Introduction to model-based learning

Pierre-Alexandre Mattei

Inria, Université Côte d'Azur

Other lectures will be given by Charles Bouveyron and Aude Sportisse

Menu of this intro

1. What's a statistical model? Why are they useful?
2. A recap on probability distributions
3. A recap on inference techniques, in particular maximum likelihood

Statistical models?

- A **statistical model** is a family of probability distribution that we want to train to perform a suitable task

- We'll begin with a few examples. The most common machine learning task is **classification**, where we want to predict a **label** $y \in \{0, 1\}$ based on some **features** $x \in \mathcal{X}$.

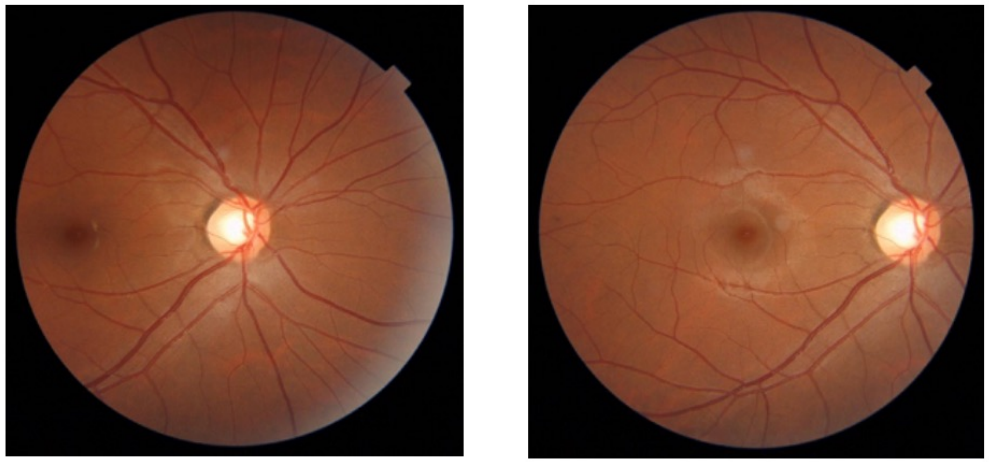
- Let's look at a specific example: **predicting if a patient has an illness** called diabetic retinopathy using an image of their retina.

Statistical models?

- A **statistical model** is a family of probability distribution that we want to train to perform a suitable task

- We want to predict a **label** $y \in \{0, 1\}$ based on some **features** $x \in \mathcal{X}$.

No apparent retinopathy



Optic disc as the center | The macula is the center

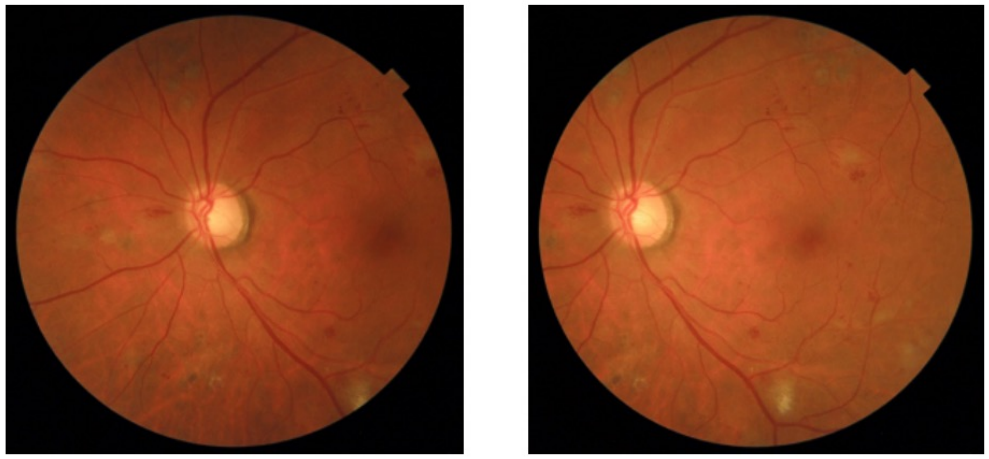
https://isbi.deepdr.org/data.html

Statistical models?

- A **statistical model** is a family of probability distribution that we want to train to perform a suitable task

- We want to predict a **label** $y \in \{0, 1\}$ based on some **features** $x \in \mathcal{X}$.

Severe Non-proliferative diabetic retinopathy



Optic disc as the center

The macula is the center

https://isbi.deepdr.org/data.html

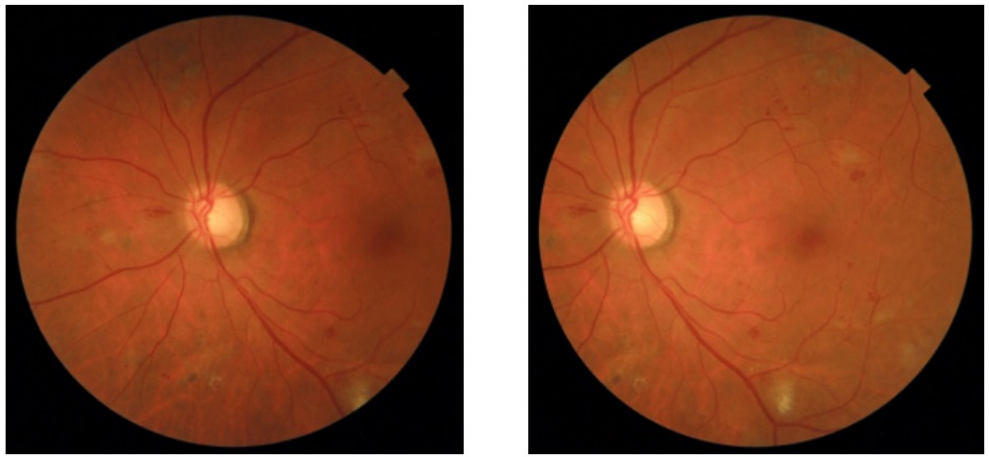
Statistical models?

Q: What probability distribution would we like to learn here?

- A **statistical model** is a family of probabili…e want to train to perform a suitable task

- We want to predict a **label** $y \in \{0, 1\}$ based on some **features** $x \in \mathcal{X}$.

Severe Non-proliferative diabetic retinopathy



Optic disc as the center

The macula is the center

https://isbi.deepdr.org/data.html

Statistical models?

Q: What probability distribution would we like to learn here?

- A **statistical model** is a family of probabil e want to train to perform a suitable task
- We want to predict a **label** $y \in \{0, 1\}$ based on some **features** $x \in \mathcal{X}$.
- The quantity

$$p_{\text{data}}(y = 1 | x = \text{  })$$

is the probability of being sick, and is exactly what we need for classification.

Statistical models?

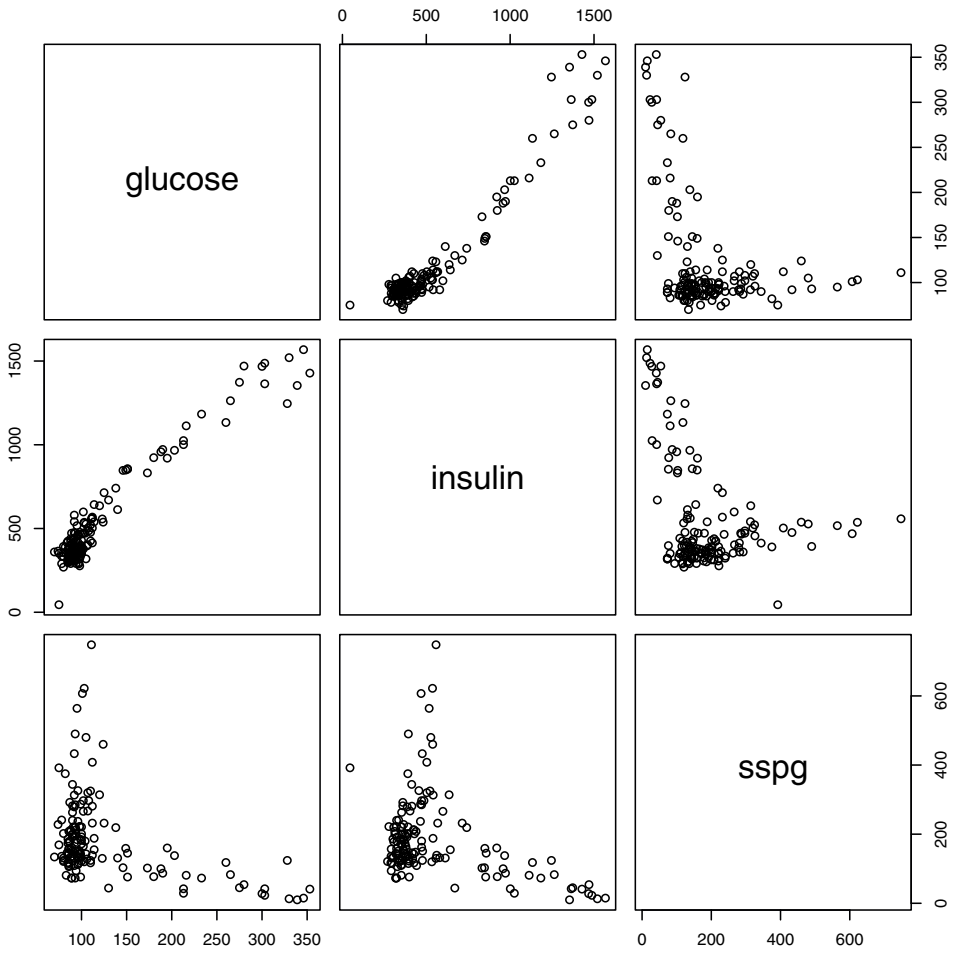
- What is nice about knowing $p_{\text{data}}(y = 1 | x =$  $)$ is that it

would allow us to **quantify the uncertainty of our predictions**, in other words, to « know when we don't know ».

