IFT 7030,
Machine Learning for Signal Processing
Week2: Probability Refresher

Cem Subakan

UNIVERSITÉ LAVAL

Mila

Logistics

- Are you all on teams?
  - Est-ce que vous êtes tous sur teams?
- Second Lab is on friday! (12.30-14.20)
  - Session de lab ce vendredi. (12.30-14.20)
- Class Project Proposal: the deadline is early October. We will explain the class project proposal.
  - Proposition de projet de classe : la date limite est début octobre. On va expliquer la proposition de projet de classe.
- Please create an OpenReview account (`https://openreview.net`) using your Laval (or other institutional) email.
  - Créez un compte OpenReview (`https://openreview.net`) avec votre courriel Laval (ou autre courriel institutionnel).

Today's refresher

- Probability Calculus
  - Calculus de probabilité
- Discrete and Continuous random variables
  - Variables Aléatoires Discret, Continu
- Multivariate Random variables
  - Variables Aléatoires Multidimensionnelles

Today's refresher

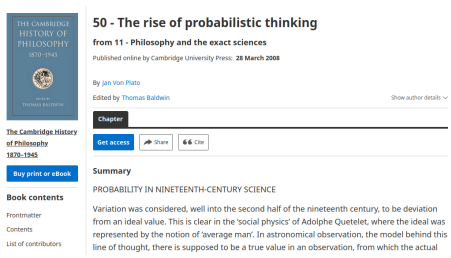
- Probability Calculus
  - Calculus de probabilité
- Discrete and Continuous random variables
  - Variables Aléatoires Discret, Continu
- Multivariate Random variables
  - Variables Aléatoires Multidimensionnelles
- Again, you don't need to understand everything today.
  - Vous n'etes pas obligés d'absorber tout maintenant. Les choses vont trouver leur context (j'espère au moins..)

What's probability / C'est quoi la probabilité là?

- It's a measure of belief.
  - C'est une mesure de certitude.

What's probability / C'est quoi la probabilité là?

- It's a measure of belief.
  - C'est une mesure de certitude.
- People didn't always think in probabilities. (e.g. Newtonian physics or your typical SP book :P)
  - On reflechissait pas de manière probabilistique toujours.



The Probabilistic Revolution: Ideas in the sciences

Lorenz Krüger, Gerd Gigerenzer, Mary S. Morgan
MIT Press, 1990 - Science - 480 pages
Winner in the category of Psychology in the 1987 Professional/Scholarly Publishing Annual Awards Competition presented by the Association of

This monumental work traces the rise, the transformation, and the diffusion of probabilistic and statistical thinking in the nineteenth and twentieth centuries. complete and so successful.
More »

Goals of probability / Les buts du probabilité

- Characterize stochastic processes
  - On characterize processus stochastique
- Understanding a die, coin toss / Comprendre un dé, lancer à pile ou face
- What will be my next word? / Qu'est-ce que je vais dire maintenant?

Table of Contents

Let's start with an example

Let's start with an example

Let's start with an example

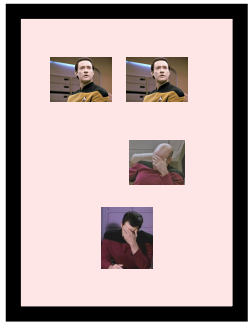
Let's start with an example

- Characters $= \{\text{data}, \text{picard}, \text{riker}\}$
- Boxes $= \{\text{red}, \text{yellow}\}$

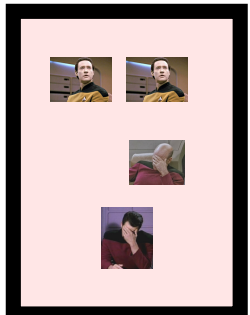
Let's start with an example

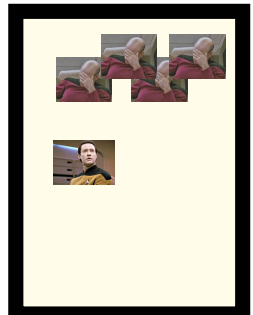
- Characters $= \{\text{data}, \text{picard}, \text{riker}\}$
- Boxes $= \{\text{red}, \text{yellow}\}$
- Let's pick a character photo, and then ask questions!
  - Prenons une photo, pis posons des questions!

Questions



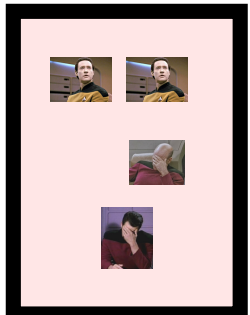
Questions

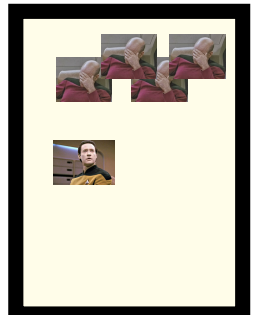




- What is the probability of picking ‘data’?
  - Quelle est la probabilité de choisir ‘data’?

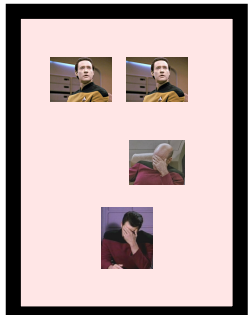
Questions

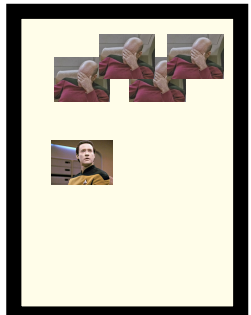




- What is the probability of picking ‘data’?
  - Quelle est la probabilité de choisir ‘data’?
- What is the probability of picking from the red box?
  - Quelle est la probabilité de choisir de la boite rouge?

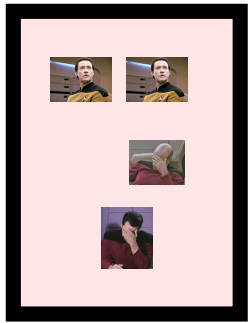
Questions





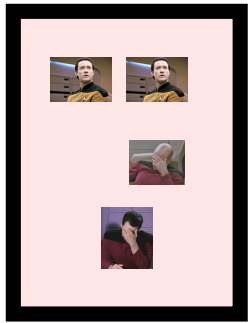
- What is the probability of picking ‘data’?
  - Quelle est la probabilité de choisir ‘data’?
- What is the probability of picking from the red box?
  - Quelle est la probabilité de choisir de la boite rouge?
- What is the probability of picking data, given red box?
  - Quelle est la probabilité de choisir data, étant donné la boite rouge?

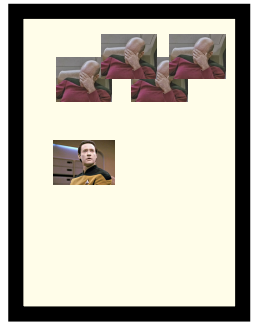
Questions



- What is the probability of picking 'data'?
  - Quelle est la probabilité de choisir 'data'?
- What is the probability of picking from the red box?
  - Quelle est la probabilité de choisir de la boite rouge?
- What is the probability of picking data, given red box?
  - Quelle est la probabilité de choisir data, étant donné la boite rouge?
- What is the probability of picking data and red box?
  - Quelle est la probabilité de choisir data et la boite rouge?

