Fondamentaux de l'Apprentissage Automatique
Latent Variable Models

Lecturer: Yann Chevaleyre
Scribe: Zhe HUANG

Lecture n°10 #
07/12/2023

1 Introduction

Latent variables appear pervasively across various domains. However, in this lecture, we will specifically focus on generative problems within the scope of unsupervised learning. Generative problems are characterized by the task of learning a distribution over data $\{x_1, \ldots, x_N\} \in \mathbb{R}^d$, with the aim of generating new data points that are representative of this learned distribution.

Latent variable models are a cornerstone in the field of probabilistic modeling, providing a robust framework for unveiling the hidden structure in observed data. These models are particularly prominent in generative problems, though their utility spans a wide range of applications. Central to these models is the concept of a latent variable, z_i, which embodies the missing information of an observed instance x_i.

1.1 Clustering Setting Example

In clustering applications, each data point x_i is typically associated with a feature set, while the latent variable z_i signifies the group membership, encapsulating the data's implicit characteristics that determine its clustering. Such latent variables are invaluable for identifying the intrinsic groupings within the data, which may not be observable through direct analysis.

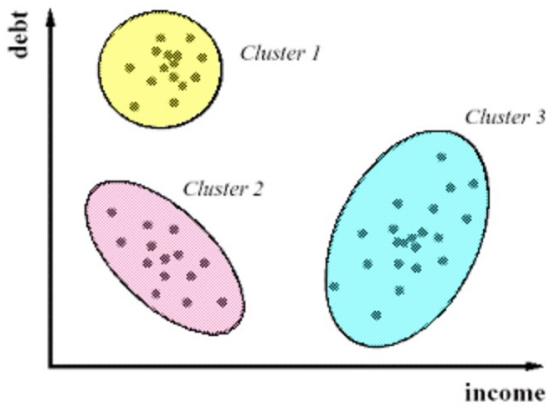

FIGURE 1 – Example of clustering as latent variables

1.2 Dimensionality Reduction Example

In dimensionality reduction, latent variables prove to be pivotal. They extract and compress the quintessential details from data that may span across many dimensions, simplifying

it into a more comprehensible, reduced form. Envision handling a dataset laden with a multitude of features ; this is where the intricacy lies. Designated as z_i, these latent variables silently sift through the data to crystallize the fundamental patterns, linkages, and variabilities present in the initial data points, denoted by x_i. This condensation is crucial, facilitating a clearer and more straightforward interpretation and analysis of complex, multidimensional data.

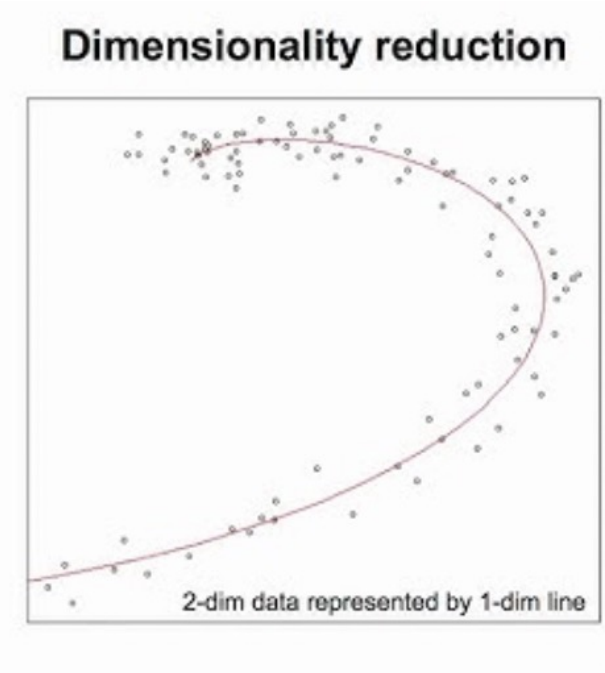

FIGURE 2 – Example of dimensionality redction with latent variables

1.3 Missing Data Example

Working with incomplete datasets is a frequent dilemma in statistical analysis. To tackle this, we employ latent variables z_i as mathematical proxies for missing entries. These variables are formulated based on the statistical properties derived from the observed portions of the data. Through algorithms like Expectation-Maximization (EM) or matrix factorization, we can estimate z_i to predict and impute the missing values.