Statistical models?

- An example that we'll study a lot in this course in **clustering.**

- It's the same problem as classification, but **we don't have labels** in our training data, and we often don't even know the number of classes (aka **clusters**).

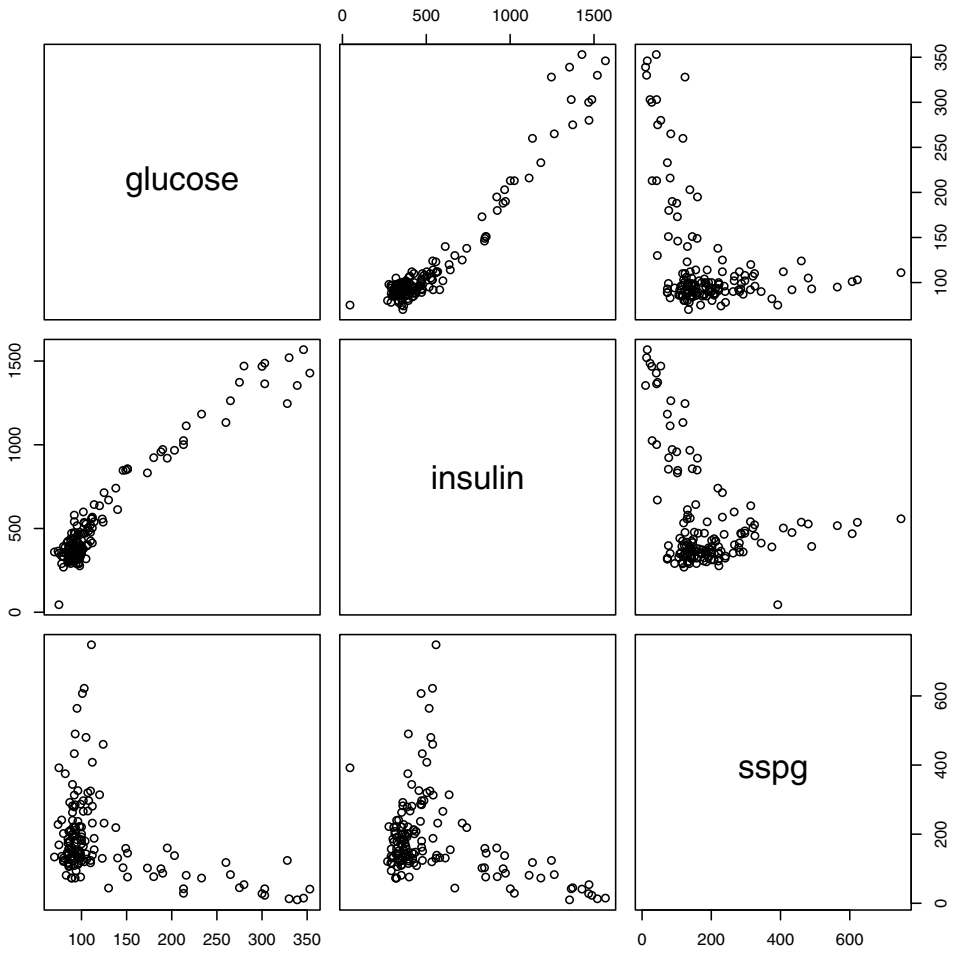


Statistical models?

- An example that we'll study a lot in this course in **clustering.**

- It's the same problem as classification, but **we don't have labels** in our training data, and we often don't even know the number of classes (aka **clusters**).

- Here are some electronic health records. How can we group them?

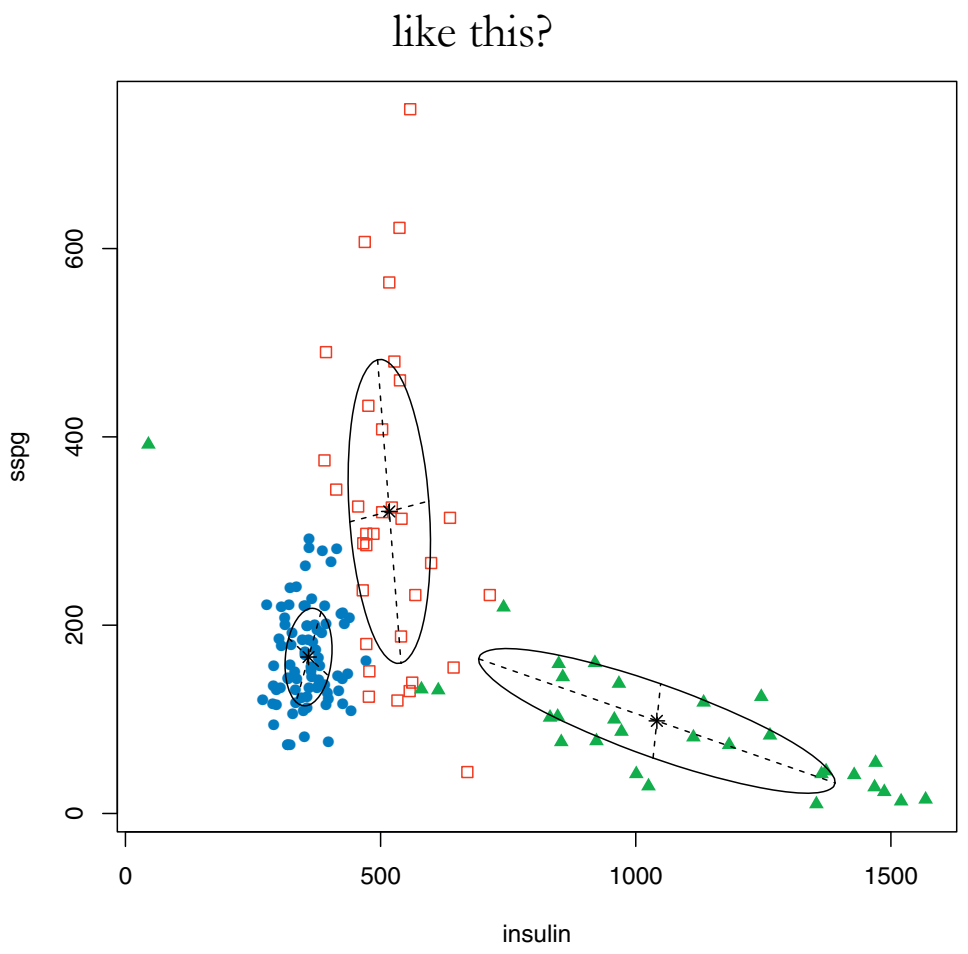


Statistical models?

- An example that we'll study a lot in this course in **clustering.**

- It's the same problem as classification, but **we don't have labels** in our training data, and we often don't even know the number of classes (aka **clusters**).

- Here are some electronic health records. How can we group them?