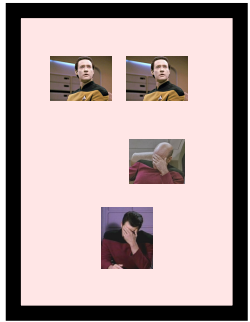
Questions

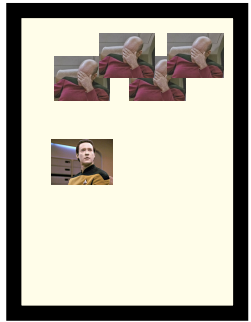




- What is the probability of picking 'data'?
  - Quelle est la probabilité de choisir 'data'? 1/3 maybe?
- What is the probability of picking from the red box?
  - Quelle est la probabilité de choisir de la boite rouge? 4/9 maybe?
- What is the probability of picking data, given red box?
  - Quelle est la probabilité de choisir data, étant donné la boite rouge? 2/4?
- What is the probability of picking data and red box?
  - Quelle est la probabilité de choisir data et la boite rouge? 2/9 or 2/8? I am getting confused..

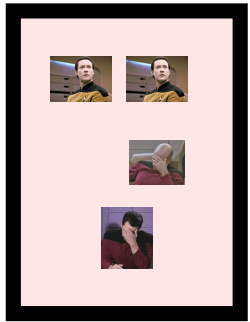
Let's do some modeling / Faisons du modeling

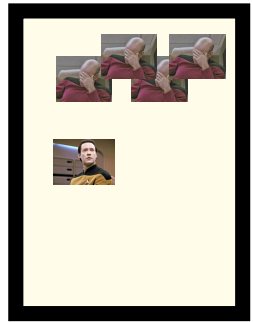




	data	picard	riker
red box			
yellow box			

Let's do some modeling / Faisons du modeling

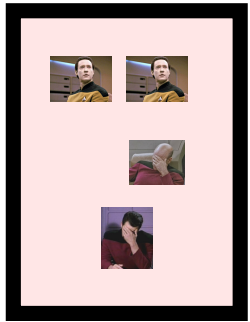


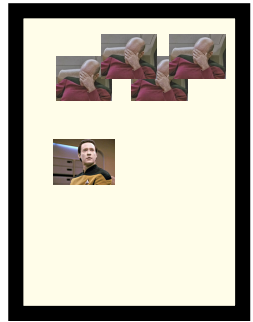


	data	picard	riker
red box	2/4	1/4	1/4
yellow box			

What are these probabilities do you think? / Quelles sont cettes probabilités?

Let's do some modeling / Faisons du modeling





	data	picard	riker
red box	2/4	1/4	1/4
yellow box	1/5	4/5	0/5

What are these probabilities do you think? / Quelles sont cettes probabilités?

Probability Tables

- These are conditional probabilities. That is $p(c|b)$.
  - Ces cont des probabilités conditionelles.

	data	picard	riker
red box			
yellow box			

Probability Tables

- These are conditional probabilities. That is $p(c|b)$.
  - Ces cont des probabilités conditionelles.

	data	picard	riker
red box	2/4	1/4	1/4
yellow box			

- How do we get to $p(c, b)$, that is the joint probability table?
  - Comment obtiens-t-on la table jointe de probabilité?

Probability Tables

- These are conditional probabilities. That is $p(c|b)$.
  - Ces cont des probabilités conditionelles.

	data	picard	riker
red box	2/4	1/4	1/4
yellow box	1/5	4/5	0/5

- How do we get to $p(c, b)$, that is the joint probability table?
  - Comment obtiens-t-on la table jointe de probabilité?
- We define a prior distribution on the boxes.
  - On definit une distribution a-priori sur les boites.

$$p(c, b) = \begin{matrix} p(b = rd) \\ p(b = yw) \end{matrix} \odot \underbrace{\begin{matrix} 2/4 & 1/4 & 1/4 \\ 1/5 & 4/5 & 0/5 \end{matrix}}_{p(c|b)}$$

Probability Tables

- These are conditional probabilities. That is $p(c|b)$.
  - Ces cont des probabilités conditionelles.

	data	picard	riker
red box	2/4	1/4	1/4
yellow box	1/5	4/5	0/5

- How do we get to $p(c, b)$, that is the joint probability table?
  - Comment obtiens-t-on la table jointe de probabilité?
- We define a prior distribution on the boxes.
  - On definit une distribution a-priori sur les boites.

$$p(c, b) = \begin{matrix} p(b = rd) \\ p(b = yw) \end{matrix} \odot \underbrace{\begin{matrix} 2/4 & 1/4 & 1/4 \\ 1/5 & 4/5 & 0/5 \end{matrix}}_{p(c|b)}$$

- Einsten notation anyone?
  - Notation Einstein?
- How do we pick the prior then?
  - Comment chosit-on le prior?

Joint distribution and beyond

- Kinda makes sense to use 0.5 red 0.5 yellow.
  - Fait du sens 0.5 0.5.

$$\begin{matrix} 2/8 & 1/8 & 1/8 \\ 1/10 & 4/10 & 0/10 \end{matrix} = \begin{matrix} 0.5 \\ 0.5 \end{matrix} \odot \begin{matrix} 2/4 & 1/4 & 1/4 \\ 1/5 & 4/5 & 0/5 \end{matrix}$$

Joint distribution and beyond

- Kinda makes sense to use 0.5 red 0.5 yellow.
  - Fait du sens 0.5 0.5.

$$\begin{matrix} 2/8 & 1/8 & 1/8 \\ 1/10 & 4/10 & 0/10 \end{matrix} = \begin{matrix} 0.5 \\ 0.5 \end{matrix} \odot \begin{matrix} 2/4 & 1/4 & 1/4 \\ 1/5 & 4/5 & 0/5 \end{matrix}$$

- This gives us the answer to the 'and' question. (e.g. red box and data)
  - Ça nous donne 'et' (ex: boite rouge et data)

Joint distribution and beyond

- Kinda makes sense to use 0.5 red 0.5 yellow.
  - Fait du sens 0.5 0.5.

$$\begin{matrix} 2/8 & 1/8 & 1/8 \\ 1/10 & 4/10 & 0/10 \end{matrix} = \begin{matrix} 0.5 \\ 0.5 \end{matrix} \odot \begin{matrix} 2/4 & 1/4 & 1/4 \\ 1/5 & 4/5 & 0/5 \end{matrix}$$

- This gives us the answer to the 'and' question. (e.g. red box and data)
  - Ça nous donne 'et' (ex: boite rouge et data)
- Let's answer the question, what's the probability of getting data?
  - C'est quoi la probabilité de choisir data?

Marginal distribution

- We can simply sum over one variable.
  - On calcule la somme sur un variable.

$$
\begin{aligned}
p(c = \text{data}) = & \sum_{b \in \{r,y\}} p(c = \text{data}, b) \\
= & p(c = \text{data}, b = \text{red}) + p(c = \text{data}, b = \text{yellow}) \\
= & 1/4 + 1/10 = 7/20
\end{aligned}
$$

- But hmmm, this is not the same as $1/3$ as we thought earlier... What's happening?
  - Pas la meme resultat qu'avant.. Qu'est-ce qui se passe là là?

Bayesian vs Frequentist Approach

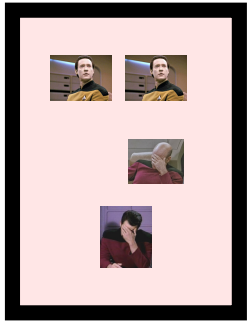
- Frequentist Approach Estimates the Probabilities, Doesn't assume anything.
  - Approche frequentist estime les probabilités. On suppose pas un modèle.

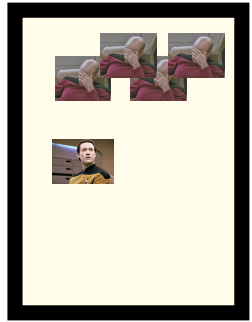
	data	picard	riker
red box			
yellow box			

- Note that $p(c = data) = 1/3$ as we thought before.
  - Notez, le probabilité marginal de data est $1/3$ comme prévu avant.

Bayesian vs Frequentist Approach

- Frequentist Approach Estimates the Probabilities, Doesn't assume anything.
  - Approche frequentist estime les probabilités. On suppose pas un modèle.