(Usually, the best learning algorithm to estimate latent variables is the Expectation-Maximisation.)

1.4 Probabilistic Machine Learning

Probabilistic machine learning frameworks are designed to handle uncertainty inherent in data. They operate under the assumption that data arises from probability distributions characterized by latent variables. This contrasts with deterministic models by accounting for the randomness present in real-world data collection and generation processes.

In this probabilistic setting, we often assume data points are generated from a parametric model, which includes latent variables to capture complex, underlying phenomena not directly observable. The process involves estimating these latent variables and the parameters

of the probabilistic model that best explain the observed data. Commonly, this estimation is achieved through optimization techniques such as Maximum Likelihood Estimation (MLE) or Bayesian methods such as Maximum A Posteriori (MAP) estimation.

1.4.1 Generative Models

Consider a generative model for a set of data points where each point X_i is derived from a function of latent variables Z_i, typically governed by a distribution, such as a Gaussian for continuous data. Noise is also factored into the model to account for variations and measurement errors. The model can be expressed mathematically as :

$$X_i = g(Z_i) + \epsilon_i \tag{1}$$

where g is a function mapping latent variables to the observed data space and ϵ_i represents the noise.

The probability distribution of the observed data given the latent variables, $P(X|Z)$, and the distribution of the data itself, $P(X)$, are integral to understanding the generative process. These distributions enable the computation of the marginal likelihood of the observed data and the posterior distribution of the latent variables, refining our model and predictions about unseen data.

1.4.2 Learning from Data

The aim is to learn the distribution of X from the dataset, enabling prediction and inference about both observed and unobserved phenomena. This is articulated through the likelihood function and the prior distribution over the latent variables. The complete data likelihood, incorporating latent variables, is given by :

$$P(X) = \int P(X|Z)P(Z)dZ \tag{2}$$

Bayesian inference then updates our knowledge about the latent variables in light of observed data using Bayes' theorem :

$$P(Z|X) = \frac{P(X|Z)P(Z)}{P(X)} \tag{3}$$

These generative models may include unobserved variables. We call these **Latent Variables**.

2 Gaussian mixtures and the EM algorithm

2.1 Univariate Gaussian Distribution

The Univariate Gaussian Distribution, commonly known as the Normal Distribution, is a fundamental concept in statistics, encapsulating a single random variable. It is characterized by two main parameters : the mean (μ), which locates the center of the distribution, and the standard deviation (σ), which measures the spread of the data around the mean. The Probability Density Function (PDF) for this distribution is given by the Gaussian function :

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4}$$

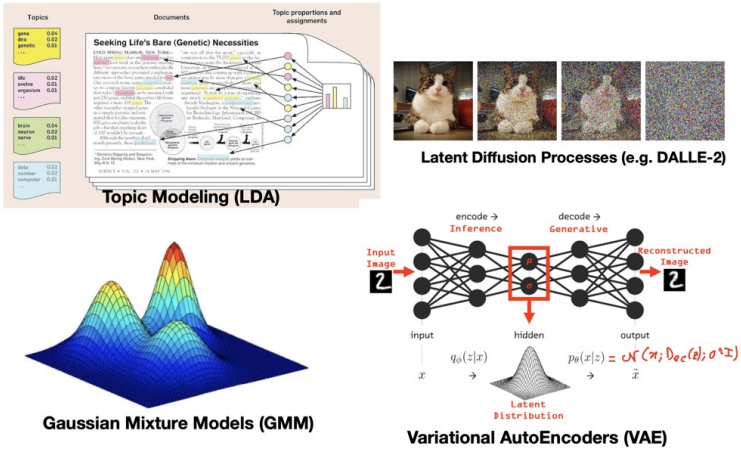

FIGURE 3 – Example of generative models with Latent Variables

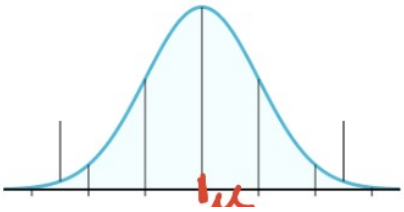

FIGURE 4 – The probability density function of the Univariate Gaussian Distribution.

2.2 Multivariate Normal Distribution

Expanding upon the Univariate Gaussian, the Multivariate Normal Distribution generalizes to higher dimensions, encapsulating a vector of random variables. This distribution is parametrized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, which together describe the location and shape of the distribution in multidimensional space.

