Unsupervised Learning with discrete latent variable models

Nicolas Jouvin
nicolas.jouvin@inrae.fr
https://nicolasjouvin.github.io/

M2 Data-Science 2023-2024

Organization

Thursdays 8h30 - 11h45, this room.

6×3h classes

1h30 class + 1h30 practical session (except today)

Important: you need one computer/person for practical sessions

Evaluation

1. CC: assiduity, practical session
2. Final exam on Friday 12th January, 2024
3. $\max(\mathrm{Exam}, \mathrm{mean}(\mathrm{Exam}, \mathrm{CC}))$

Bibliography & relevant sources

- Kevin P. Murphy (2022). *Probabilistic Machine Learning: An introduction*. MIT Press
- Trevor Hastie et al. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA
- Christopher M. Bishop (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer

Some relevant lecture/slides on the topic for a different point-of-view (⚠notations)

- S. Robin lectures
-

Introduction

Types of statistical learning

Supervised

Data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with y_i an output (*response*) and x_i some features (*covariates*).
The goal is to learn a good predictor $\hat{f}$ such that $y_i \approx \hat{f}(x_i)$ that generalizes well on new data.

Unsupervised (this course)

The data $\mathcal{D} = \{x_i\}_{i=1}^n$ The goal is to learn "interesting" and hidden structure in the data to

- partition the data, aka clustering
- visualize/compress the data, aka dimension reduction

Generative models: posit a statistical model on the distribution of (X_i)

Many flavors in modern ML

semi-supervised, self-supervised, reinforcement learning, multi-task, etc.

What this course is about...

(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ we will assume the generative process of X involves an unobserved (latent) variable Z

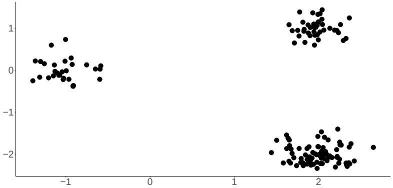
What this course is about...

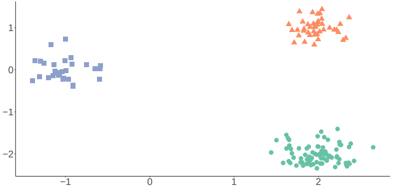
(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ we will assume the generative process of X involves an unobserved (latent) variable Z

Clustering

X is an unlabeled observation and Z its group membership





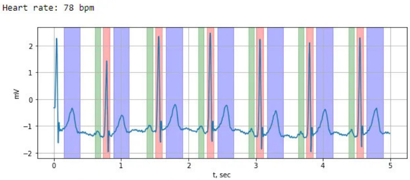
What this course is about...

(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ we will assume the generative process of X involves an unobserved (latent) variable Z

Time series segmentation

X is the temporal signal and Z the cardiac phase



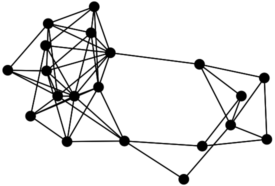
Example of ECG annotation, source: `https://medium.com/data-analysis-center/56f8b9abd83a`

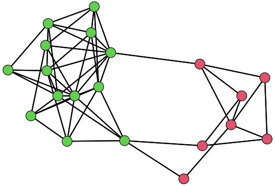
What this course is about...

(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ we will assume the generative process of X involves an unobserved (latent) variable Z

Node clustering in a network





X is the graph (connection between node) and Z the group of the node (community)

Course outline

1. **Fundamentals of Bayesian statistics**
2. **Clustering with mixture models**
3. **Inference in latent variable models: the EM algorithm**
4. **Hidden Markov Models**
5. **Stochastic Block Model: an introduction to variational inference**

Fundamentals of Bayesian statistics

Bayes formula

Frequentist inference

Assumption: the observation $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is a realization of a random vector $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ with distribution $p_{\theta^\star}$.

Posit: a statistical model $\{p_\theta,\ \theta \in \Theta\}$, *i.e.* a family of parametric distribution on $\mathcal{X}^n$

Goal: Provide an estimate $\hat{\theta}$ of $\theta^\star$. [1]

Maximum-likelihood estimation

Find the model, hence θ, that maximizes the probability of having seen the data

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \log p_\theta(x_1, \ldots, x_n) \tag{MLE}$$

[1] and eventually derive theoretical guarantees such as convergence and confidence intervals on $\hat{\theta}_n(X_1, \ldots, X_n)$ (*e.g.* via central limit theorem)

The Bayesian paradigm

Maximum-likelihood and frequentist statistics produces *point estimates*

Paradigm shift: random parameters

Parameters θ are no longer treated as deterministic but as *random* quantities. The *prior* distribution, denoted as $\pi(\theta)$, encodes knowledge & uncertainty we have on the parameters **before** seeing new data.
$\rightsquigarrow$ the goal is to update this a priori knowledge when new data comes: this is the essence of Bayes formula.