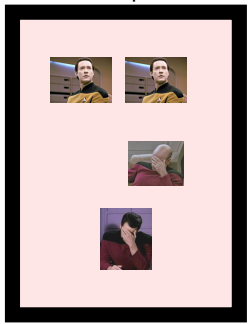


	data	picard	riker
red box	2/9	1/9	1/9
yellow box			

- So, what we did before was bonkers?
  - Alors, ce qu'on a calculé avant était faux?

Bayesian vs Frequentist Approach

- Frequentist Approach Estimates the Probabilities, Doesn't assume anything.
  - Approche frequentist estime les probabilités. On suppose pas un modèle.



	data	picard	riker
red box	2/9	1/9	1/9
yellow box	1/9	4/9	0/9

- So, what we did before was bonkers?
  - Alors, ce qu'on a calculé avant était faux?

Bayesian Approach

- We do not always observe a lot of data.
  - On n'observe pas toujours beaucoup de données.
- Bayesian approach is a modeling approach. We construct a model to describe the process.
  - Approche Bayesien est le cheminement du modeling. On construit un modèle.
- All models are wrong. But some are useful.
  - Tout les modèles sont faux, mais quelqu'uns sont utiles.

$$p(c, b) = \begin{matrix} p(b = rd) \\ p(b = yw) \end{matrix} \odot \underbrace{\begin{matrix} 2/4 & 1/4 & 1/4 \\ 1/5 & 4/5 & 0/5 \end{matrix}}_{p(c|b)}$$

  - This one can be useful, if we want to construct a general model!
    - Peut-etre ce modèle est utile aussi, si on veut construire un modèle générale.

Bayesian vs Frequentist

- Will a météor hit earth?
  - Est-ce qu'un météor va frapper la terre?

Bayesian vs Frequentist

- Will a météor hit earth?
  - Est-ce qu'un météor va frapper la terre?
- Frequentist: Let's wait until N is large.
  - Fréquentist: Attendons que N soit large.

Bayesian vs Frequentist

- Will a météor hit earth?
  - Est-ce qu'un météor va frapper la terre?
- Frequentist: Let's wait until N is large.
  - Fréquentist: Attendons que N soit large.
- Bayesian: Let's construct a model, and try to predict.
  - Bayesien: Ça sera bien de construire un modèle, et prédire.

Probability Calculus

- Sum Rule:

$$p(c) = \sum_b p(c, b)$$

$$p(b) = \sum_c p(c, b)$$

- Product Rule:

$$p(c, b) = p(c|b)p(b) = p(b|c)p(c)$$

- Sum to 1:

$$\sum_c p(c) = 1$$

$$\sum_b p(b) = 1$$

$$\sum_{b,c} p(c, b) = 1$$

Probability Calculus

- Sum Rule:

$$p(c) = \sum_b p(c, b)$$

$$p(b) = \sum_c p(c, b)$$

- Product Rule:

$$p(c, b) = p(c|b)p(b) = p(b|c)p(c)$$

- Sum to 1:

$$\sum_c p(c) = 1$$

$$\sum_b p(b) = 1$$

$$\sum_{b,c} p(c, b) = 1$$

Joint $p(c, b)$ as a matrix — marginals are the row / column sums:

$p(b) = \sum_c p(c, b)$

	data	picard	riker	$p(b)$
red box	2/9	1/9	1/9	4/9
yellow box	1/9	4/9	0/9	5/9
$p(c)$	3/9	5/9	1/9	

$p(c) = \sum_b p(c, b)$

Product Rule

- The joint factorizes as a (row) marginal times a conditional:

$$p(b)\,p(c|b) = p(c,b)$$

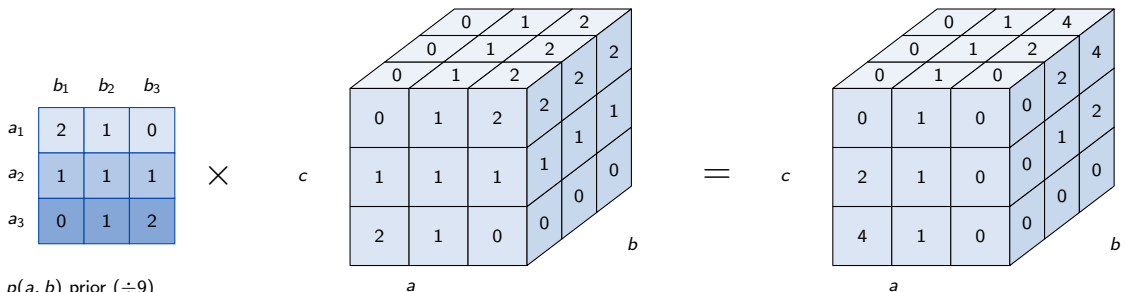
$p(b)$ $\quad$ $p(c|b)$ (rows sum to 1) $\quad$ $p(c,b)$

Each row of the joint = that box's probability $p(b)$ times its conditional $p(c|b)$.

Product Rule (chain rule) — 3D cube

- Each distribution over (a, b, c) is a $3 \times 3 \times 3$ cube (values shown on the visible faces):

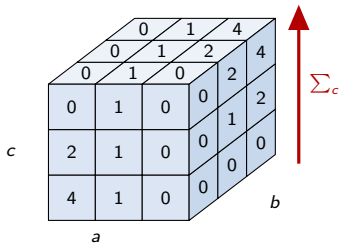
$$\underbrace{p(a,b)}_{\div 9}\ \underbrace{p(c|a,b)}_{\div 3} = \underbrace{p(a,b,c)}_{\div 27}$$



Cube faces: top $= c_3$, front $= b_1$, right $= a_3$. Each joint cell = prior $\times$ conditional.

Sum Rule — 3D cube & tensor index notation

- Marginalising = summing the joint cube over one axis. Sum over c (collapse the vertical axis):



$p(a, b, c)$ joint cube ($\div 27$)

$\xrightarrow{\sum_c}$

	b_1	b_2	b_3
a_1	6	3	0
a_2	3	3	3
a_3	0	3	6

$p(a, b) = \sum_c p(a, b, c)$ ($\div 27$, i.e. the prior $p(a, b)$)

Index (tensor) notation

The joint is a rank-3 tensor P_{abc}; marginalising = contracting an index:

$$p(a, b) = \sum_c p(a, b, c) \iff P_{ab} = \sum_c P_{abc} = P_{abc}\,\mathbf{1}_c.$$

Contracting an index lowers the rank ($3 \to 2$); e.g. $p(a) = \sum_{b,c} P_{abc} = P_{abc}\,\mathbf{1}_b\mathbf{1}_c$.

Bayes Rule

- We invert the conditioning.
  - On reverse la partie sachant.

$$p(b|c) = \frac{p(c|b)p(b)}{\sum_b p(c|b)p(b)}$$

$$= \frac{\overbrace{p(c|b)}^{\text{likelihood}}\overbrace{p(c)}^{\text{prior}}}{\underbrace{p(c)}_{\text{normalizing constant}}}$$

Bayes Rule

- We invert the conditioning.
  - On reverse la partie sachant.

$$p(b|c) = \frac{p(c|b)p(b)}{\sum_b p(c|b)p(b)}$$

$$= \frac{\overbrace{p(c|b)}^{\text{likelihood}}\overbrace{p(c)}^{\text{prior}}}{\underbrace{p(c)}_{\text{normalizing constant}}}$$

Let's calculate the posterior

- What is the probability of picking from red box, given data? $p(b|c)$.
  - Quelle est la probabilité de choisir de la boite rouge, sachant data?

	data	picard	riker
red	2/9	1/9	1/9
yellow	1/9	4/9	0/9
$p(c)$	3/9	5/9	1/9

$p(c, b)$, with $p(c) = \sum_b p(c, b)$

Let's calculate the posterior

- What is the probability of picking from red box, given data? $p(b|c)$.
  - Quelle est la probabilité de choisir de la boite rouge, sachant data?