The Probability Density Function (PDF) of the Multivariate Normal Distribution is defined as :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{5}$$

where $\mathbf{x}$ is a k-dimensional random vector, $\boldsymbol{\mu}$ is the mean vector, and Σ is the $k \times k$ covariance matrix.

When Σ is the identity matrix, the covariance between any pair of variables is zero, implying that they are uncorrelated, and the PDF simplifies to :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{k}{2}}} \exp\left(-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}\|^2\right) \tag{6}$$

This special case describes a spherical distribution where the variance in each dimension is equal, and the variables are independent.

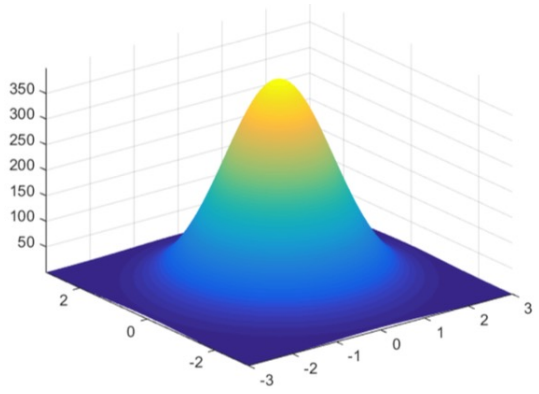

FIGURE 5 – Gaussian Distribution in 2 variables

2.3 Gaussian Mixture Models

The Gaussian Mixture Model (GMM) is among the simplest latent variable models and is akin to Linear Discriminant Analysis (LDA) without class information. The LDA for two classes is essentially a generative model with two Gaussians, each class being modeled by a different Gaussian distribution.

Linear Discriminant Analysis serves as a foundation for understanding Gaussian Mixture Models. LDA assumes that different classes generate data based on Gaussian distributions with class-specific means but a shared covariance matrix. For a two-class problem, the data generation process for LDA can be modeled as :

$$X_i|Y_i = k \sim \mathcal{N}(\mu_k, \Sigma), \quad k \in \{1, 2\}$$

where X_i is the observed data, Y_i is the class label, μ_k is the mean vector for class k, and Σ is the common covariance matrix across classes.

GMM extends LDA to scenarios lacking explicit class labels by utilizing a mixture of Gaussians to model the data probabilistically. In GMM, each component of the mixture corresponds to a latent class, and data points are considered to be generated from this mixture.

A GMM with two components is defined by the following :

$$Z_i \sim \text{Ber}(\pi) + 1$$

$$X_i \sim \begin{cases} \mathcal{N}(\mu_a, \Sigma_a), & \text{if } Z_i = 1, \\ \mathcal{N}(\mu_b, \Sigma_b), & \text{if } Z_i = 2. \end{cases}$$

The observed dataset is $\{X_1, X_2, \ldots, X_N\}$, and Z_i are the latent variables indicating the component from which X_i was generated.

2.4 Computing the Log Likelihood of a GMM Model

To compute the log likelihood of a Gaussian Mixture Model (GMM), we consider a model comprising K Gaussian distributions $\mathcal{N}(\mu_k, \Sigma_k)$, each with an associated prior probability π_k for $k \in \{1 \ldots K\}$. Let θ represent the parameter set for the GMM, comprising all means μ_k, covariance matrices Σ_k, and prior probabilities π_k.

Given a dataset $\mathbf{X} = \{x_1, \ldots, x_N\}$ in $\mathbb{R}^N$, the log likelihood $\mathcal{LL}(\theta)$ of the model parameterized by θ is expressed as :

$$\mathcal{LL}(\theta) = \log P_\theta(\mathbf{X}) = \sum_{i=1}^{N} \log P_\theta(x_i) \tag{7}$$

$$= \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} P_\theta(x_i | z_i = k) P_\theta(z_i = k) \right) \tag{8}$$

$$= \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \mathcal{N}(x_i | \mu_k, \Sigma_k) \pi(k) \right) \tag{9}$$

The maximum likelihood estimate $\hat{\theta}$ is the set of parameters that maximizes $\mathcal{L}(\theta)$:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{LL}(\theta) \tag{10}$$

Note that for $K = 1$, computing $\hat{\theta}$ is straightforward as it involves well-understood optimization of a single Gaussian distribution. However, for $K > 1$, the optimization becomes non-convex, often challenging to compute directly.