A bit of history...

The terminology *Bayesian* has been coined that way thanks to the work of Reverend Thomas Bayes (1701-1761) and his posthumous *essay in view of solving the doctrine of chance*. Pierre-Simon Laplace independently proposed a version in 1774.
N.B. : this course will not settle the somewhat sterile debate "Bayesian VS Frequentist".

Bayes formula

Equipped with a prior $\pi(\theta)$, we posit an observational model on $X \mid \theta \iff$ the likelihood. Bayesian modelisation essentially adds one layer to frequentist models : the prior.

$$
\begin{aligned}
&1. \quad \theta \sim \pi, && \text{(prior)} \\
&2. \quad X \mid \theta \sim p(\cdot \mid \theta) = p_\theta && \text{(likelihood)}.
\end{aligned}
$$

The posterior

Given a realization x, we update our prior via a new distribution called the *posterior*:

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{Z}, \qquad \text{(Bayes formula)}$$

Here, $Z = \int_\Theta p(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta$ is a normalization constant, independent of θ. Thus, it is common to write[a]

$$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta)$$

[a] Although computing this normalization constant is generally a challenging task in Bayesian statistics.

Choosing a prior

Expert knowledge

The prior π may be used to represent any available expert knowledge on θ.

Conjugate priors

When the prior π and the posterior $\pi(\cdot \mid x)$ belong to the same family of distributions (e.g. Gaussian, Beta, etc.), then we say that the prior is *conjugate* to the observational model $p(x \mid \theta)$. Skip to an example

Conjugate priors are widely used as they greatly simplify computations.

Uninformative prior

When the prior equally charges Θ we say that the prior is *uninformative*, noted $\pi(\theta) \propto 1$. Obviously, $\pi \propto 1$ does not always define a proper p.d.f. (consider $\Theta = \mathbb{R}$). Still, as long as the posterior is well defined (*i.e.* the normalization constant Z exists and is finite) then we can still use the posterior $\pi(\theta \mid x)$ and the prior is *improper*.

Example of conjugacy: the Beta-Binomial model (1)

Experiment & question Given a sequence of independent coin flips $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, determine the probability of getting tail.

Observational model: the likelihood
Given a probability of tail θ, we model the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ as *i.i.d.* Bernoulli $X_i \sim Ber(\theta)$ so that

$$p(\boldsymbol{X} \mid \theta) = \prod_{i=1}^{n} Ber(x_i \mid \theta) = \theta^{\sum_i x_i}(1-\theta)^{\sum_i 1-x_i}.$$

Choice of a prior
We use Beta distribution with support $\Theta = [0,1]$

$$\pi(\theta) = Beta(a,b) \propto \mathbf{1}_{[0,1]}(\theta)\theta^{a-1}(1-\theta)^{b-1}.$$

a and b are called *hyper-parameters* and they control our level of a priori

- $a = b = 1$: uniform on $[0,1]$ (uninformative)
- $a = b > 1$: in favor of a balanced coin, the greater a, the stronger the prior
- $a > b$ (resp. $a < b$): in favor of tail (resp. head).

Example of conjugacy: the Beta-Binomial model (2)