	data	picard	riker
red	2/9	1/9	1/9
yellow	1/9	4/9	0/9
$p(c)$	3/9	5/9	1/9

$p(c, b)$, with $p(c) = \sum_b p(c, b)$

divide each column by $p(c)$

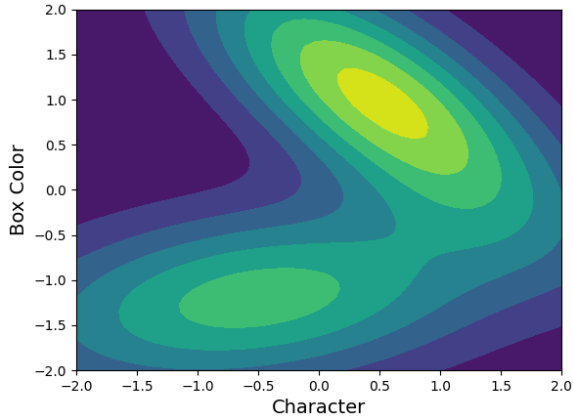
	data	picard	riker
red	2/3	1/5	1
yellow	1/3	4/5	0
sum	1	1	1

$p(b|c)$ (columns sum to 1)

Table of Contents

Continuous Random Variables

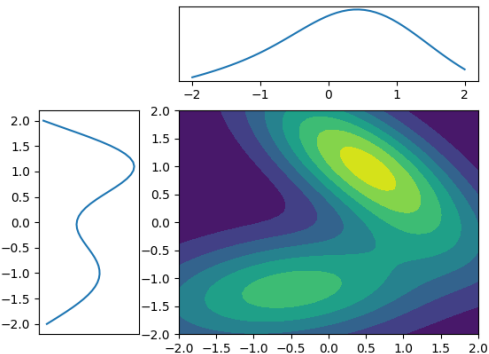
- What if we have infinite number of different characters in TNG (and beyond), and we explore all the color spectrum for the boxes?
  - Imaginons qu'on avait un numbre infini de personnages, et pis on veut explorer tout les couleurs dans le spectre la lumiere.



Same rules!

- Well, the same probability rules apply.
  - Bon, on a les memes reglès de probabilité.
- Sum Rule:

$$p(c) = \int p(c, b)db, \quad p(b) = \int p(c, b)dc$$



Same rules

- Product Rule

$$p(c, b) = p(c|b)p(b)$$

  - Note that the above are continuous objects now.
    - Notez que les objects sont des fonctions continu maintenant.

- Bayes Rule

$$p(b|c) = \frac{p(c|b)p(b)}{\int p(c|b)p(b)db} = \frac{\overbrace{p(c|b)}^{\text{likelihood}}\overbrace{p(c)}^{\text{prior}}}{\underbrace{\int p(c, b)db}_{\text{normalizing constant}}}$$

Probability Density Function

- $p(c)$ no longer gives probability values, but rather density values.
  - on n'a plus des valeur de probabilités $p(c)$

Probability Density Function

- $p(c)$ no longer gives probability values, but rather density values.
  - on n'a plus des valeur de probabilités $p(c)$
- $0 \leq p(c) \leq \infty$

Probability Density Function

- $p(c)$ no longer gives probability values, but rather density values.
  - on n'a plus des valeur de probabilités $p(c)$
- $0 \leq p(c) \leq \infty$
- $\int_{\infty}^{\infty} p(c) = 1$

Probability Density Function

- $p(c)$ no longer gives probability values, but rather density values.
  - on n'a plus des valeur de probabilités $p(c)$
- $0 \leq p(c) \leq \infty$
- $\int_{\infty}^{\infty} p(c) = 1$
- $P(x_a \leq c \leq x_b) = \int_{x_a}^{x_b} p(c)dc$

Probability Density Function

- $p(c)$ no longer gives probability values, but rather density values.
  - on n'a plus des valeur de probabilités $p(c)$
- $0 \leq p(c) \leq \infty$
- $\int_{\infty}^{\infty} p(c) = 1$
- $P(x_a \leq c \leq x_b) = \int_{x_a}^{x_b} p(c)dc$
- $P(c \leq x) = \int_{-\infty}^{x} p(c)dc$
  $\rightarrow$ This is the Cumulative Distribution Function!

Some common measures 1

- Expectation

$$\mathbb{E}[f(x)] = \int f(x)p(x)dx$$

  - Or for discrete distributions

$$\mathbb{E}[f(x)] = \sum_x f(x)p(x)$$

- Expected value (Center of gravity)

$$\mathbb{E}[x] = \int xp(x)dx$$

Some common measures 2

- Variance

$$\begin{aligned}\text{var}(x) =& \mathbb{E}[(x - \mathbb{E}[x])^2] = \int (x - \mathbb{E}[x])^2 p(x) dx \\ =& \mathbb{E}[x^2] - (\mathbb{E}[x])^2\end{aligned}$$

- Variance for Multivariate Distributions (we will call the result covariance matrix)
  - La matrice de covariance

$$\begin{aligned}\text{covar}(x) =& \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top] \\ =& \mathbb{E}[xx^\top] - \mathbb{E}[x](\mathbb{E}[x])^\top\end{aligned}$$

Table of Contents

List of popular distributions

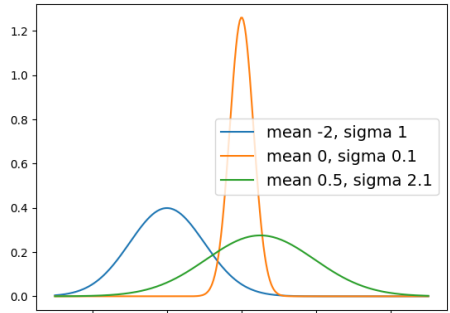
- Gaussian
  - It's everywhere. C'est partout. Apprenez-le avec votre coeur!
- Poisson
  - Useful for count data. Utile quand on compte
- Gamma
  - Useful for non-negative continuous data. Utile pour valeur non-negatifs.
- Laplacian
  - Sometimes shows up. Ça arrive des fois.
- Beta-Dirichlet
  - Useful to model discrete distributions. Ca modelise les distributions discrets.
- Exponential Family
  - A general form of distributions. Une forme générale des distributions.

The mighty Gaussian

- The bell curve / Normal distribution

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

- This is essentially a distribution built around the scaled euclidean distance. (The other stuff is so that the thing is normalized)
  - Dans le fond, c'est une distribution bati autour de la distance euclidienne à l'échelle. Les autres trucs sont là pour que la chose est normalisée.

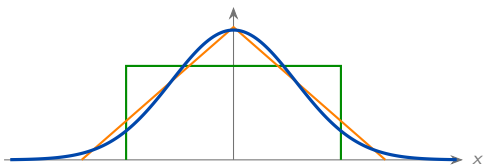


Central Limit Theorem

- The average of many i.i.d. random variables becomes Gaussian — whatever their own distribution.
  - La moyenne de plusieurs variables i.i.d. devient gaussienne, peu importe leur loi.
- For X_i i.i.d. with mean μ and variance σ^2:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1), \qquad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- This is why the Gaussian shows up everywhere (noise, measurement errors, ...).



n=1 (uniform) n=2 $n\to\infty$ (Gaussian)

Distribution of the standardized sum, sharpening to a bell.

