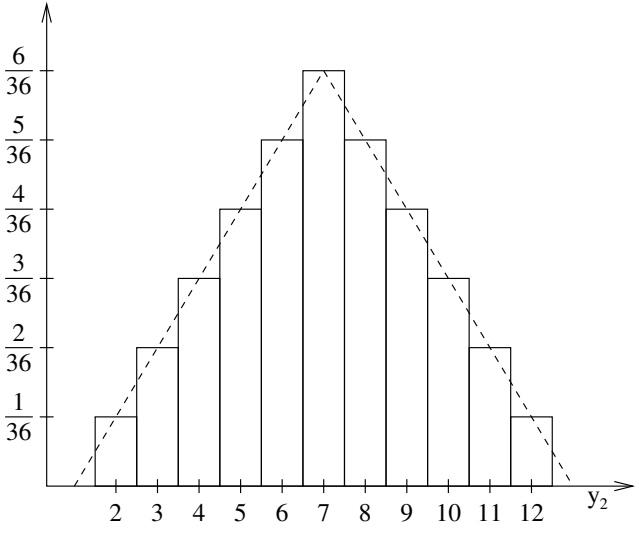
$\Longrightarrow$ on compte la somme des résultats des dés

Somme pour 1 jet de dé



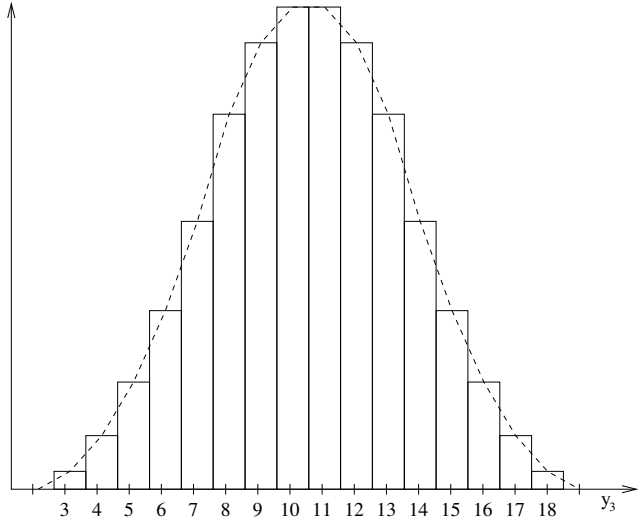
Loi normale = limite d'autres lois (2/4)

Somme pour 2 jets de dés



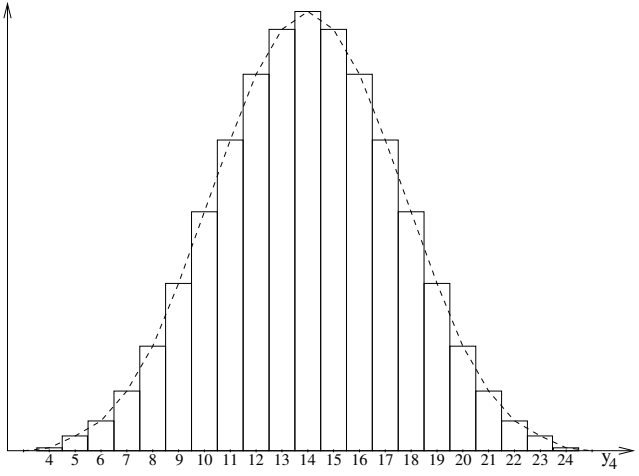
Loi normale = limite d'autres lois (3/4)

Somme pour 3 jets de dés



Loi normale = limite d'autres lois (4/4)

Somme pour 4 jets de dés



Loi normale en pratique

Théorème

$X \sim \mathcal{N}(\mu; \sigma^2)$

Alors la variable $Y = aX + b$ obéit à la loi $\mathcal{N}(a\mu + b; a^2\sigma^2)$.