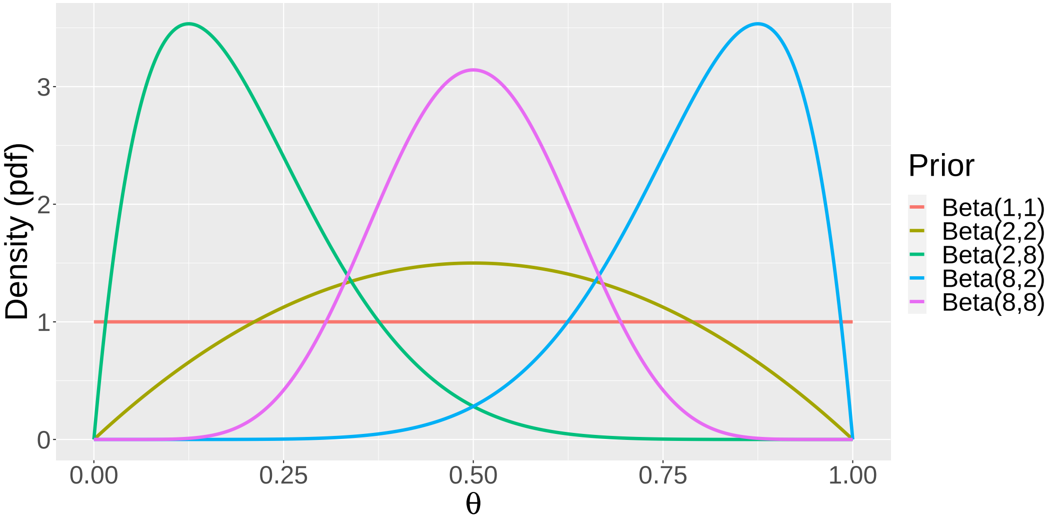


Figure: Graph of the p.d.f. $Beta(\cdot \mid a, b)$ for different values of a and b.

Example of conjugacy: the Beta-Binomial model (3)

We seek to derive the posterior, and we directly have

$$\begin{aligned}\pi(\theta \mid \boldsymbol{X}) &\propto p(\boldsymbol{X} \mid \theta)\pi(\theta),\\ &\propto \theta^{\sum_i x_i}(1-\theta)^{\sum_i 1-x_i}\theta^{a-1}(1-\theta)^{b-1}\mathbf{1}_{[0,1]}(\theta),\\ &\propto \theta^{a+\sum_i x_i}(1-\theta)^{b+\sum_i 1-x_i}\mathbf{1}_{[0,1]}(\theta).\end{aligned}$$

We recognize the p.d.f of a Beta distribution

$$\theta \mid \boldsymbol{X} \sim Beta\left(a+\sum_i X_i, b+n-\sum_i X_i\right)$$

Remarks :

1. a and b act as *pseudo-counts* for head and tails, smoothing the estimates when n is small.
2. This conjugacy between the Beta prior and the binomial model always hold : property of the model (prior + likelihood) and not our specific experiment.

Bayesian decision theory

Bayesian point estimates

Having derived the posterior: how do we provide point estimates $\hat{\theta}$?

Cost function

A *cost function* is a function $C : \Theta \times \Theta \in \mathbb{R}_+$ where $C(\eta, \theta)$ is the "cost of predicting η for a parameter θ. Some examples

- $C(\eta, \theta) = (\eta - \theta)^p$ (L^p-loss)
- $C(\eta, \theta) = \mathbf{1}_{\eta \neq \theta}$ (0-1 loss)

Bayesian estimator

Remember that θ is random. For a given model and observation x, the Bayesian estimator is the one that minimizes the average cost under the posterior distribution:

$$\hat{\theta} \in \arg\min_{\eta} \left\{ \mathbb{E}_{\theta \sim \pi(\cdot|x)} \left[C(\eta, \theta) \right] = \int_{\Theta} C(\eta, \theta) \pi(\theta \mid x) \, \mathrm{d}\theta \right\}. \qquad \text{(Bayes estimator)}$$

Posterior Mean, Median & Mode

Different cost functions leads to different Bayes estimator among which

1. posterior mean $\hat{\theta} = \mathbb{E}[\theta \mid x]$ corresponds to the L^2-loss
2. posterior median $\hat{\theta}$ such that $\pi(\theta \geq \hat{\theta} \mid x) = \pi(\theta \leq \hat{\theta} \mid x) = 0.5$ (L^1-loss)
3. **posterior mode (aka MAP)**: $\hat{\theta} \in \arg\max_\theta \pi(\theta \mid x)$ (0-1 loss)

Maximum a posteriori is one of the most popular

- reduces to an optimization problem
- log-prior can be interpreted in a frequentist setting as a regularizer for MLE

$$\log \pi(\theta \mid x) = cte + \underbrace{\log p_\theta(x)}_{\text{likelihood}} + \underbrace{\log \pi(\theta)}_{\text{regularizer}}$$

Credibility regions

The posterior may also be used for uncertainty quantification by computing regions $\mathcal{R} \subset \Theta$ s.t. $\pi(\theta \in \mathcal{R} \mid x) = \int_{\mathcal{R}} \pi(\theta \mid x)\, \mathrm{d}\theta = 1 - \alpha$

Latent variable models

Incomplete data models

Most often, the observations are involved in complicated (biological, ecological, physical) processes, with many unobserved variables and complex dependency structure.