What students want

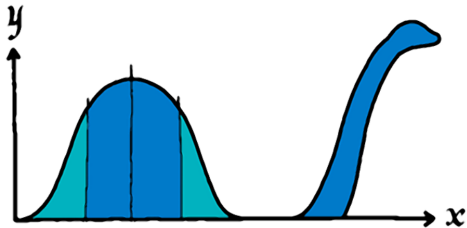


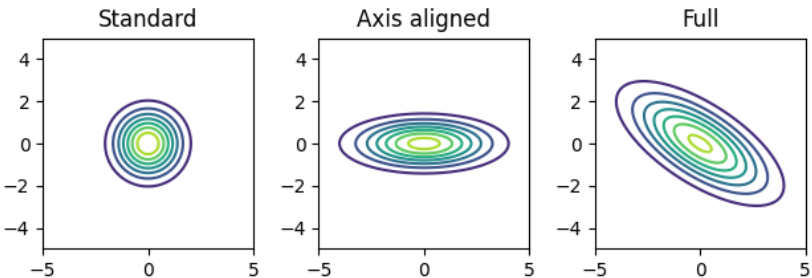
fig 1.0 The Extended Bell Curve.

– by Tang Yau Hoong

Multivariate Gaussian

- Same distribution but defined over vectors
  - La meme distribution mais définit sur les vecteurs

$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

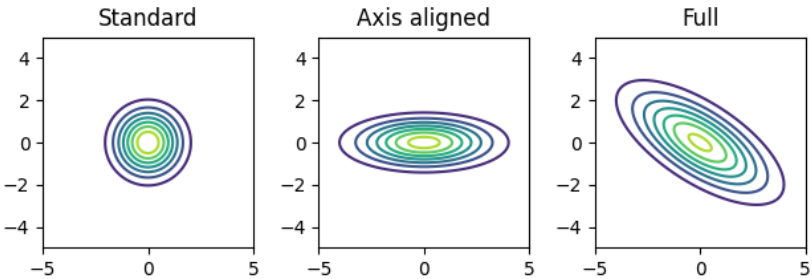


- Mu: $\mu := \mathbb{E}[x] \in \mathbb{R}^L$
- Sigma: $\Sigma := \mathbb{E}[xx^\top] - \mathbb{E}[x]\mathbb{E}[x]^\top \in \mathbb{R}^{L \times L}$

Multivariate Gaussian

- Same distribution but defined over vectors
  - La meme distribution mais définit sur les vecteurs

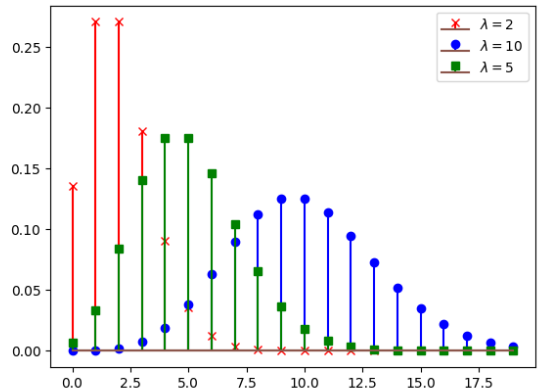
$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$



- Mu: $\mu := \mathbb{E}[x] \in \mathbb{R}^L$
- Sigma: $\Sigma := \mathbb{E}[xx^\top] - \mathbb{E}[x]\mathbb{E}[x]^\top \in \mathbb{R}^{L \times L}$
- Note that $x^\top \Sigma x \geq 0$, $\forall x$. x is positive-semi-definite.

Poisson Distribution

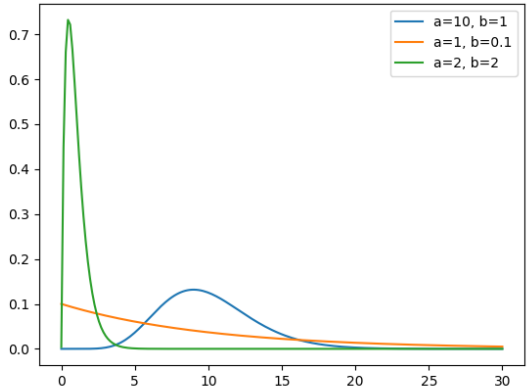
- It's useful to model count data (it's a discrete distribution)
  - C'est utile pour modéliser le comptage (c'est une distribution discrete)



- $\mathcal{PO}(x, \lambda) = \exp(-\lambda)\frac{\lambda^x}{x!}$
- $\mathbb{E}[x] = \lambda$

Gamma Distribution

- Good for modeling non-negative data
  - C'est utile pour modéliser du data non-negatif



- $\mathcal{G}(x; a, b) = \frac{1}{\Gamma(a)} \beta^{\alpha} x^{a-1} \exp(-\beta x)$
- $\mathbb{E}[x] = \frac{a}{b}$

General Discrete Distribution

- The density is written with the following expression.
  - La densité est écrit avec l'expression suivante

$$\text{Discrete}(x, \pi) = \prod_{k=1}^{K} \pi_k^{[k=x]}$$

- $[k = j]$ is an Iverson bracket. Output is one when the argument is true.
  - C'est une notation Iverson. La sortie est 1 quand l'argument est vrai.
- Note that $x \in \{0, 1\}^K$.
- $\mathbb{E}[x] = \pi$

Dirichlet Distribution

- This is a distribution over discrete probability distributions (so continuous distribution). (π parameter in the slide before)
  - Ça donne une distribution sur des distributions discrets.
- This is a distribution defined over a simplex. That is $\sum_i x_i = 1$.

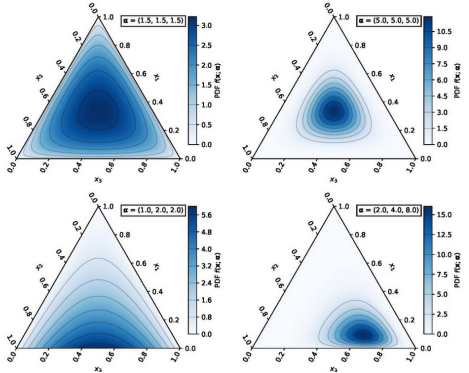


Image Taken from wikipedia.

- $\text{Dirichlet}(x; \mathbf{a}) = \frac{1}{B(\mathbf{a})} \prod_i x_i^{a_i - 1}$.
- $\mathbb{E}[x_i] = \frac{x_i}{\alpha_0}$, $\alpha_0 := \sum_i \alpha_i$.

Exponential Family

- The general form

$$p(x|\theta) = h(x) \exp\left(\eta(\theta) T(x) - A(\theta)\right)$$

- The distributions we discussed fit in exponential family. You can see that by pattern matching.
  - Les distributions dont on a parlé sont dans la famille exponentielle.

Exponential Family

- The general form

$$p(x|\theta) = h(x) \exp\left(\eta(\theta) T(x) - A(\theta)\right)$$

- The distributions we discussed fit in exponential family. You can see that by pattern matching.
  - Les distributions dont on a parlé sont dans la famille exponentielle.
- $T(x)$, is the sufficifient statistics. If you have that, you don't need to know anything else.
  - $T(x)$ décrit la distribution au complet.

Exponential Family

- The general form

$$p(x|\theta) = h(x) \exp\left(\eta(\theta) T(x) - A(\theta)\right)$$

- The distributions we discussed fit in exponential family. You can see that by pattern matching.
  - Les distributions dont on a parlé sont dans la famille exponentielle.
- $T(x)$, is the sufficifient statistics. If you have that, you don't need to know anything else.
  - $T(x)$ décrit la distribution au complet.
- Conjugate priors: Some prior-likelihood pairs yield the same type in the posterior distribution. Useful for Bayesian inference. (Gaussian-Gaussian, Gamma-Poisson, Dirichlet-Discrete, more..)
  - On aura le meme type de prior que le posterior. Utile dans l'inférence Bayesienne.