$\Longrightarrow$ toute transformée affine d'une variable aléatoire suivant une loi normale suit aussi une loi normale

Loi normale en pratique

Théorème

$X \sim \mathcal{N}(\mu; \sigma^2)$

Alors la variable $Y = aX + b$ obéit à la loi $\mathcal{N}(a\mu + b; a^2\sigma^2)$.

$\Longrightarrow$ toute transformée affine d'une variable aléatoire suivant une loi normale suit aussi une loi normale

Corollaire

- X une variable aléatoire obéissant à une loi $\mathcal{N}(\mu; \sigma^2)$

  $\Longrightarrow Z = \frac{X - \mu}{\sigma}$ suit la loi $\mathcal{N}(0; 1)$
- Z suit une loi normale centrée (à cause de la moyenne en 0) réduite (à cause du σ^2 égal à 1)

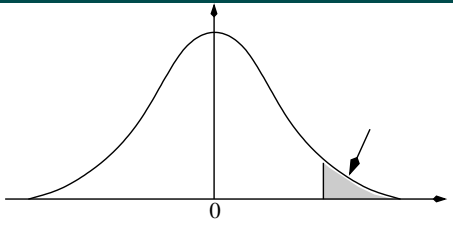
Loi normale en pratique (2)

Théorème

$X_1 \sim \mathcal{N}(\mu_1; \sigma_1^2), \qquad X_2 \sim \mathcal{N}(\mu_2; \sigma_2^2)$

Si les variables sont indépendantes, alors la variable $Y = X_1 + X_2$ obéit à la loi $\mathcal{N}(\mu_1 + \mu_2; \sigma_1^2 + \sigma_2^2)$.

Table de la loi normale centrée réduite



z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0859	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0466	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

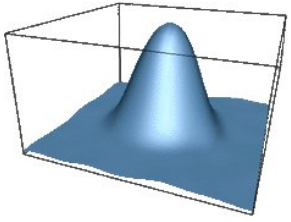
Loi normale bi-dimensionnelle

Définition : loi normale bi-dimensionnelle

- couple de variables (X, Y)
- densité dans $\mathbb{R}^2$:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

où $\rho = \frac{cov(X,Y)}{\sigma_x\sigma_y}$ = **coefficient de corrélation linéaire**



Théorème central-limite

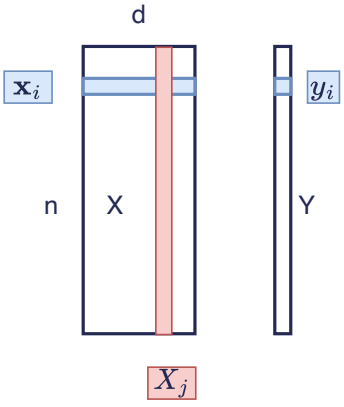
Théorème central-limite

- $(X_n)_{n\in\mathbb{N}}$: suite de variables
  - de même loi
  - d'espérance μ
  - de variance σ^2
  - **mutuellement** indépendantes
- alors la suite des moyennes empiriques centrées réduites $\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}$ tend en loi vers la loi normale centrée réduite :

$$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{loi} \mathcal{N}(0,1)$$

Naive Bayes

Notations et représentation des données



X matrice des données

- composée de n individus $\mathbf{x}_i \in \mathcal{X}$
- presque toujours, $\mathcal{X} = \mathbb{R}^d$

Y étiquettes des données, $y_i \in \mathcal{Y}$

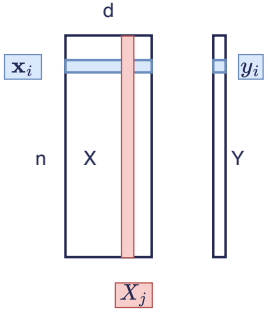
- $y_i \in \mathbb{R}$ $\Rightarrow$ régression
- $y_i \in \{1, \ldots, C\}$ $\Rightarrow$ classification en C catégories