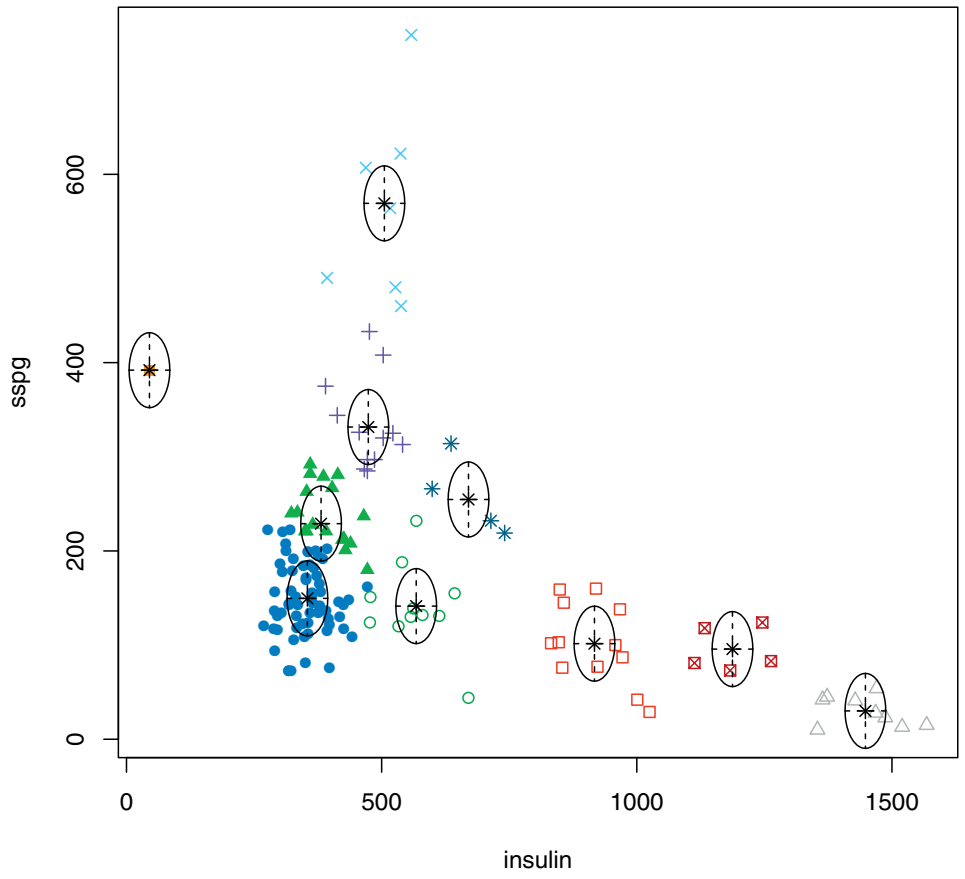
Statistical models?

- An example that we'll study a lot in this course in **clustering.**

- It's the same problem as classification, but **we don't have labels** in our training data, and we often don't even know the number of classes (aka **clusters**).

- Here are some electronic health records. How can we group them?

…or like this?

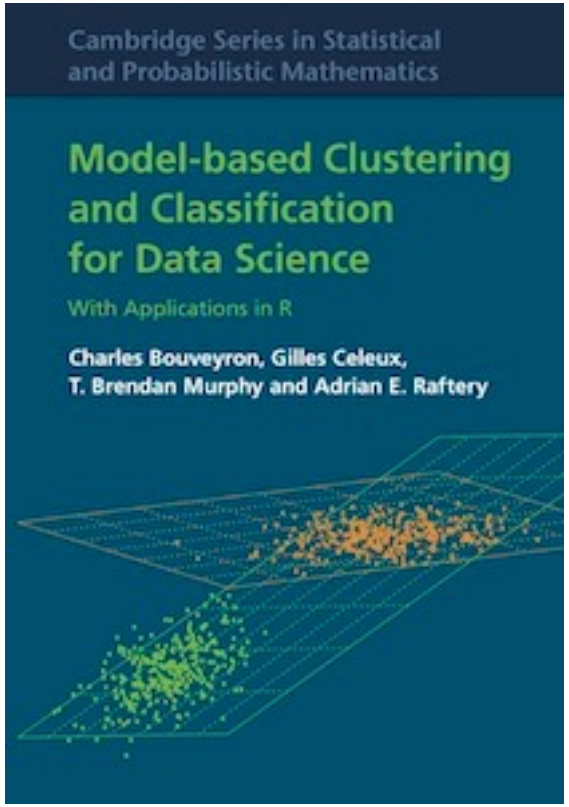


Statistical models?

- An example that we'll study a lot in this course in **clustering.**

- Charles Bouveyron co-wrote a book on the subject, and we'll refer to it often!

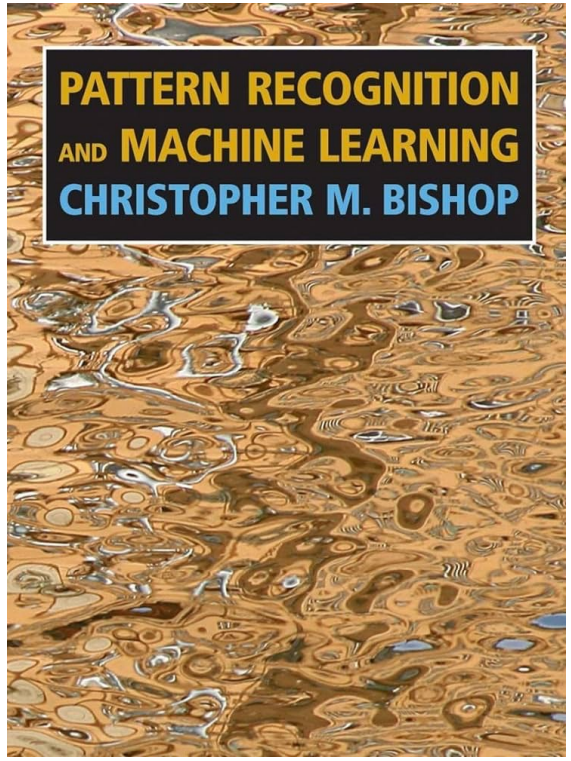
- It's freely available on Charles's website



Statistical models?

- Speaking of books, a classic book that covers most of machine learning and is very model-based at heart is the one by Chris Bishop. We'll refer to it a few times

- It's freely available on Bishop's webpage.



Why statistical models?

- Another example: we observe some data $x_1, \ldots, x_n \in \mathcal{X}$, for instance the CelebA dataset

Why statistical models?

- We observe some data $x_1, \ldots, x_n \in \mathcal{X}$, for instance the CelebA dataset

- Goal:

➢train an algorithm that will be able to generate new and plausible images!

Why on earth would we want to generate new and plausible images?

- Because it's cool!

Why on earth would we want to generate new and plausible images?

- Because it's cool!

Why on earth would we want to generate new and plausible images?

- Because it's cool!

Why on earth would we want to generate new and plausible images?

➢Because it can impact society unexpectedly quickly!

TIME ← THE BEST INVENTIONS OF 2022

Artificial Imagination

OpenAI DALL-E 2

Why on earth would we want to generate new and plausible images?

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

➢Because it can impact society unexpectedly quickly!

TIME ← THE BEST INVENTIONS OF 2022

Artificial Imagination

OpenAI DALL-E 2

Why on earth would we want to generate new and plausible images?

OPINION
GUEST ESSAY

Noam Chomsky: The False Promise of ChatGPT

➢Because it can impact society unexpectedly quickly!

TIME ← THE BEST INVENTIONS OF 2022

Artificial Imagination

OpenAI DALL-E 2

The Economist

Science & technology | Generative AI

Large, creative AI models will transform lives and labour markets

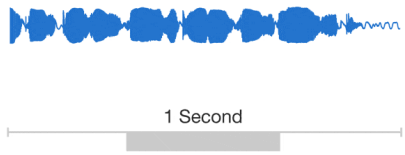
It's not just about pretty images and texts!

WaveNet: A Generative Model for Raw Audio

Aäron van den Oord **Sander Dieleman** **Heiga Zen**[†]

Karen Simonyan **Oriol Vinyals** **Alex Graves**

Nal Kalchbrenner **Andrew Senior** **Koray Kavukcuoglu**

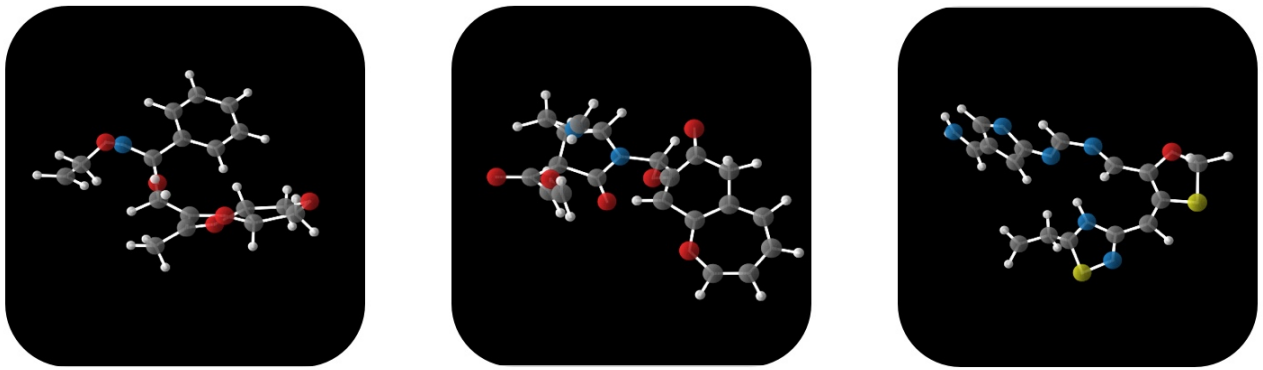


Character Controllers Using Motion VAEs

HUNG YU LING, University of British Columbia, Canada
FABIO ZINNO, Electronic Arts Vancouver, Canada
GEORGE CHENG, Electronic Arts Vancouver, Canada
MICHIEL VAN DE PANNE, University of British Columbia, Canada

Equivariant Diffusion for Molecule Generation in 3D

Emiel Hoogeboom [*1] **Victor Garcia Satorras** [*1] **Clément Vignac** [*2] **Max Welling** [1]

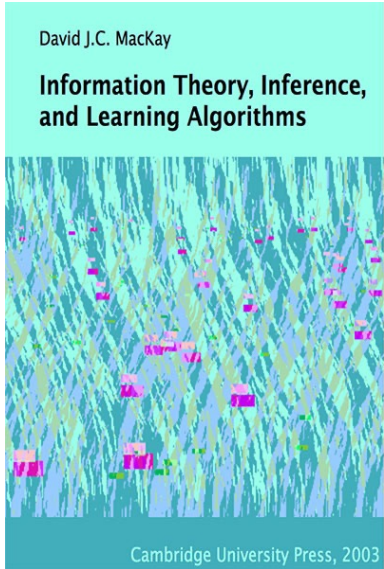


Towards formalising the problem

- Our goal is to generate new data points that are plausible.
- The kind of mathematical object that does just that is a probability distribution.