If you want other distributions

- http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

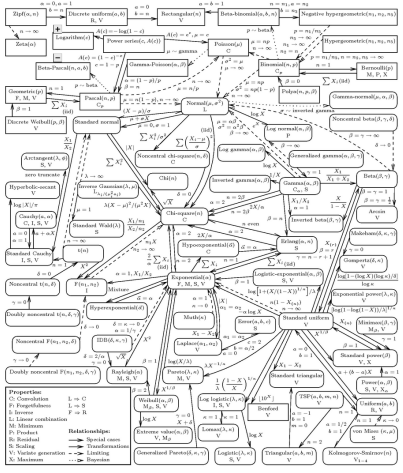


Table of Contents

Parameter Estimation

- Ok, but so what?
  - We want to explain the data with them.
    - On aimerait expliquer les données avec ces differents types de distributions.
- For that, we need to fit these distributions to data. (Note that these distributions are basically models)
  - On doit fitter ces distributions aux données. (Notez que ces distributions sont essentiellement des modèles)
- Fitting a model means doing 'parameter estimation'.
  - Fitter un modèle veut dire 'estimation des paramètres'.
- We have different ways of going about parameter estimation.
  - On a different cheminents pour aboutir à cela.

Parameter Estimation

- Given some independent samples
  - Étant donné des échantillons indépendents

$$X = \{x_1, x_2, \ldots, x_N\}$$

Parameter Estimation

- Given some independent samples
  - Étant donné des échantillons indépendents

$$X = \{x_1, x_2, \ldots, x_N\}$$

- And a model

$$p(x; \theta)$$

Parameter Estimation

- Given some independent samples
  - Étant donné des échantillons indépendents

$$X = \{x_1, x_2, \ldots, x_N\}$$

- And a model

$$p(x; \theta)$$

- We estimate a set of parameters θ.

Maximum Likelihood

- The likelihood is defined as,
  - On définit le likelihood comme le suivant:

$$p(X;\theta) = p(x_1;\theta)p(x_2;\theta)\dots p(x_N;\theta) = \prod_{i=1}^{N} p(x_i;\theta)$$

- The problem to solve:
  - Le problème à resoudre:

$$\widehat{\theta}_{ML} = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i;\theta)$$

- Any idea how to solve it?
  - Des idées?

Calculate the gradient

- We calculate the gradient, and set it to zero, solve for θ.
  - On va calculer le gradient, et resoudre pour θ.
- We work in the log domain to make things simpler. log is monotonic function so it's fine.
  - On travaille dans le domaine de logarithme pour faire les choses plus simple.

$$\begin{aligned}\log p(X;\theta) &= \log \prod_{i=1}^{N} p(x_i;\theta) \\ &= \sum_{i=1}^{N} \log p(x_i;\theta)\end{aligned}$$

- Let's do an example.
  - Faisons une exemple.

Estimate μ with ML

$$\begin{aligned}\log p(X;\theta) &= \sum_i \log \mathcal{N}(x_i;\mu,\Sigma) \\ &= \sum_i \left(-\frac{1}{2}(\mathbf{x}_i-\mu)^\top \Sigma^{-1}(\mathbf{x}_i-\mu)\right) \\ &\quad -\frac{1}{2}\log\det\Sigma - \frac{K}{2}\log 2\pi\end{aligned}$$

Estimate μ with ML

$$
\begin{aligned}
\log p(X;\theta) &= \sum_i \log \mathcal{N}(x_i;\mu,\Sigma) \\
&= \sum_i \left(-\frac{1}{2}(\mathbf{x}_i-\mu)^\top \Sigma^{-1}(\mathbf{x}_i-\mu) \right) \\
&\quad - \frac{1}{2}\log\det\Sigma - \frac{K}{2}\log 2\pi \\
&=^{+} \sum_i \left(-\frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1} x_i - 2\mu^\top \Sigma^{-1} x_i + \mu^\top \Sigma^{-1}\mu \right)
\end{aligned}
$$

Estimate μ with ML

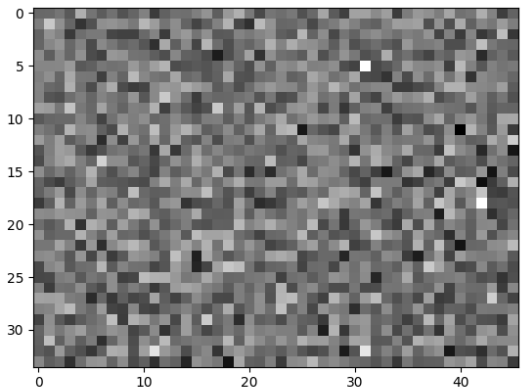
$$
\begin{aligned}
\log p(X;\theta) &= \sum_i \log \mathcal{N}(x_i;\mu,\Sigma) \\
&= \sum_i \left(-\frac{1}{2}(\mathbf{x}_i-\mu)^\top \Sigma^{-1}(\mathbf{x}_i-\mu) \right) \\
&\quad - \frac{1}{2}\log\det\Sigma - \frac{K}{2}\log 2\pi \\
&=^{+} \sum_i \left(-\frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1} x_i - 2\mu^\top \Sigma^{-1} x_i + \mu^\top \Sigma^{-1}\mu \right) \\
\frac{\partial \log p(X;\theta)}{\partial \mu} &= \sum_i \left(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu \right) \\
&= -2\Sigma^{-1}\left(\sum_{i=1}^{N} x_i \right) + 2N\Sigma^{-1}\mu \\
&\rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} x_i
\end{aligned}
$$

All this to get just this?

- Well, it's not always this straightforward result.
  - On ne va pas avoir un résultat si simple.
- In most cases we will not have a closed-form result. (Gradient Descent anyone?)
  - Dans le majorité de cas on n'aura pas un résultat simple.

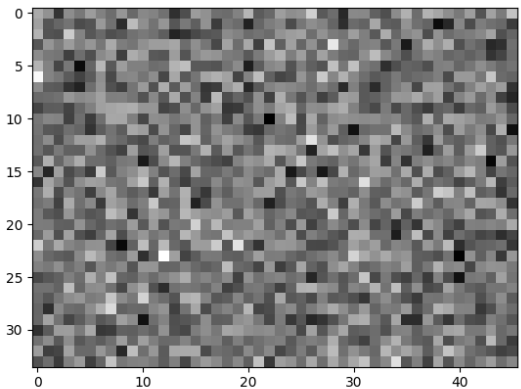
Estimation in Action

- Let's say we observe these data
  - Disons qu'observe cela



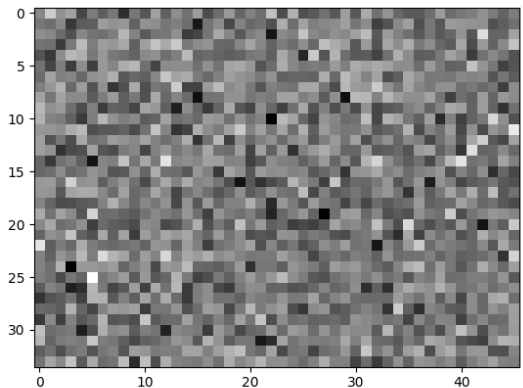
Estimation in Action

- Let's say we observe these data
  - Disons qu'observe cela



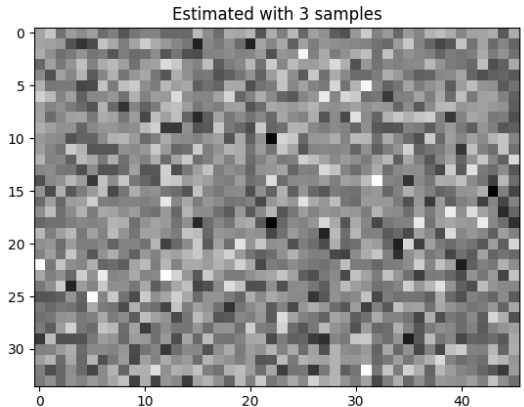
Estimation in Action

- Let's say we observe these data
  - Disons qu'observe cela



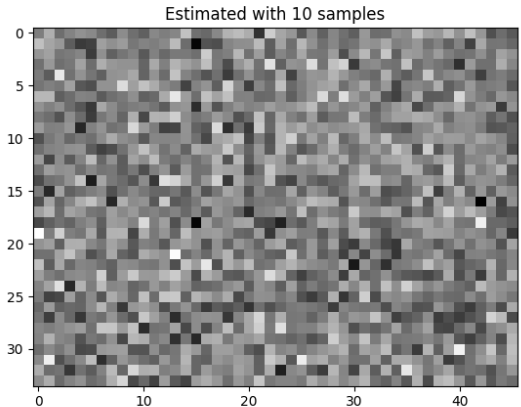
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



