Chapter 8, Part 1: Latent Variable Models

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

Course Feedback

Lectures Feedback Form

Part C Classes

MSc Classes

Overview

So far we have focused mostly on supervised learning, we now cover an important class of unsupervised (or sometimes semi–supervised) approaches: **latent variable models** (LVMs)

LVMs are a probabilistic **generative** modeling approach to unsupervised learning

They provide a mechanism for reasoning about **how** data is generated, from which we can derive a variety of useful tasks:

- Representation/feature learning
- Clustering
- Density estimation
- Learning generative models

Latent Variable Models

A LVM is a generative model where each datapoint x_i has a corresponding **latent** variable z_i

The underlying assumption is that each datapoint is represented by its corresponding latent variable, that is (given our model) it provides the information needed to generate the datapoint

z_i is often lower dimensional, simpler, or more interpretable than x_i

For data $\mathbf{X} = \{x_i\}_{i=1}^n$ and latents $\mathbf{Z} = \{z_i\}_{i=1}^n$, a LVM is given by

$$p_\theta(\mathbf{X}, \mathbf{Z}) = p_\theta(\mathbf{Z}) \prod_{i=1}^n p_\theta(x_i|z_i) \tag{1}$$

Here θ are global variables that are usually treated deterministically

Occasionally we take a Bayesian approach for θ as well, giving

$$p(\theta, \mathbf{X}, \mathbf{Z}) = p(\theta)p(\mathbf{Z}|\theta) \prod_{i=1}^n p(x_i|z_i, \theta)$$

Example: Mixture Models

Mixture models are LVMs that assume $\mathbf{X}$ was created by sampling iid from K distinct populations called **mixture components**

We thus have the following generative model for $i = 1, \ldots, n$:

$$Z_i \sim \text{Discrete}(\pi_1, \ldots, \pi_K) \qquad \text{i.e., } \mathbb{P}(Z_i = k) = \pi_k$$
$$X_i | Z_i = k \sim f_k(x; \lambda_k)$$

where $\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$ are known as **mixture weights** (or mixing proportions) and $f_k(x; \lambda_k)$ is a local model for the k^{th} **mixture component** with parameters λ_k. Here $\theta = \{\pi_k, \lambda_k\}_{k=1}^{K}$

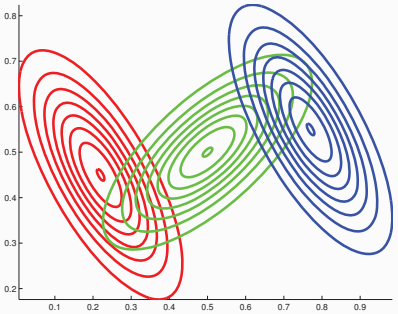
Mixture models can can be used for both clustering and density estimation; we observe $\mathbf{X}$ and want to learn about θ and $\mathbf{Z}$

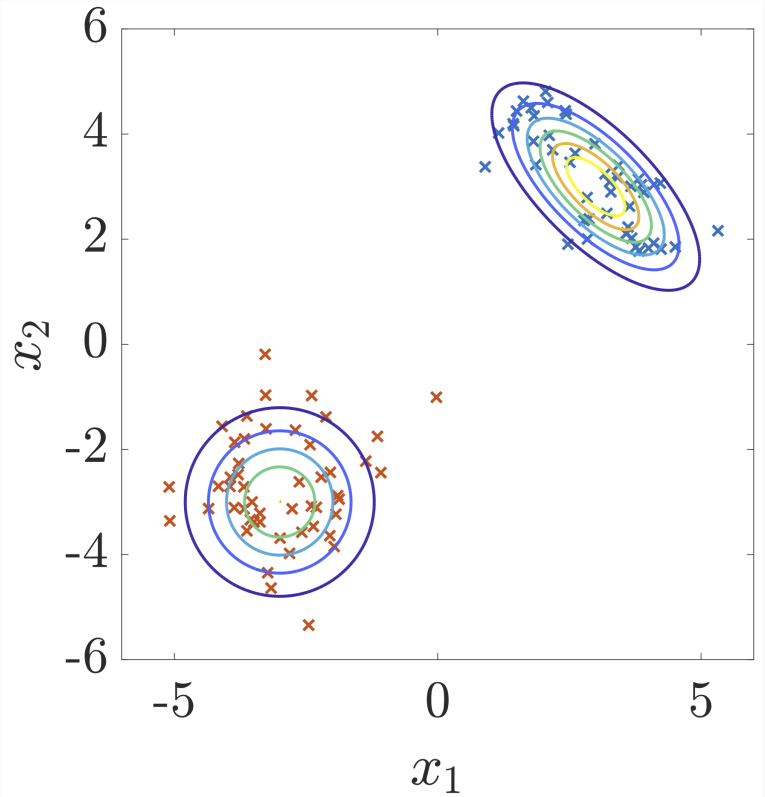
Example: Gaussian Mixture Models

A common choice of f_k is a Gaussian, yielding a Gaussian mixture model (GMM) for which $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ and