Towards formalising the problem

- Our goal is to generate new data points that are plausible.
- The kind of mathematical object that does just that is a probability distribution.
- Our goal will thus be to **learn a probability distribution that could have generated the data**.
- « a generative model describes a process that is assumed to give rise to some data » - DJC MacKay



Towards formalising the problem

- Our goal is to generate new data points that are plausible.
- The kind of mathematical object that does just that is a probability distribution.
- Our goal will thus be to **learn a probability distribution that could have generated the data**.
- Models are « small worls » - Jimmie Savage

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- Bernoulli distribution with parameter $\pi \in [0,1]$:
  - $\mathcal{X}=\{0,1\}$, so the subsets of $\mathcal{X}$ are $\emptyset, \{0\}, \{1\}, \{0,1\}$
  - $\mathbb{P}(\emptyset) = 0$
  - $\mathbb{P}(\{1\}) = \pi$
  - $\mathbb{P}(\{0\}) = 1 - \pi$
  - $\mathbb{P}(\{0,1\}) = 1$

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- Gaussian distribution with zero mean and variance 1/2:
  - $\mathcal{X} = \mathbb{R}$, so there are many styles of subsets of $\mathcal{X}$
    - ➢ Let's look at some simple ones: let $a \in \mathbb{R}$
  - $\mathbb{P}(\{a\}) = \mathbb{P}(\{b\}) = 0$
  - $\mathbb{P}([-a, a]) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt$

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- Dirac measure at a certain point $x_0 \in \mathcal{X}$
  - $\delta_{x_0}(\{x_0\}) = 1$
  - $\delta_{x_0}(\{x\}) = 0$ if $x \neq x_0$
  - For a subset $\mathcal{S} \subset \mathcal{X}$, $\delta_{x_0}(\mathcal{S}) = \begin{cases} 1 \text{ if } x_0 \in \mathcal{S} \\ 0 \text{ if } x_0 \notin \mathcal{S} \end{cases}$

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- Empirical distribution of training dataset of distinct points $x_1, \ldots, x_n \in \mathcal{X}$
  - $\mathbb{P}_n(\{x_i\}) = \frac{1}{n}$ for all i
  - For a subset $\mathcal{S} \subset \mathcal{X}$,

$$\mathbb{P}_n(\mathcal{S}) = \frac{|i \in \{1, \cdots, n\} \text{ such that } x_i \in \mathcal{S}|}{n}$$

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- It can't be just any function over subsets! **Probability measures have to verify a few intuitive properties** called Kolgogorov's axioms:
  - The function $\mathbb{P}$ is nonnegative
  - $\mathbb{P}(\mathcal{X}) = 1$
  - For any countable collection of disjoint subsets $\mathcal{S}_i \subset \mathcal{X}$,

$$\mathbb{P}(\bigcup_i \mathcal{S}_i) = \sum_i \mathbb{P}(\mathcal{S}_i)$$

What's a probability distribution?

- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- We can do some operations with distributions. For example a convex combination $\alpha\mathbb{P}_1 + (1-\alpha)\mathbb{P}_2$ will be a new distribution called the **mixture of $\mathbb{P}_1$ and $\mathbb{P}_2$** with mixing coefficient $\alpha \in (0,1)$.
  - **Q:** We saw 4 examples of distributions right before. Is one of them a mixture of another one?

What's a probability distribution?

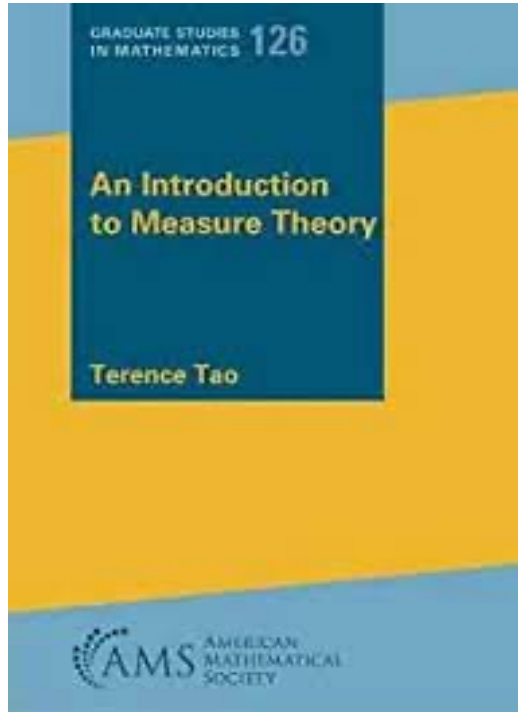
- Our data lives in a space $\mathcal{X}$
- A **probability distribution** (aka **measure**) is a function $\mathbb{P}$ that takes subsets of $\mathcal{X}$ and output how likely they are

- We can do some operations with distributions. For example a convex combination $\alpha\mathbb{P}_1 + (1-\alpha)\mathbb{P}_2$ will be a new distribution called the **mixture of $\mathbb{P}_1$ and $\mathbb{P}_2$** with mixing coefficient α.
  - **A:** The empirical distribution is a mixture of Diracs!

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

A lot was swept under the rug

- The past few slides were a very quick recap of measure-theoretic probability theory.
- I have clearly swept many (many) things under the rug, like issues related to measurability, the definition of a probability space…
- These things matter in ML, and also for DGMs, but they are not crucial either.



The issue with probability distributions

- Probability distributions are the right object to model random phenomena, but they are **not very convenient, because defining functions whose inputs are subsets is not easy** !

The issue with probability distributions

- Probability distributions are the right object to model random phenomena, but they are **not very convenient, because defining functions whose inputs are subsets is not easy** !

- Because of this, we often use other tools that allow to recover the probability distribution, but are more convenient:
  - The **cumulative distribution function (CDF)**
  - The **probability density function (PDF)**

The issue with probability distributions

- Probability distributions are the right object to model random phenomena, but they are **not very convenient, because defining functions whose inputs are subsets is not easy** !

- Because of this, we often use other tools that allow to recover the probability distribution, but are more convenient:
  - The **cumulative distribution function (CDF)**
  - The **probability density function (PDF)**

- Both CDF and PDF will be functions of elements of $\mathcal{X}$, instead of subsets

The cumulative distribution function

- Mostly easy to define when $\mathcal{X} = \mathbb{R}$:

$$\Phi(x) = \mathbb{P}(]-\infty, x]) = \mathbb{P}(X \leq x)$$

where X is a random variable with distribution $\mathbb{P}$.

The cumulative distribution function

- Mostly easy to define when $\mathcal{X} = \mathbb{R}$:

$$\Phi(x) = \mathbb{P}(]-\infty, x]) = \mathbb{P}(X \leq x)$$

where X is a random variable with distribution $\mathbb{P}$.