Apprentissage automatique

A partir des données, construire une fonction f telle que :

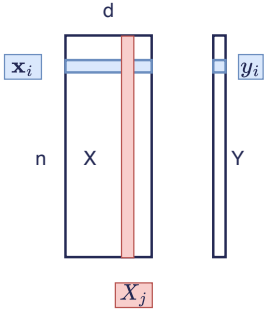
$$\forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, \qquad f(\mathbf{x}) \approx y$$

Algorithme naïf



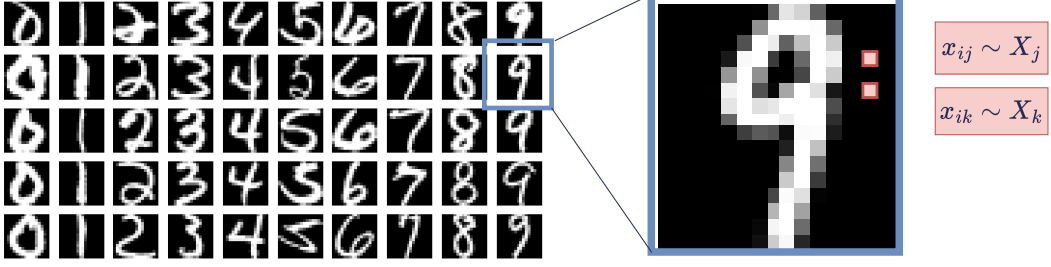
Hypothèse d'indépendance des variables descriptives X_j

Algorithme naïf



Hypothèse d'indépendance des variables descriptives X_j

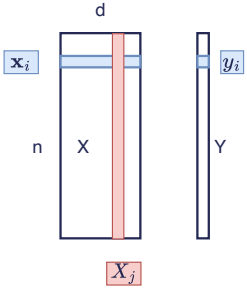
Pourquoi c'est très naïf ?



$x_{ij} \sim X_j$

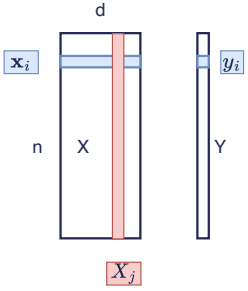
$x_{ik} \sim X_k$

Hypothèse variable par variable



1. Choix loi de probabilité pour (une ou toutes les) X_j
   e.g Bernoulli pour une image binaire : $X_j \sim Ber(p_j)$

Hypothèse variable par variable



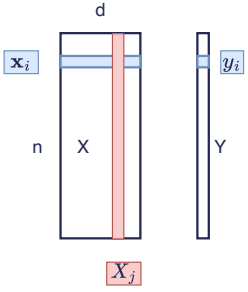
1. Choix loi de probabilité pour (une ou toutes les) X_j
   e.g Bernoulli pour une image binaire : $X_j \sim Ber(p_j)$

   $P(X_j = 1) = p_j \qquad P(X_j = 0) = 1 - p_j$

   Vraisemblance de l'observation x_{ij} :
   $P(X_j = x_{ij}) = p_j^{x_{ij}}(1 - p_j)^{(1-x_{ij})}$

2. Une variable descriptive $X_j \Rightarrow 1$ paramètre p_j
   On regroupe les paramètres : $\Theta = \{p_1, \ldots, p_d\}$
3. Optimisation des paramètres par max de vraisemblance

Hypothèse variable par variable



1. Choix loi de probabilité pour (une ou toutes les) X_j
   e.g Bernoulli pour une image binaire : $X_j \sim Ber(p_j)$

   $P(X_j = 1) = p_j \qquad P(X_j = 0) = 1 - p_j$

   Vraisemblance de l'observation x_{ij} :
   $P(X_j = x_{ij}) = p_j^{x_{ij}}(1 - p_j)^{(1-x_{ij})}$