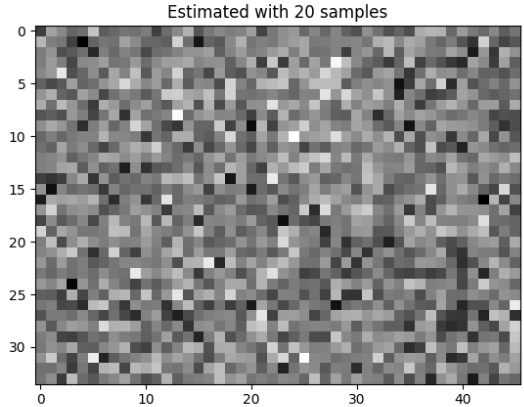
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



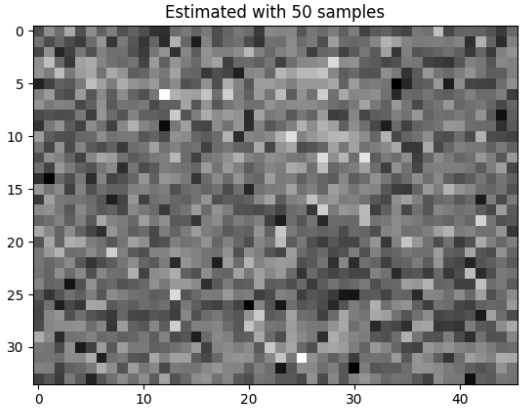
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



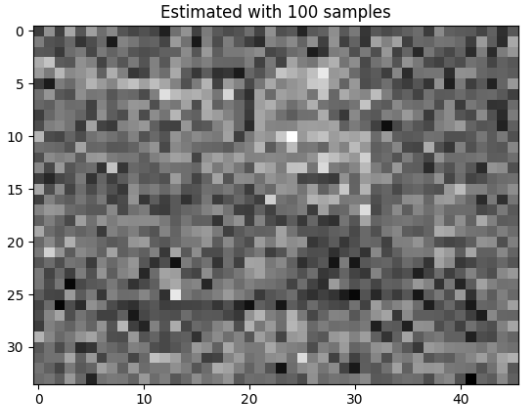
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



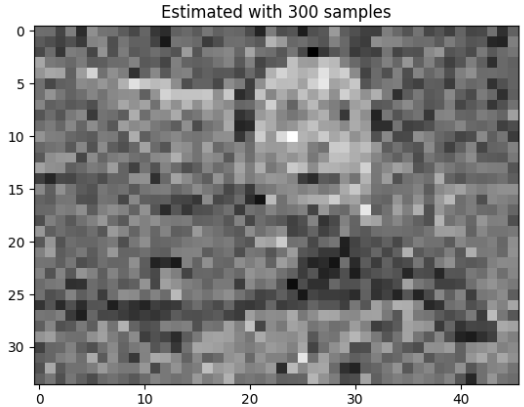
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



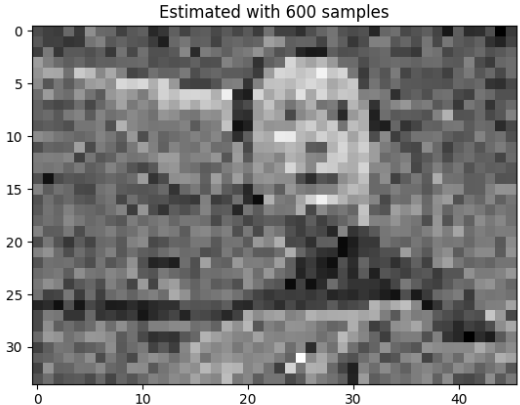
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



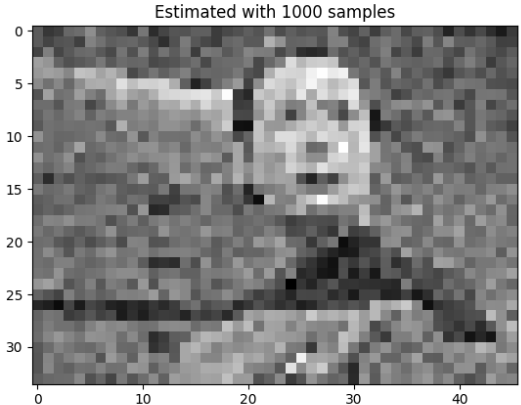
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



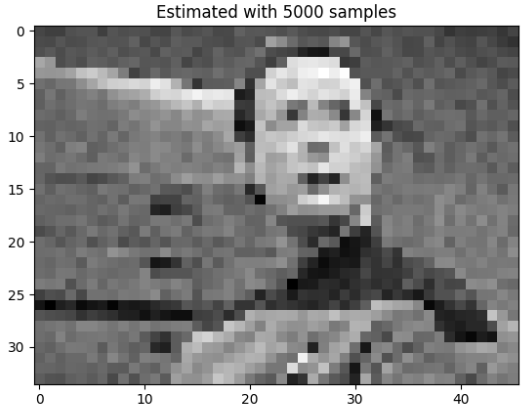
Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



Let's fit a Gaussian!

- We will fit a Gaussian, and plot the mean parameter, for different N.



Maximum A Posteriori Estimation (MAP)

- Sometimes we want to inject prior information.
  - Ça arrive que des fois on veux injecter de l'information a priori.
- We will do that by using a prior
  - On va incorporer une distribution a priori.

Maximum A Posteriori Estimation (MAP)

- Sometimes we want to inject prior information.
  - Ça arrive que des fois on veux injecter de l'information a priori.
- We will do that by using a prior
  - On va incorporer une distribution a priori.
- Note that,

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

- Then we can,

$$\arg\max_{\theta} p(X|\theta)p(\theta)$$

Estimate μ with MAP

$$\begin{aligned}\log p(X;\theta) + \log p(\theta) &= \sum_i \log \mathcal{N}(x_i;\mu,\Sigma) + \log \mathcal{N}(\mu;\mu_0,\Sigma_0) \\ &= \sum_i \left(-\frac{1}{2}(\mathbf{x}_i-\mu)^\top \Sigma^{-1}(\mathbf{x}_i-\mu)\right) \\ &\quad -\frac{1}{2}\log\det\Sigma - \frac{K}{2}\log 2\pi + \left(-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0)\right)\end{aligned}$$

Estimate μ with MAP

$$
\begin{aligned}
\log p(X;\theta) + \log p(\theta) &= \sum_i \log \mathcal{N}(x_i;\mu,\Sigma) + \log \mathcal{N}(\mu;\mu_0,\Sigma_0) \\
&= \sum_i \left(-\frac{1}{2}(\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu) \right) \\
&\quad - \frac{1}{2}\log\det\Sigma - \frac{K}{2}\log 2\pi + \left(-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0) \right) \\
&=^+ \sum_i \left(-\frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1} x_i - 2\mu^\top \Sigma^{-1} x_i + \mu^\top \Sigma^{-1}\mu \right) + \frac{1}{2}\mu^\top \Sigma^{-1}\mu - \mu^\top \Sigma_0^{-1}\mu_0
\end{aligned}
$$