- $\forall x \in \mathcal{X}, \Phi(x) \geq 0$
- $\lim_{x \to -\infty} \Phi(x) = \mathbb{P}(\emptyset) = 0$
- $\lim_{x \to \infty} \Phi(x) = \mathbb{P}(\mathbb{R}) = 1$

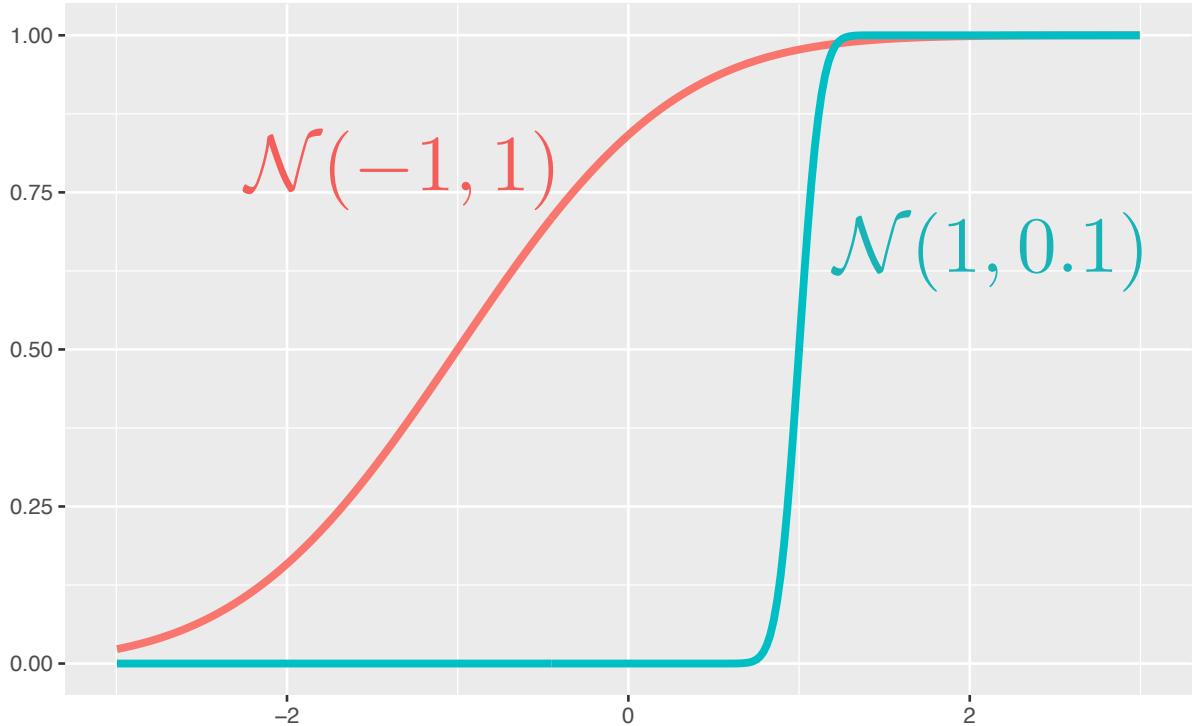
The cumulative distribution function

- Mostly easy to define when $\mathcal{X} = \mathbb{R}$:

$$\Phi(x) = \mathbb{P}(]-\infty, x]) = \mathbb{P}(X \leq x)$$

where X is a random variable with distribution $\mathbb{P}$.

- $\forall x \in \mathcal{X}, \Phi(x) \geq 0$
- $\lim_{x \to -\infty} \Phi(x) = \mathbb{P}(\emptyset) = 0$
- $\lim_{x \to \infty} \Phi(x) = \mathbb{P}(\mathbb{R}) = 1$



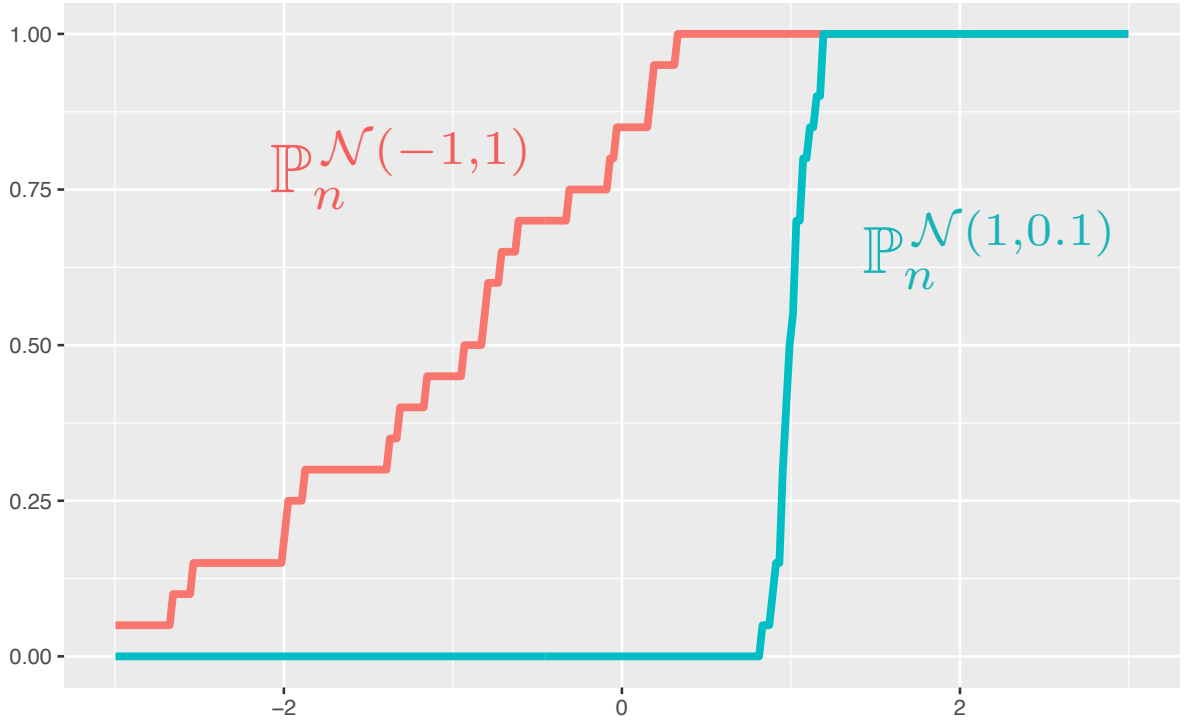
The cumulative distribution function

- Mostly easy to define when $\mathcal{X} = \mathbb{R}$:

$$\Phi(x) = \mathbb{P}(]-\infty, x]) = \mathbb{P}(X \leq x)$$

where X is a random variable with distribution $\mathbb{P}$.

$n = 20$



- $\forall x \in \mathcal{X}, \Phi(x) \geq 0$
- $\lim_{x \to -\infty} \Phi(x) = \mathbb{P}(\emptyset) = 0$
- $\lim_{x \to \infty} \Phi(x) = \mathbb{P}(\mathbb{R}) = 1$

The cumulative distribution function

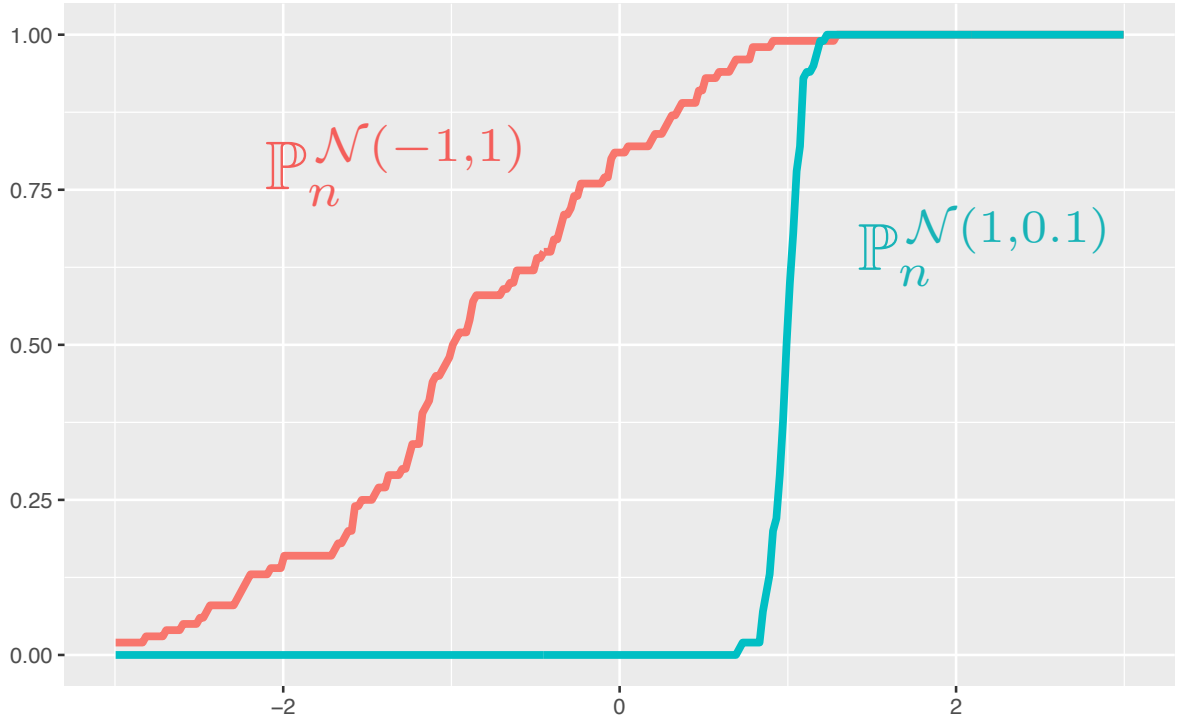
- Mostly easy to define when $\mathcal{X} = \mathbb{R}$:

$$\Phi(x) = \mathbb{P}(]-\infty, x]) = \mathbb{P}(X \leq x)$$

where X is a random variable with distribution $\mathbb{P}$.