- $\boldsymbol{X}$ observed random variables
- $\boldsymbol{Z}$ unobserved (latent/hidden) variables
- θ unknown parameters

An attempt at defining latent variables (creds. to S. Robin)

- Frequentist setting:

$$\text{latent variables} = \text{random but unobserved}, \quad \text{parameters} = \text{fixed}$$

- Bayesian setting:

$$\text{both latent variables and parameters} = \text{random}$$

but

$$\#\ \text{latent variable} \simeq \#\ \text{data}, \quad \#\ \text{parameters} \ll \#\ \text{data}$$

Different types of likelihoods

In this course, we place ourselves in the frequentist setting, using MLE inference. Although Bayesian extension of the proposed models are common.

Complete data likelihood

Joint likelihood of the whole random process $(\boldsymbol{X}, \boldsymbol{Z})$ with given parameters θ.

$$p_\theta(\boldsymbol{X}, \boldsymbol{Z}) = p_\theta(\boldsymbol{X} \mid \boldsymbol{Z}) p_\theta(\boldsymbol{Z}).$$

$\rightsquigarrow$ tractable in many models, but we do not observe $\boldsymbol{Z}$!

Observed data likelihood

Marginal likelihood of the observed random variables $\boldsymbol{X}$

$$p_\theta(\boldsymbol{X}) = \int_{\mathcal{Z}} p_\theta(\boldsymbol{X}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}^{a}$$

$\rightsquigarrow$ only involves the observed $\boldsymbol{X}$, but not always tractable.

[a] When $\mathcal{Z}$ is discrete, replace $\int$ by $\sum$

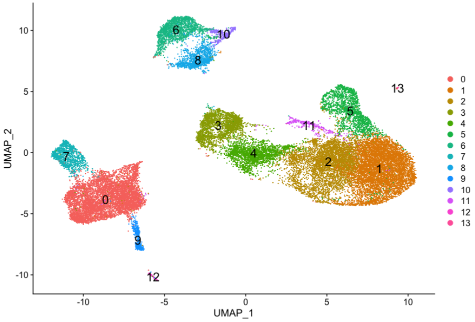
Clustering with mixture models

Motivation

Sometimes our data is organized in sub-population: groups of individuals we call *clusters*.

Example

In modern biology, discovering cell-types via their gene expression profile is an important task.



When the groups are unknown, we call the task of discovering them *clustering*[2]

[2] as opposed to classification in a supervised context

Mathematical context

We search for an optimal partition of $\boldsymbol{x} = \{x_1, \dots, x_n\}$ into K groups.

Definition: partition

A partition $\boldsymbol{C} = \{C_1, \dots, C_K\}$ of $\{1, \dots, n\}$ is a set of sets s.t.

$$\bigcup_k C_k = \{1, \dots, n\}, \qquad \forall k \neq l, \quad C_k \cap C_l = \emptyset.$$

Alternative encoding of the partition

For each individual $i = 1, \dots, n$, we define its *cluster membership* $z_i \in \{0, 1\}^K$

$$k = 1, \dots, K, \quad z_{ik} = \begin{cases} 1 & \text{if i belongs to cluster k,} \\ 0 & \text{otherwise} \end{cases}.$$

The set $\boldsymbol{Z} = \{z_1, \dots, z_n\}$ represents a partition of $\{1, \dots, n\}$. This particular encoding is sometimes referred to as one-hot encoding.

Clustering criteria

"Optimality" implies the definition of some criterion $L \iff$ assumptions on the nature of clusters. Methods can be roughly split in two

Similarity-based methods

Design L via geometric notions of similarity between x_i's, favoring *e.g.*

- elliptic clusters
- convex clusters
- connected clusters

Statistical methods

Consider the partition $\boldsymbol{Z}$ as a latent variable and posit a generative model $p_\theta(\boldsymbol{X}, \boldsymbol{Z})$
$\rightsquigarrow$ Clustering becomes an inference problem of finding $\hat{\boldsymbol{Z}}$.

There are connections between both !

K-means

The K-means problem

K-means seeks clusters well concentrated around their centroids $\mu_k := \frac{1}{|C_k|}\sum_{i \in C_k} x_i$ by minimizing

$$\underset{\boldsymbol{C}}{\arg\min}\left\{L(\boldsymbol{C},\boldsymbol{X}) = \sum_{k=1}^{K}\sum_{i\in C_k}\|x_i - \mu_k\|_2^2\right\} \qquad \text{(K-means problem)}$$

- Good news: discrete problem $\rightsquigarrow$ there exists an optimum $\boldsymbol{C}^\star$.
- **Bad news**: there are K^n possible partitions $\rightsquigarrow$ enumeration is not an option.