2. Une variable descriptive $X_j \Rightarrow 1$ paramètre p_j
   On regroupe les paramètres : $\Theta = \{p_1, \ldots, p_d\}$

3. Optimisation des paramètres par max de vraisemblance

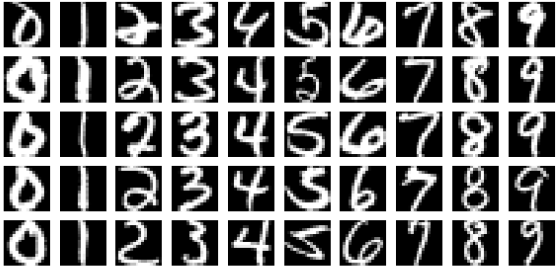
   Echantillon i.i.d + NB $\Rightarrow \mathcal{L}(X) = \prod_{i=1}^{n} \prod_{j=1}^{d} P(x_{ij}|\Theta)$

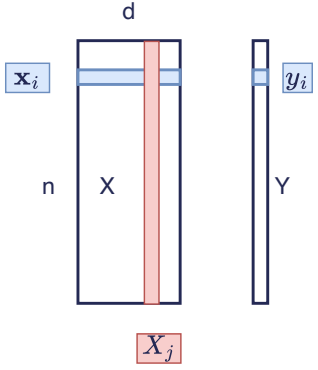
   Optimisation : $p_j^\star = \arg\max_{p_j} \mathcal{L}(X)$

Apprentissage statistique

Identification des paramètres optimaux correspondant aux observations

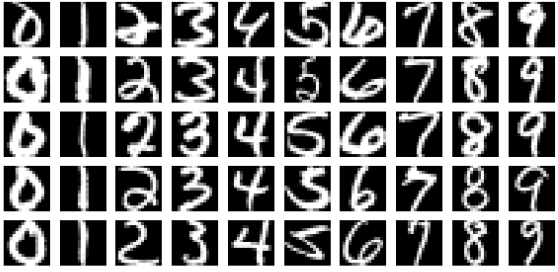
Cas de la classification bayesienne naïve

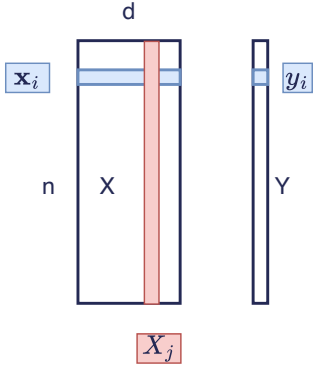




- 1 classe $C = 1$ sous-ensemble de données $=$ 1 modèle optimisé ($=$ un ensemble de paramètre Θ_c)
- C ($\times d$) problèmes d'optimisation distincts
- Combien de paramètres avec une modélisation de Bernoulli sur $d = 256$ pixels ?

Cas de la classification bayesienne naïve





- 1 classe $C = 1$ sous-ensemble de données $=$ 1 modèle optimisé ($=$ un ensemble de paramètre Θ_c)
- C ($\times d$) problèmes d'optimisation distincts
- Combien de paramètres avec une modélisation de Bernoulli sur $d = 256$ pixels ?
- $\Theta_c = \{p^\star_{c,1}, \ldots, p^\star_{c,d}\}$ et $\Theta = \{\Theta_1, \ldots, \Theta_c, \ldots, \Theta_C\} \Rightarrow$ 2560 paramètres

Calcul de la vraisemblance

- Pour une **valeur descriptive** , sous l'hypothèse de Bernoulli :

$$P(X_j = x_{ij}) = P(X_j = x_{ij}|p_j) = p_j^{x_{ij}}(1-p_j)^{(1-x_{ij})}$$

- Pour un **individu** , avec indépendance des variables descriptives :

$$P(\mathbf{x}_i) = P(\mathbf{x}_i|\Theta) = \prod_{j=1}^{d} P(X_j = x_{ij})$$

- Pour l'**échantillon** entier :
$\mathcal{L}(X) = \prod_{i=1}^{n}\prod_{j=1}^{d} P(x_{ij}|\Theta) = \prod_{i=1}^{n}\prod_{j=1}^{d} p_j^{x_{ij}}(1-p_j)^{(1-x_{ij})}$