- $\forall x \in \mathcal{X}, \Phi(x) \geq 0$
- $\lim_{x \to -\infty} \Phi(x) = \mathbb{P}(\emptyset) = 0$
- $\lim_{x \to \infty} \Phi(x) = \mathbb{P}(\mathbb{R}) = 1$

$n = 100$



The cumula ive ist ibution function

- Mostly easy to

$= \mathbb{P}(X \leq x)$

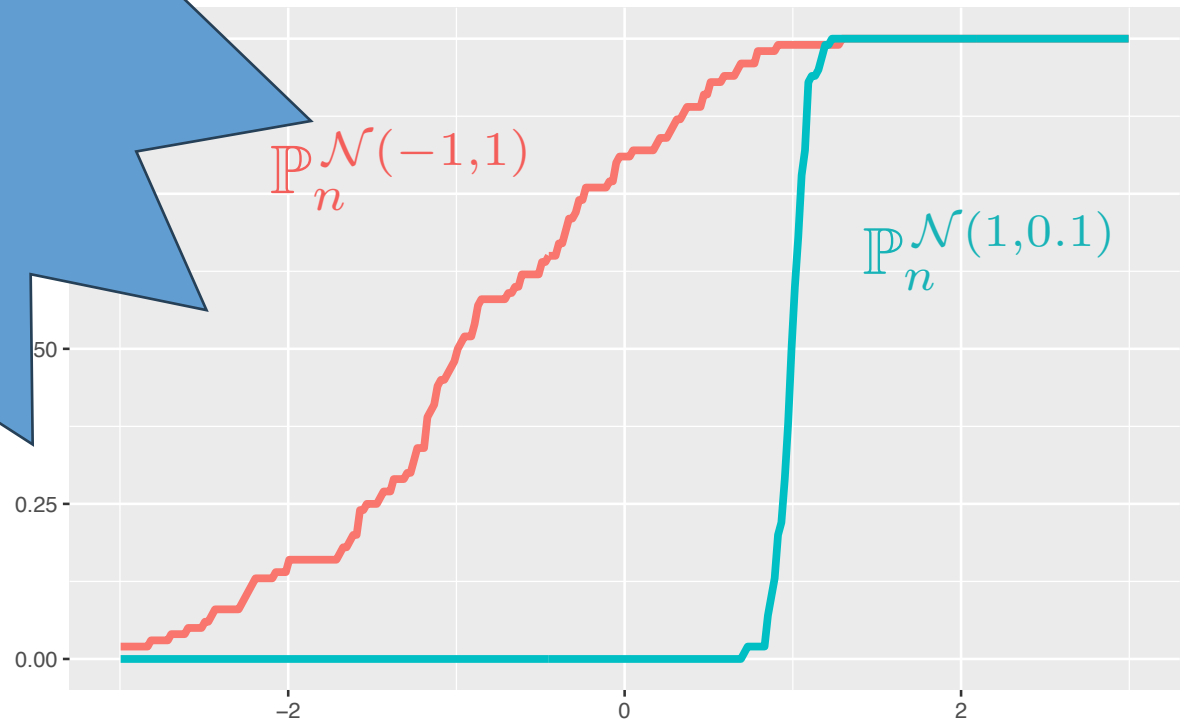
where X is a $\mathbb{P}$.

- $\forall x \in \mathcal{X}, \Phi($
- $\lim_{x \to -\infty} \Phi(x) =$
- $\lim_{x \to \infty} \Phi(x) = \mathbb{P}(\mathbb{R}) = 1$

This illustrates the fact that $\mathbb{P}_n \to \mathbb{P}$

Q: do you know this result under a different name?

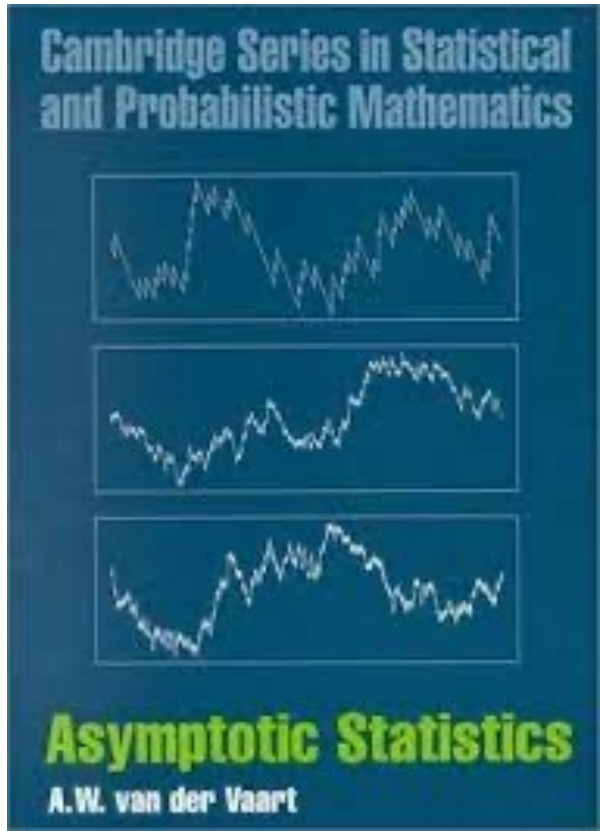
$n = 100$



The law of large numbers

- The fact that the empirical distribution converges to the theoretical one is (a variant of) the law of large numbers. It is sometimes called the **Glivenko-Cantelli theorem**.

- The law of large numbers comes in many flavours, the only one that is really important in ML is the **strong law of large numbers**, that we'll see in a few slides.



The density function

- Often (but not always, as we'll see in the deep learning course 😭), there exists a function $p: \mathcal{X} \to \mathbb{R}^+$ that allows to compute probabilities using integrals: for any subset $\mathcal{S} \subset \mathcal{X}$,

$$\mathbb{P}(\mathcal{S}) = \int_{\mathcal{S}} p(x)dx$$

- This function p is called the density of $\mathbb{P}$

- This integral is with respect to a base measure dx on $\mathcal{X}$, so it typically means that it will just be a **finite sum when $\mathcal{X}$ is discrete**, and a **Lebesgue integral when $\mathcal{X} = \mathbb{R}^D$**

The density function of a Bernoulli

- $p: \mathcal{X} \to \mathbb{R}^+$ depends on the parameter $\pi \in [0,1]$, so **we will denote it $p_\pi(x)$ or $p(x \mid \pi)$** to make the dependence clear. We will often do it when densities (or distributions in general) depend on a parameter.

- The density is defined as $p_\pi(x) = \begin{cases} \pi \text{ if } x = 1 \\ 1 - \pi \text{ if } x = 0 \end{cases}$

The density function of a Bernoulli

- $p: \mathcal{X} \to \mathbb{R}^+$ depends on the parameter $\pi \in [0,1]$, so **we will denote it $p_\pi(x)$ or $p(x\,|\pi)$** to make the dependence clear. We will often do it when densities (or distributions in general) depend on a parameter.

- The density is defined as $p_\pi(x) = \begin{cases} \pi \text{ if } x = 1 \\ 1 - \pi \text{ if } x = 0 \end{cases}$
- There another (sometimes more convenient) way to write it:

$$p_\pi(x) = \pi^x (1-\pi)^{1-x}$$

The density function of a Bernoulli

- $p: \mathcal{X} \to \mathbb{R}^+$ depends on the parameter $\pi \in [0,1]$, so **we will denote it $p_\pi(x)$ or $p(x\,|\pi)$** to make the dependence clear. We will often do it when densities (or distributions in general) depend on a parameter.