In fact, K-means problem is a **nonconvex NP-hard** problem and one need to resort to fast heuristics.
⚠With a slight abuse, we drop distinction between K-means problem and heuristics to solve it.

The K-means algorithm (MacQueen 1967)

Draw centroids $\mu_1, \dots, \mu_K$ at random among the sample $\boldsymbol{X}$ and

1. Assign each point to its closest centroid

$$C_k \leftarrow \left\{ i \,:\, \|x_i - \mu_k\|_2^2 = \min_j \|x_j - \mu_k\|_2^2 \right\}$$

2. recompute centroids as the barycenter of each center

$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} y_i$$

3. Go to 1 until clusters (hence barycenters) are unchanged

Properties of the algorithm

K-means is a greedy algorithm which

- monotonically decreases the criterion
- converges in a finite number of iterations
- will get stuck in local minima of L (non-convex)

$\rightsquigarrow$ In practice, we try several restarts with different random inits.

Extensions

Kmeans++ initialization matter ! $\rightsquigarrow$ stop drawing centroids at random

- Choose μ_1 uniformly among the sample
- then sequentially do for each $k = 2, \dots, K$
  - compute weight $w_i := \min_{j<k} \|x_i - \mu_j\|_2^2$
  - Choose μ_k among the sample with proba $\propto w_i$

Optimality bounds can be obtained (Arthur et al. 2007)

Sparse K-means include variable selection, useful when x_i in dimension $d \gg n$

Kernel K-means compute distance between $\phi(x_i)$ with $\phi : \mathcal{X} \to \mathcal{H}$ a *feature map*.

Mixture models

Probabilistic view on clustering

The partition is now seen as a set of discrete latent variables $\boldsymbol{Z} = \{z_1, \ldots, z_n\}$

Denote $\pi = (\pi_1, \ldots, \pi_K)$ the (unknown) cluster proportions, we have

$$p_\pi(z_{ik} = 1) = \pi_k \iff z_i \sim \mathcal{M}(1, \pi)$$

Mixture models

For all $i = 1, \ldots, n$, mixture models suppose that (z_i, x_i) are drawn *i.i.d.* according to the two-stage hierarchical model

1. $Z_i \sim \mathcal{M}_K(1, \pi)$
2. $X_i \mid \{z_{ik} = 1\} \sim p_{\gamma_k}$

The model parameters are $\theta = \{\pi_k, \gamma_k\}_{k=1}^K$ and p_γ can be any parametric distribution over X_i.

Clusters are sometimes called *components*

$\rightsquigarrow$ **general** and **flexible** framework, adapt to nature of the data (discrete, continuous, mixed-type)via p_γ

Observed (marginal) likelihood

Properties: independence

In a mixture model, $(Z_i)_i$ are *i.i.d.* and $(X_i)_i$ also are *i.i.d.*

Observed likelihood

$$
\begin{aligned}
p_\theta(\boldsymbol{X}) = \sum_{z_1,\dots,z_n} p_\theta(\boldsymbol{Z}, \boldsymbol{X}) &= \sum_{z_1,\dots,z_n} \prod_{i=1}^{n} p_\theta(X_i \mid z_i) p_\theta(z_i), \\
&= \prod_{i=1}^{n} \sum_{z_i} p_\gamma(X_i \mid z_i) p_\theta(z_i), \\
&= \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k \log p_{\gamma_k}(X_i) \right).
\end{aligned}
$$

$\rightsquigarrow$ the marginal distribution of X_i is a convex combination (*mixture*) of the K base distributions $(p_{\gamma_k})_k$, with weights π_k.

Complete likelihood

Properties: conditional independence

In a mixture model, $(X_i)_i \perp\!\!\!\perp \boldsymbol{Z}$ and $(Z_i)_i \perp\!\!\!\perp \boldsymbol{X}$, **but not** identically distributed

Complete log-likelihood

$$\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) = \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{X} \mid \boldsymbol{Z}) = \sum_{i=1}^{n} \log p_\pi(Z_i) + \log p_\gamma(X_i \mid Z_i),$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} Z_{ik} \left[\log \pi_k + \log p_{\gamma_k}(X_i)\right].$$

Posterior distribution of $Z \mid X$

For $i = 1, \ldots, n$, $Z_i \mid X_i \sim \mathcal{M}_K(1, \tau_i)$ with

$$\tau_{ik} := p_\theta(z_{ik} = 1 \mid X_i) \propto \pi_k p_{\gamma_k}(X_i)$$

Notice that τ_i also depends on the parameters θ.