2.5 GMM with $K > 1$

2.5.1 The Complete Log Likelihood

For Gaussian Mixture Models (GMM) with more than one component ($K > 1$), maximizing the likelihood directly is challenging due to the model's complexity. An alternative approach is to optimize the Complete Log Likelihood (CLL), which serves as a surrogate to the true likelihood. This method assumes that we have knowledge of the latent variables $z_1, \ldots, z_N$, although in practice, these are also unknown and must be estimated.

The CLL is defined as the logarithm of the joint probability of the observed data $x_1, \ldots, x_N$ and the latent variables $z_1, \ldots, z_N$, and is given by :

$$\mathcal{CLL}(\theta, z_1, \ldots, z_N) = \log P_\theta(x_1, \ldots, x_N, z_1, \ldots, z_N) \tag{11}$$

Where θ includes the mixture weights, mean vectors, and covariance matrices of all K Gaussian components.

2.5.2 The Expected Complete Log Likelihood

In practive, we cannot compute the latent variables $z_1, \ldots, z_N$. But if for each data point i, we introduce an estimate $q_i(k)$ for the probability $p(\mathbf{z}_i = k | \mathbf{x}_i; \boldsymbol{\theta})$, then the Expected Complete Log Likelihood (ECLL) can be computed.

The mathematical expression for the ECLL is :

$$\mathcal{L}(q_i, \boldsymbol{\theta}_i) = \mathbb{E}_{\mathbf{z}_i \sim q_i} \left[\log P_\theta(\mathbf{x}_i, \mathbf{z}_i) \right] = \sum_{k=1}^{K} q_i(k) \log \left(P_\theta(\mathbf{x}_i | \mathbf{z}_i = k) \cdot \pi_k \right) \tag{12}$$

$$\text{ECLL} = \delta(q, \boldsymbol{\theta}) = \sum_{i=1}^{N} \alpha(q_i, \boldsymbol{\theta}_i, \lambda) = \mathbb{E}_{\mathbf{z}_1 \sim q_1, \ldots, \mathbf{z}_N \sim q_N} \left[\mathcal{CLL}(\boldsymbol{\theta}, \mathbf{z}_1, \ldots, \mathbf{z}_N) \right] \tag{13}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} q_i(k) \log \left[P_\theta(\mathbf{x}_i | \mathbf{z}_i = \ell) \pi_k \right] \tag{14}$$

2.6 The EM Algorithm

The Expectation-Maximization (EM) algorithm is a cornerstone in the statistical analysis of latent variable models. It iteratively optimizes the likelihood function when the model includes unobserved latent variables.

1. **Initialization :** Choose initial parameters $\hat{\theta}$ arbitrarily.
2. **E-Step :** Estimate the latent variable distribution $p(z_i = k | x_i; \hat{\theta})$ for each observation i and each latent class k.
3. **M-Step :** Maximize the expected complete log-likelihood with respect to $\hat{\theta}$.

During the E-Step, we calculate the posterior probabilities $q_i(k)$ which serve as weights for the observations in the subsequent optimization of the M-Step. The M-Step then updates the parameters to maximize the weighted likelihood, effectively refining the model with respect to the observed data.

Details : For a Gaussian Mixture Model with K components, the parameter set includes mean vectors, covariance matrices, and mixing coefficients. The algorithm's goal is to find the set of parameters that maximize the likelihood of the observed data, given the Gaussian mixture model.

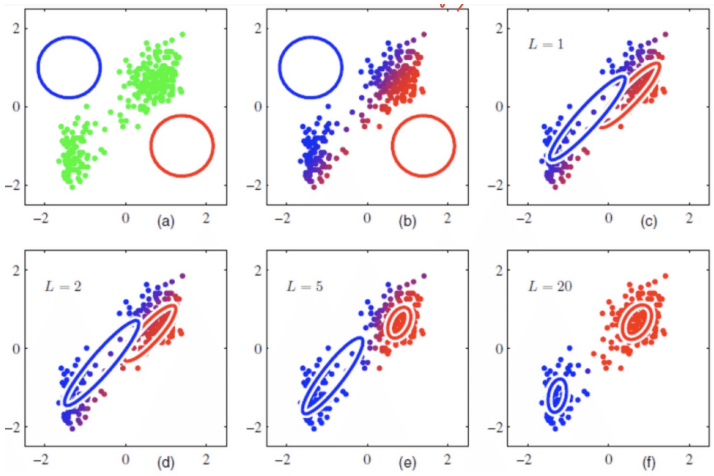

FIGURE 6 – The EM Algorithm

3 Variational Analysis of the EM algorithm

We wanted to optomise the likelihood $p_\theta(x_1, ..., x_N)$, but instead we optimise the surrogate : the expected Complete Log Likelihood (ECLL), $\mathcal{L}(q, \theta)$. What is the link ?