$$p_\theta(z_i|\mathbf{X}) = \frac{\pi_{z_i}\mathcal{N}\left(x_i; \mu_{z_i}, \Sigma_{z_i}\right)}{\sum_{k=1}^K \pi_k \mathcal{N}\left(x_i; \mu_k, \Sigma_k\right)}$$

$$p_\theta(\mathbf{X}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \mathcal{N}\left(x_i; \mu_k, \Sigma_k\right)$$





Credit: Murphy, 2012, Ch. 11.

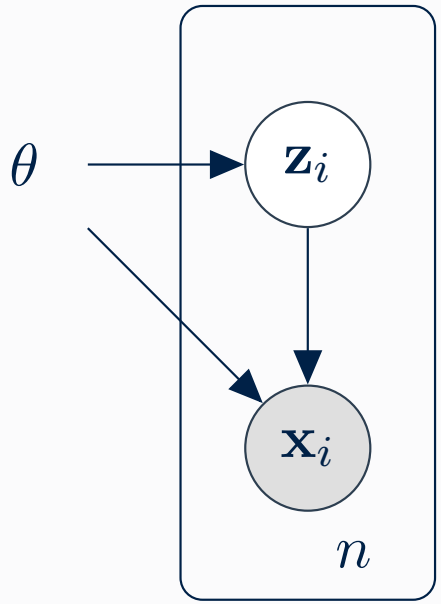
Factorized Latent Variable Models

If we assume that our data is i.i.d. given θ, this implies that all the z_i should also be conditionally independent given θ

This is typically a reasonable assumption when the data has no natural ordering

The resulting modal is known as a **factorized** LVM and has joint distribution

$$p_\theta(\mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} p_\theta(z_i) p_\theta(x_i|z_i)$$
$$= \prod_{i=1}^{n} p_\theta(x_i) p_\theta(z_i|x_i)$$



Why Factorized LVMs?

Lots of unsupervised data factorizes over datapoints and we can use this to learn very powerful models if we have lots of data

Glow: Generative flow with invertible 1x1 convolutions. Kingma and Dhariwal. 2018

Example: Images Transfer with GANs

Karras, Laine, and Aila. arXiv:1812.04948. 2019

Example: Multi-Modal Learning (e.g. Dall-E)

Hierarchical Text-Conditional Image Generation with CLIP Latents, Ramesh et al. 2022

Model Learning and Inference

There are two key tasks we need to perform for a factorized LVM:

- **Model learning**: we want to find the best configuration for θ
- **Inference**: for a given θ, we want to calculate the posterior

$$p_\theta(\mathbf{Z}|\mathbf{X}) \propto \prod_{i=1}^{n} p_\theta(z_i)p_\theta(x_i|z_i)$$

Note that this is **separable** in the sense that

$$p_\theta(z_i|\mathbf{X}) = p_\theta(z_i|x_i) \propto p_\theta(z_i)p_\theta(x_i|z_i)$$

$$p_\theta(\mathbf{X}) = \prod_{i=1}^{n} p_\theta(x_i)$$

Model Learning / Type–II Maximum Likelihood

To do model learning, we can try to directly optimize the marginal likelihood w.r.t. θ

This is know as **type–II maximum likelihood** or **maximum marginal likelihood (MML)** estimation

$$\theta^* = \arg\max_\theta p_\theta(\mathbf{X}) = \arg\max_\theta \mathbb{E}_{p_\theta(\mathbf{Z})}\left[p_\theta(\mathbf{X}|\mathbf{Z})\right]$$

For factorized LVMs this simplifies to

$$\theta^* = \arg\max_\theta \prod_{i=1}^n p_\theta(x_i) = \arg\max_\theta \sum_{i=1}^n \log p_\theta(x_i)$$

Next time we will start looking into how we can (approximately) perform both MML estimation and inference