Estimate μ with MAP

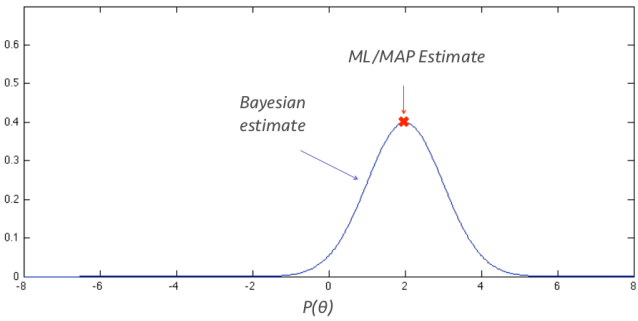
$$
\begin{aligned}
\log p(X;\theta) + \log p(\theta) &= \sum_i \log \mathcal{N}(x_i;\mu,\Sigma) + \log \mathcal{N}(\mu;\mu_0,\Sigma_0) \\
&= \sum_i \left(-\frac{1}{2}(\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu) \right) \\
&\quad - \frac{1}{2}\log\det\Sigma - \frac{K}{2}\log 2\pi + \left(-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0) \right) \\
=^+ \sum_i \left(-\frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1} x_i - 2\mu^\top \Sigma^{-1} x_i + \mu^\top \Sigma^{-1}\mu \right) &+ \frac{1}{2}\mu^\top \Sigma^{-1}\mu - \mu^\top \Sigma_0^{-1}\mu_0 \\
\frac{\partial \log p(X;\theta)}{\partial \mu} &= \sum_i \left(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu \right) + 2\Sigma_0^{-1}\mu - 2\Sigma_0^{-1}\mu_0 \\
&= -2\Sigma^{-1}\left(\sum_{i=1}^N x_i \right) + 2N\Sigma^{-1}\mu + 2\Sigma_0^{-1}\mu - 2\Sigma_0^{-1}\mu_0 \\
&\rightarrow \mu = \left(N\Sigma^{-1} + \Sigma_0^{-1} \right)^{-1} \left(\Sigma^{-1}\left(\sum_{i=1}^N x_i \right) + \Sigma_0^{-1}\mu_0 \right)
\end{aligned}
$$

MAP and ML

- If $p(\theta)$ is close to uniform, MAP and ML give the same result.
  - Si le prior est uniform, MAP et ML donnent la meme résultat.

Full Bayesian Inference

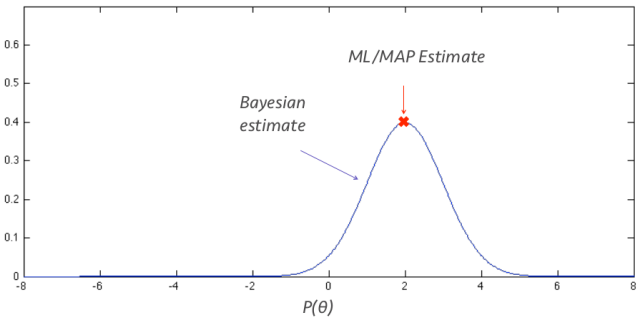
- In Full Bayesian inference we work with distribution over the parameters.
  - Dans Inférence Full-Bayesienne on travail avec une distribution complete sur les paramètres



Taken from UIUC MLSP class slides

Full Bayesian Inference

- In Full Bayesian inference we work with distribution over the parameters.
  - Dans Inférence Full-Bayesienne on travail avec une distribution complete sur les paramètres



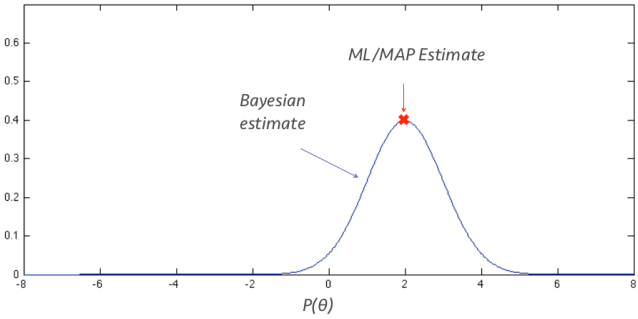
Taken from UIUC MLSP class slides

- Posterior-Predictive Distribution

$$p_{\text{pred}}(x) = \int p(x;\theta)p(\theta|x)d\theta$$

Full Bayesian Inference

- In Full Bayesian inference we work with distribution over the parameters.
  - Dans Inférence Full-Bayesienne on travail avec une distribution complete sur les paramètres



Taken from UIUC MLSP class slides

- Posterior-Predictive Distribution

$$p_{\text{pred}}(x) = \int p(x;\theta)p(\theta|x)d\theta$$

  - Notice that this is some sort of ensembling
    - C'est un genre d'ensemblent des predicteurs.

The same exercise in the same setup

- $p(X|\theta) = \prod_i \mathcal{N}(x_i; \mu, \Sigma)$
- $p(\theta) = \mathcal{N}(\mu; \mu_0, \Sigma_0)$

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

- The posterior will be Gaussian. Do it in the log-domain, do pattern matching.
  - Le posterior sera Gaussien. Tu peux faire la somme en log, et fais du pattern-matching.

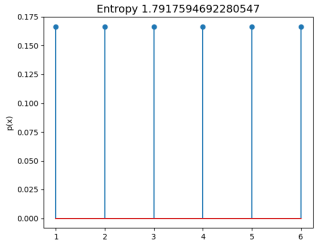
Table of Contents

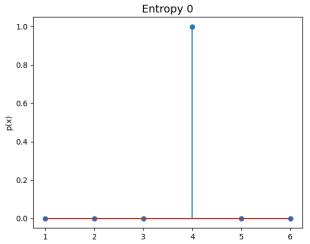
Entropy

- Is a measure of uncertainty.
  - Est un mésure d'incertitude.
- $H(x) = -\int p(x) \log p(x) dx$

Entropy

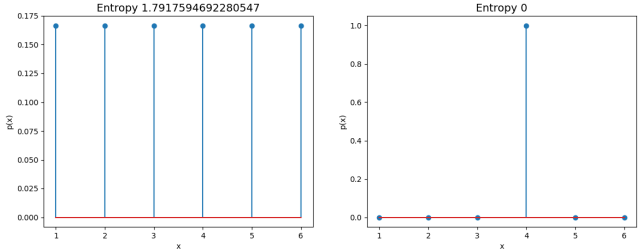
- Is a measure of uncertainty.
  - Est un mésure d'incertitude.
- $H(x) = -\int p(x) \log p(x) dx$





Entropy

- Is a measure of uncertainty.
  - Est un mésure d'incertitude.
- $H(x) = -\int p(x) \log p(x) dx$



- Kullback-Leibler (KL) divergence:

$$D(p(x)\|q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx$$
$$= \underbrace{-H(x)}_{\text{entropy of } p} + \underbrace{H(p,q)}_{\text{crossentropy } p,q}$$

Recap

- We saw the probability calculus.
  - Sum, Product, Bayes rules.
    - On vu le calculus de probabilité.

Recap

- We saw the probability calculus.
  - Sum, Product, Bayes rules.
    - On vu le calculus de probabilité.
- We learnt about typical distributions.
  - On a vu les distributions communs.

Recap

- We saw the probability calculus.
  - Sum, Product, Bayes rules.
    - On vu le calculus de probabilité.
- We learnt about typical distributions.
  - On a vu les distributions communs.
- We learnt how to do parameter estimation.
  - ML, MAP, Full-Bayesian
    - On a vu comment faire estimation de parametres.

Recommended Reading

- Bishop, Pattern Recognition and Machine Learning, Chapters 1, 2

What's Next

- All of signal processing in one lecture.
  - Tout le traitement du signal dans une session.