3.1 Linking ECLL to Likelihood

The ECLL is a sum over all data points, capturing the expected log-likelihood for each point when considering the distribution of latent variables :

$$\mathcal{L}(q, \theta) = \sum_{i=1}^{N} \mathcal{L}(q, \theta, i), \tag{15}$$

where $\mathcal{L}(q, \theta, i)$ is the expected log-likelihood contribution of the data point x_i, and q_i is the distribution of its associated latent variable.

To illustrate, let's compute the log-likelihood of an individual observation x under the model parameters θ. This involves summing over all potential latent variable states k, weighted by their estimated probabilities $q_x(k)$:

$$\log P_\theta(x) = \log \sum_{k=1}^{K} q_x(k) \cdot \frac{P_\theta(x, z = k)}{q_x(k)}. \tag{16}$$

Applying Jensen's inequality(which holds for any concave function $f, \mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$), we derive a lower bound for our log-likelihood :

$$\log P_\theta(x) \geq \mathbb{E}_{k \sim q_x} \left[\log P_\theta(x, z = k) \right], \tag{17}$$

$$\log P_\theta(x) \geq \sum_{k=1}^{K} q_x(k) \log P_\theta(x, z = k) - \sum_{k=1}^{K} q_x(k) \log q_x(k). \tag{18}$$

where the first term ECLL $= \alpha(q_x, \theta) = \sum_{k=1}^{K} q_x(k) \log P_\theta(x, z = k)$ and the second term $-\sum_{k=1}^{K} q_x(k) \log q_x(k)$ represents the Entropy. The sum of these two term establishes the Evidence Lower Bound(ELBO).

3.2 The Evidence Lower Bound (ELBO)

The ELBO serves as a cornerstone in the variational analysis of the EM algorithm. It not only provides a computationally tractable surrogate for the log-likelihood but also facilitates the optimization process by bounding the likelihood from below.

The relationship between the likelihood and the ELBO can be shown as below :

$$\ln P_\theta(x) - \text{ELBO}(x, \theta, q_x) = \ln P_\theta(x) - \mathbb{E}_{k \sim q_x} \left[\ln \frac{f(x, z = k)}{q_x(k)} \right] \tag{19}$$

$$= \mathbb{E}_{k \sim q_x} \left[\ln \frac{p_\theta(x) q_x(k)}{P_\theta(x, z = k)} \right] = \mathbb{E} \left[\ln \frac{q_x(k)}{P_\theta(z = k|x)} \right] \tag{20}$$

$$= \text{KL}(q_x(z) \parallel P_\theta(z|x)) \tag{21}$$

The Kullback-Leibler (KL) Divergence possesses several useful properties :

$$\begin{aligned}
&1. KL(q||q') = \mathbb{E}_{k \sim q} \left[\ln \frac{q(x)}{q'(x)} \right] \\
&2. KL(q||q') \geq 0 \quad \forall q, q' \\
&3. KL(q'||q) = 0 \quad \text{iff } q = q' \\
&4. KL(q'||q) \neq KL(q||q')
\end{aligned}$$

From these properties, we can derive :

$$\ln P_\theta(x) = \mathrm{ELBO}(\theta, x, q_x) \quad \text{iff} \quad q_x(z) = P_\theta(z \mid x) \tag{22}$$

$$\arg\max_{q_x(\cdot)} \mathrm{ELBO}(x, \theta, q_x) = P_\theta(z \mid x) \tag{23}$$

$$\arg\max_{\theta} \mathrm{ELBO}(x, \theta, P_\theta(z \mid x)) = \arg\max_{\theta} \ln P_\theta(x) \tag{24}$$

$$\arg\max_{\theta, q_x} \mathrm{ELBO}(x, \theta, q) = \arg\max_{\theta} \ln P_\theta(x) \tag{25}$$