- The density is defined as $p_\pi(x) = \begin{cases} \pi \text{ if } x = 1 \\ 1 - \pi \text{ if } x = 0 \end{cases}$
- There another (sometimes more convenient) way to write it:

$$p_\pi(x) = \pi^x (1-\pi)^{1-x}$$

- Minus the log of this is called the **cross-entropy loss** between x and π

$$\mathrm{CE}(x, \pi) = -\log p_\pi(x) = -x \log \pi - (1-x)\log(1-\pi)$$

The density function of discrete distributions

- For the Bernoulli, we had, for any x

$$p(x) = \mathbb{P}(\{x\})$$

- This is true more generally for any discrete distribution!
  - This is because this is a density with respect to the counting measure

The density function of discrete distributions

- For the Bernoulli, we had, for any x

$$p(x) = \mathbb{P}(\{x\})$$

- **This is true more generally for any discrete distribution!**
  - **This is because this is a density with respect to the counting measure**

- For discrete distribution, **densities are "true probabilities"** (ie, between 0 and 1 and sum to 1)

- This is unfortunately not the case for "continuous" distribution, for which, very often $\mathbb{P}(\{x\}) = 0$

The density function of a Gaussian

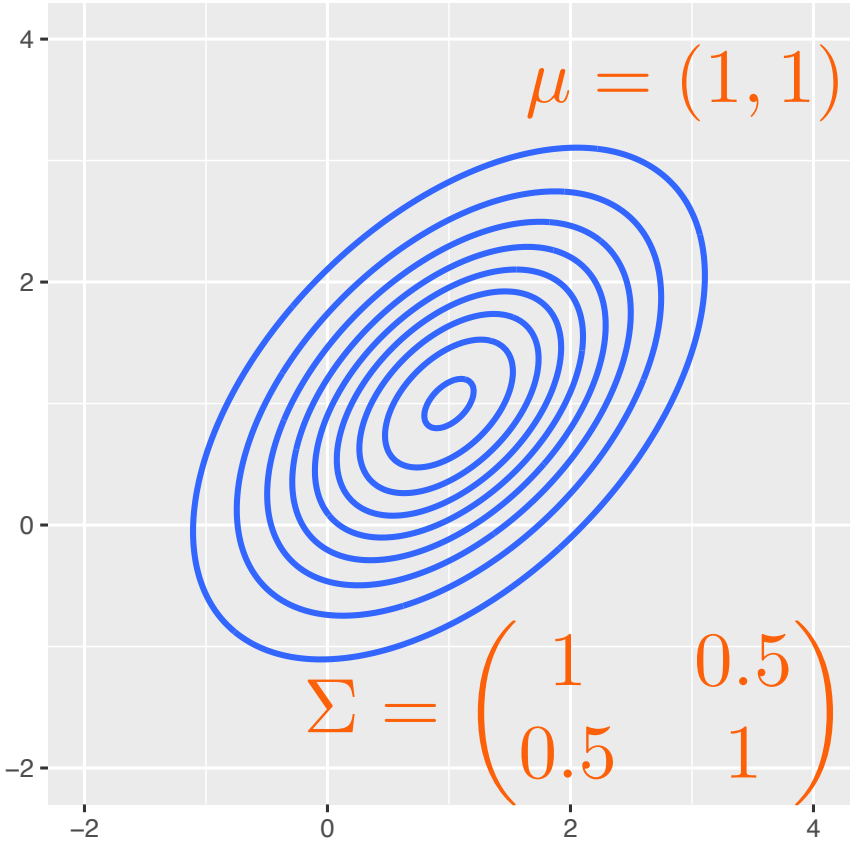
- Here, $\mathcal{X} = \mathbb{R}^D$ depends now on two parameters, $\mu \in \mathbb{R}^D$ and $\Sigma \in S_D$, where S_D is the set of positive definite matrices. Because it's a bit clumsy to write $p_{\mu,\Sigma}$, we'll write p_θ where $\theta = (\mu, \Sigma)$.

$$p_\theta(x) = \frac{1}{\sqrt{2\pi \det \Sigma}} e^{(x-\mu)^T \Sigma^{-1} (x-\mu)/2}$$

The density function of a Gaussian

- Here, $\mathcal{X} = \mathbb{R}^D$ depends now on two parameters, $\mu \in \mathbb{R}^D$ and $\Sigma \in S_D$, where S_D is the set of positive definite matrices. Because it's a bit clumsy to write $p_{\mu,\Sigma}$, we'll write p_θ where $\theta = (\mu, \Sigma)$.

$$p_\theta(x) = \frac{1}{\sqrt{2\pi \det \Sigma}} e^{(x-\mu)^T \Sigma^{-1} (x-\mu)/2}$$



- The axes of the ellipsis are aligned with the eigenvectors of Σ

The density function of a Gaussian

- Minus the log of the pdf of the Bernoulli was the **cross-entropy loss**, is there something similar with the Gaussian?

- Yes! If we assume that $\Sigma = \sigma^2 I_D$, we find

$$-\log p_\theta(x) = \frac{1}{2\sigma^2} \underbrace{\sum_{j=1}^{D} (x_j - \mu_j)^2}_{\text{squared error between } \mu \text{ and } x} + \frac{1}{2} \log(\sqrt{2\pi \det \Sigma})$$

The density function of a Gaussian

- Minus the log of the pdf of the Bernoulli was the **cross-entropy loss**, is there something similar with the Gaussian?

- Yes! If we assume that $\Sigma = \sigma^2 I_D$, we find

$$-\log p_\theta(x) = \frac{1}{2\sigma^2} \underbrace{\sum_{j=1}^{D} (x_j - \mu)^2}_{\text{squared error between } \mu \text{ and } x} + \frac{1}{2} \log(\sqrt{2\pi \det \Sigma})$$

- There's a deep link **Gaussian-Mean squared error (MSE)**, we'll come back to this several times (for VAEs & diffusion models in particular!).

The density function of a mixture

- $\mathbb{P}_1$ admits a density p_1
- $\mathbb{P}_2$ admits a density p_2

  ➢ both w.r.t. the same measure (e.g. both are continuous or discrete)

- $\alpha \in [0,1]$

➢ **Then, $\alpha\mathbb{P}_1 + (1-\alpha)\mathbb{P}_2$ has density $\alpha p_1 + (1-\alpha)p_2$**

The density function of the data?

- A fondamental distribution in ML is the empirical distribution of our dataset $x_1, \dots, x_n \in \mathbb{R}^D$, that we defined earlier as a **mixture of Diracs**

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

- Note that is we assume that the $x_1, \dots, x_n$ are random variables, $\mathbb{P}_n$ will be a random variable too!
- **Q:** What's the density of $\mathbb{P}_n$?

The density function of the data?

- A fondamental distribution in ML is the empirical distribution of our dataset $x_1, \dots, x_n \in \mathbb{R}^D$, that we defined earlier as a **mixture of Diracs**

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

- Note that is we assume that the $x_1, \dots, x_n$ are random variables, $\mathbb{P}_n$ will be a random variable too!
- **Q:** What's the density of $\mathbb{P}_n$?
  - **Trick question!** When the sample space is $\mathbb{R}^D$, **Dirac distributions have no densities** wrt the Lebesgue measure.
  - That means that $p_n(x)$ **will have no meaning** 😰

Independence and densities

- Densities are super convenient for conveying many probabilistic concepts
- For instance, some random variables $x_1, \ldots, x_n \in \mathcal{X}$ with joint density $p(x_1, \ldots, x_n)$ are independent when the density can be written

$$p(x_1, \cdots, x_n) = p_1(x_1) \cdots p_n(x_n)$$

where $p_i(x_i) = \int p(x_1, \cdots, x_n) dx_{-i}$ is called a **marginal density**.

Independence and densities

- Densities are super convenient for conveying many probabilistic concepts
- For instance, some random variables $x_1, \ldots, x_n \in \mathcal{X}$ with joint density $p(x_1, \ldots, x_n)$ are independent when the density can be written

$$p(x_1, \cdots, x_n) = p_1(x_1) \cdots p_n(x_n)$$

where $p_i(x_i) = \int p(x_1, \cdots, x_n) dx_{-i}$ is called a **marginal density**.

- When they are additionally **identically distributed**, the formula becomes

$$p(x_1, \cdots, x_n) = p_1(x_1) \cdots p_1(x_n)$$

The density and expectations

- Densities are super convenient for conveying many probabilistic concepts
- For instance, the expected value of a random variable with density p is

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

The density and expectations

- Densities are super convenient for conveying many probabilistic concepts
- For instance, the expected value of a random variable with density p is

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

- This can be extended to any function of our random variable

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

➢Sometimes known as the **law of the unconscious statistician** (LOTUS)

Expectations and the strong LLN

- The **strong law of large numbers** is one of the few asymptotic stats results that are often useful in ML

- It goes as follows: if $x_1, \ldots, x_n$ are independent and identically distributed with common and finite mean μ, then

$$\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow[\text{a.s}]{n\to\infty} \mu$$

- This is the basis of the idea of **Monte Carlo**: we can use the average of the data to approximate the true mean.