$\mathcal{L}(X) \Rightarrow \log \mathcal{L}(X)$

La vraisemblance a en générale vocation à être dérivée pour trouver les paramètres optimaux... Comme le log est une fonction croissante :

$$\arg\max_{\Theta} \mathcal{L}(X) = \arg\max_{\Theta} \log \mathcal{L}(X)$$

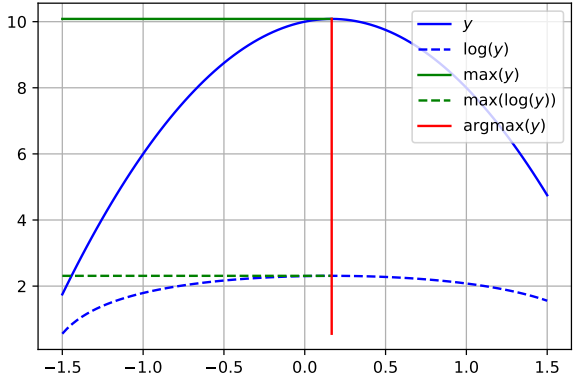
$\Rightarrow$ On travaille donc sur la log-vraisemblance, bien plus facile à dériver

Vraisemblance *vs* log-Vraisemblance

$$\mathcal{L}(X) = \prod_{i=1}^{n} \prod_{j=1}^{d} p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}$$

$$\log \mathcal{L}(X) =$$

$$\sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} \log(p_j) (1 - x_{ij}) \log(1 - p_j)$$



Laquelle des deux préférez-vous dériver par rapport à p_j ?

Apprentissage du modèle

Comment résoudre :
$p_j^\star = \arg\max_{p_j} \mathcal{L}(X) = \arg\max_{p_j} \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} \log(p_j)(1 - x_{ij}) \log(1 - p_j)$?

Solution 1

$$\frac{\partial \mathcal{L}_j(X)}{\partial p_j} = 0 \Leftrightarrow \ldots$$

$$p_j^\star = \ldots$$

Solution 2

Apprentissage du modèle

Comment résoudre :
$p_j^\star = \arg\max_{p_j} \mathcal{L}(X) = \arg\max_{p_j} \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} \log(p_j)(1 - x_{ij}) \log(1 - p_j)$?

Solution 1

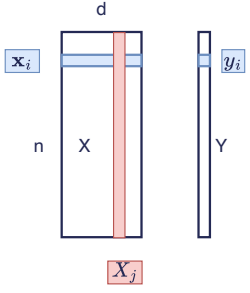
$$\frac{\partial \mathcal{L}_j(X)}{\partial p_j} = 0 \Leftrightarrow \ldots$$

$$p_j^\star = \ldots$$

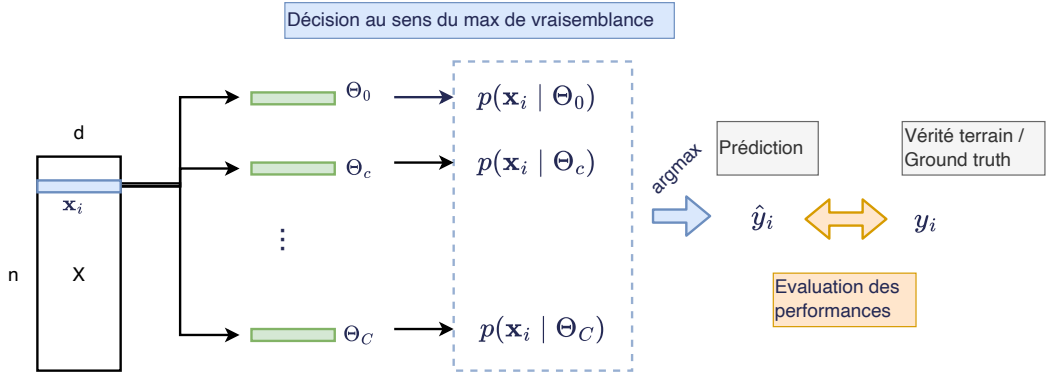
Solution 2

Je connais la loi de Bernoulli
(ou j'ai accès à wikipedia)