The learning problem

- Our data $x_1, \dots, x_n \in \mathcal{X}$ are assumed to be iid samples from an **unknown distribution $\mathbb{P}_{\text{data}}$, with density p_{data}**

- Our goal is to approximate p_{data} using a statistical model, ie a family of probability densities $(p_\theta)_{\theta \in \Theta}$

- In other words, we need to **find a good $\hat{\theta} \in \Theta$ such that $p_{\hat{\theta}} \approx p_{\text{data}}$**

- We will sketch two of these approaches: **Bayesian inference** and **maximum likelihood estimation**

Bayesian inference

- Conceptually, it is quite simple!
  - « Bayesian statistics is fundamentally boring », A.P. Dawid

- Start with a **density $p_\Theta(\theta)$ over Θ, called a prior**, it is supposed to give higher probability to the models you find best before seeing the data
  - ➢ smooth/sparse/"simple" models are more probable a priori

Bayesian inference

- Conceptually, it is quite simple!
  - « Bayesian statistics is fundamentally boring », A.P. Dawid

- Start with a **density $p_\Theta(\theta)$ over Θ, called a prior**, it is supposed to give higher probability to the models you find best before seeing the data
  - ➢ smooth/sparse/"simple" models are more probable a priori

- Compute the **posterior density** of the model given the data using Bayes's rule

$$p(\theta|x_1,\ldots,x_n) = \frac{p(x_1|\theta)\ldots p(x_n|\theta)p_\Theta(\theta)}{p(x_1,\ldots,x_n)}$$

In Bayesian mode, it's a bit clearer to write $p(x|\theta)$ instead of $p_\theta(x)$

Bayesian inference

- Conceptually, it is quite simple!
  - « Bayesian statistics is fundamentally boring », A.P. Dawid

- Start with a **density $p_\Theta(\theta)$ over Θ, called a prior**, it is supposed to give higher probability to the models you find best before seeing the data
  - ➢ smooth/sparse/"simple" models are more probable a priori

- Compute the **posterior density** of the model given the data using Bayes's rule

$$p(\theta|x_1,\ldots,x_n) = \frac{p(x_1|\theta)\ldots p(x_n|\theta)p_\Theta(\theta)}{p(x_1,\ldots,x_n)}$$

- Then, **sample from the posterior** to get good $\hat{\theta}$s!

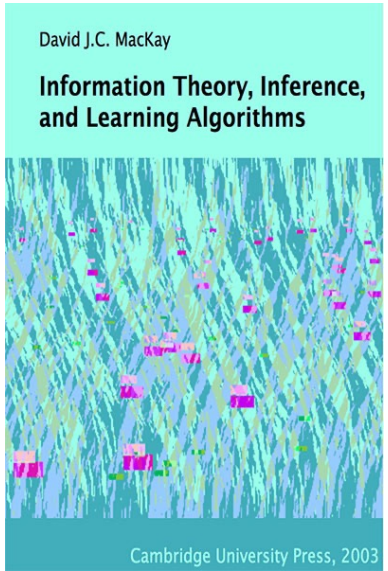
In Bayesian mode, it's a bit clearer to write $p(x|\theta)$ instead of $p_\theta(x)$

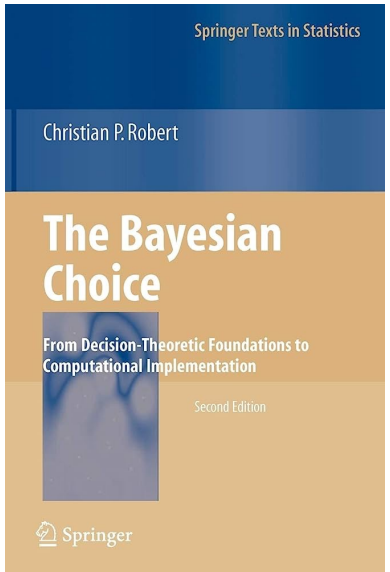
Bayesian inference?

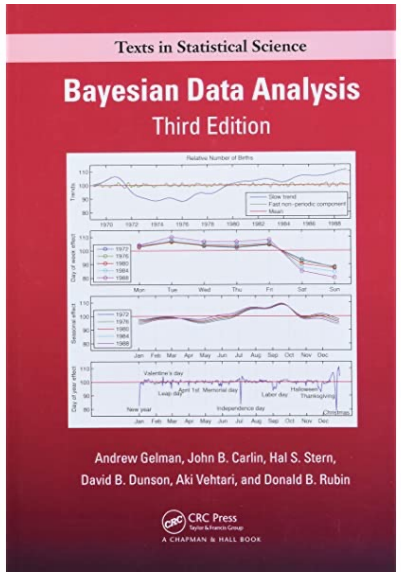
- We won't use much Bayesian inference in this course, except later for **model selection** (i.e. finding the number of clusters).

- Fortunately, there is a nice Bayesian inference course in the master's.

- Here are a few Bayesian books that I like.







Towards maximum likelihood

- We will present the workhorse of DGM estimation, **maximum-likelihood inference**, that will be used (in one form or another) for VAEs, flows, ARMs, probabilistic circuits, diffusion models…

- Let's keep in mind **find a good $\hat{\theta} \in \Theta$ such that $p_{\hat{\theta}} \approx p_{\text{data}}$**

- Idea: let's find a **distance d over models, and find the model closest to the truth:**

$$\hat{\theta} \in \text{argmin}_\theta d(p_{\text{data}}, p_\theta)$$

- **Q:** do you know distances over models? (ie probability distributions)

The Kullback-Leibler divergence

- Many choices of distances d over models will lead to interesting estimators, but as we will see, the simplest choice will the the **Kullback-Leibler divergence**, defined as

$$\mathrm{KL}(p_1, p_2) = \int_{\mathcal{X}} \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x)dx = \mathbb{E}_{X\sim p_1}\left[\log\left(\frac{p_1(X)}{p_2(X)}\right)\right].$$

  - average difference of the log-densities
  - asymmetrical, but with some nice "distance-like" properties:

1. $\mathrm{KL}(p_1, p_2) \geq 0,$
2. $\mathrm{KL}(p_1, p_2) = 0 \iff p_1 = p_2.$

Finding the best model according to the KL

- Now that we've chosen a distance, let's look at the objective again:

$$\hat{\theta} \in \operatorname{argmin}_{\theta} \mathrm{KL}(p_{\mathrm{data}}, p_{\theta})$$

Finding the best model according to the KL

- Now that we've chosen a distance, let's look at the objective again:

$$\hat{\theta} \in \operatorname{argmin}_\theta \mathrm{KL}(p_{\mathrm{data}}, p_\theta)$$

Where $\mathrm{KL}\left(p_{\mathrm{data}}, p_\theta\right) = \mathbb{E}\left[\log p_{\mathrm{data}}(X)\right] - \mathbb{E}\left[\log p_\theta(X)\right]$ with $X \sim p_{\mathrm{data}}$

Finding the best model according to the KL

- Now that we've chosen a distance, let's look at the objective again:

$$\hat{\theta} \in \operatorname{argmin}_\theta \mathrm{KL}(p_{\mathrm{data}}, p_\theta)$$

Where $\mathrm{KL}\left(p_{\mathrm{data}}, p_\theta\right) = \underbrace{\mathbb{E}\left[\log p_{\mathrm{data}}(X)\right]}_{\text{does not depend on } \theta} - \mathbb{E}\left[\log p_\theta(X)\right]$ with $X \sim p_{\mathrm{data}}$

Finding the best model according to the KL

- Now that we've chosen a distance, let's look at the objective again:

$$\hat{\theta} \in \operatorname{argmin}_\theta \mathrm{KL}(p_{\mathrm{data}}, p_\theta)$$

Where $\mathrm{KL}\left(p_{\mathrm{data}}, p_\theta\right) = \underbrace{\mathbb{E}\left[\log p_{\mathrm{data}}(X)\right]}_{\text{does not depend on } \theta} - \mathbb{E}\left[\log p_\theta(X)\right]$ with $X \sim p_{\mathrm{data}}$

- Therefore $\hat{\theta} \in \operatorname{argmax}_\theta \mathbb{E}[\log p_\theta(X)]$

Finding the best model according to the KL

- Now that we've chosen a distance, let's look at the objective again:

$$\hat{\theta} \in \text{argmin}_\theta \text{KL}(p_{\text{data}}, p_\theta)$$

Where $\text{KL}(p_{\text{data}}, p_\theta) = \underbrace{\mathbb{E}[\log p_{\text{data}}(X)]}_{\text{does not depend on } \theta} - \mathbb{E}[\log p_\theta(X)]$ with $X \sim p_{\text{data}}$

- Therefore $\hat{\theta} \in \text{argmax}_\theta \mathbb{E}[\log p_\theta(X)]$
- Unfortunately, **we can't compute this expectation**, because we don't know p_{data}. **But we do have samples from p_{data}.**

Using the data to approximate the KL

- We want to find $\hat{\theta} \in \operatorname{argmax}_\theta \mathbb{E}[\log p_\theta(X)]$

Using the data to approximate the KL

- We want to find $\hat{\theta} \in \operatorname{argmax}_\theta \mathbb{E}[\log p_\theta(X)]$

- Idea! Use the law of large numbers and the fact that we have access to the data to approximate the expectation.

$$\mathbb{E}[\log p_\theta(x_i)] \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i) = \ell(\theta)}_{\text{log-likelihood function}}$$

Finally defining the MLE

- The maximum-likelihood estimate of our model is defined as

$$\hat{\theta}_{\mathrm{MLE}} \in \operatorname{argmax} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i)$$

- The hope is that it won't be too far from the optimal KL model $\hat{\theta}$.
  - **Q:** Why can we hope? What may go wrong?