$$p_j^\star = \frac{\sum_i x_{ij}}{n}$$

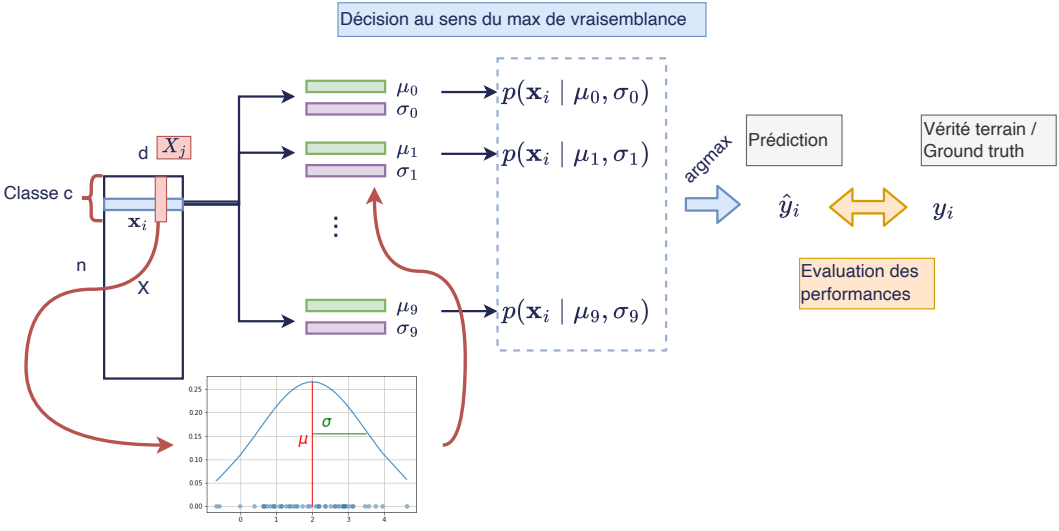


Inférence



Est-ce que la donnée $\mathbf{x}_i$ est plus vraisemblable sous le modèle de la classe 0, 1, ... ou C ?

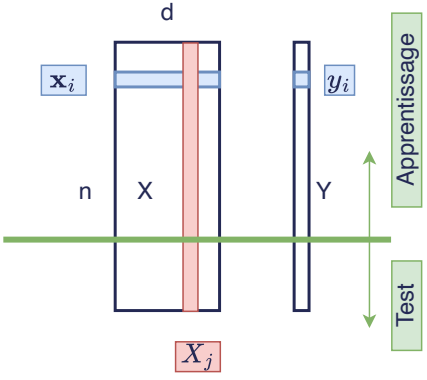
Passage à la gaussienne



Evaluation du modèle / Sélection de modèle

!! L'évaluation est aussi importante que l'apprentissage !!

- Evaluer sur les données d'apprentissage (=qui ont servi à régler les paramètres)
  $\Rightarrow$ **Tricherie, surestimation des performances**
- Evaluer sur des données vierges = OK



Problème de la répartition entre apprentissage et test

- La validation croisée

Evaluation du modèle / Sélection de modèle

!! L'évaluation est aussi importante que l'apprentissage !!

- Evaluer sur les données d'apprentissage (=qui ont servi à régler les paramètres)
  $\Rightarrow$ **Tricherie, surestimation des performances**
- Evaluer sur des données vierges = OK
- La validation croisée