Latent-Variable Models: Gaussian-Mixture & Other Cases

Qinliang Su （苏勤亮）

Sun Yat-sen University

suqliang@mail.sysu.edu.cn

Outline

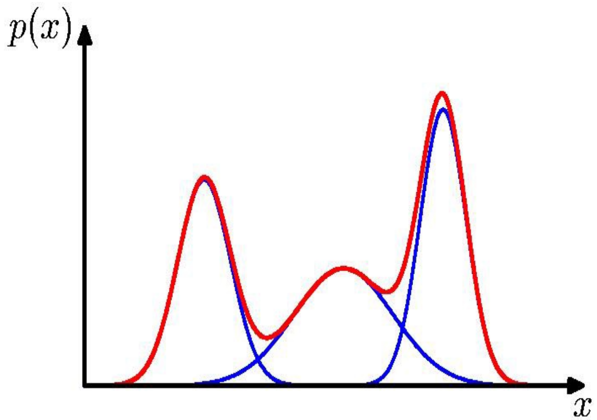
- Gaussian Mixture Distribution
- Learning the Distribution Parameters
- Other Examples of LVMs

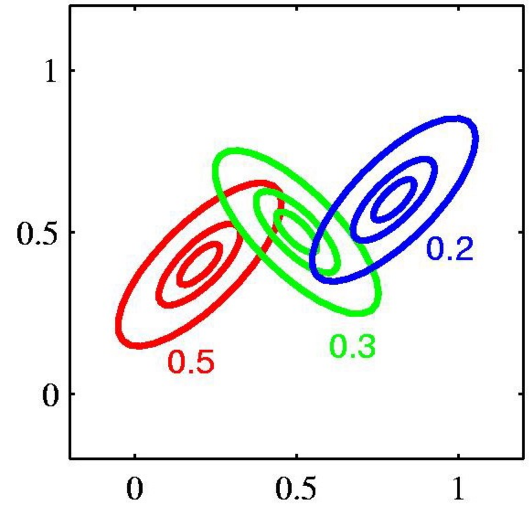
Gaussian Mixture Distributions

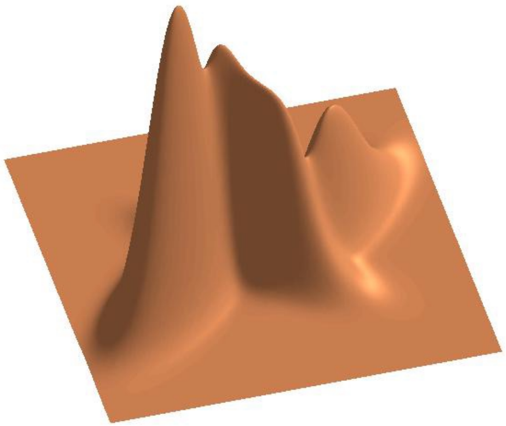
- The distribution expression

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

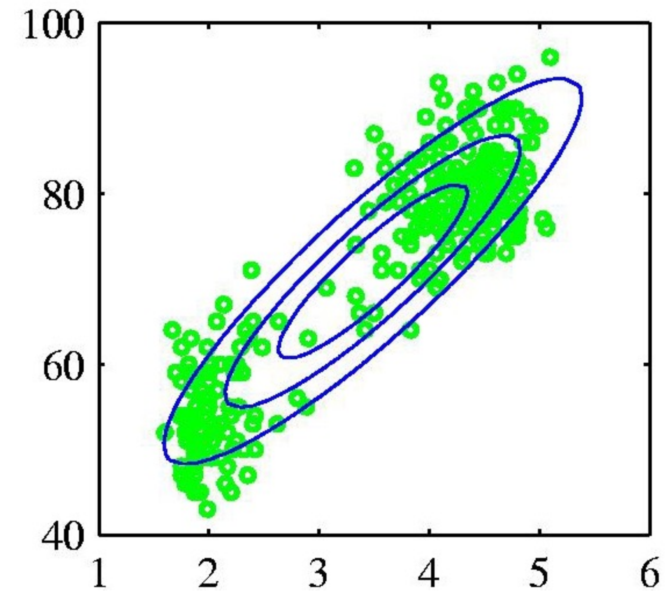
  - K is the number of Gaussian distributions
  - π_k is the weight of the *k*-th distribution with $\sum_{k=1}^{K} \pi_k = 1$
  - $\boldsymbol{\mu}_k$ *and* $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th Gaussian distribution



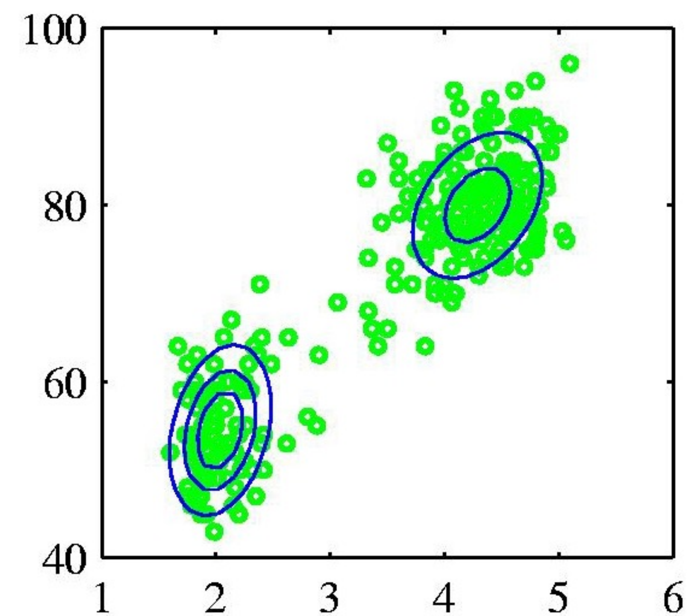




- It is very difficult to model the green points by a Gaussian distribution



- But if we model it with the mixture of two Gaussian distributions, it looks much better



Representing Gaussian Mixture Distribution as LVM

- For a latent-variable model $p(\boldsymbol{x}, \boldsymbol{z})$, if we set its conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$ and prior distribution $p(\boldsymbol{z})$ as

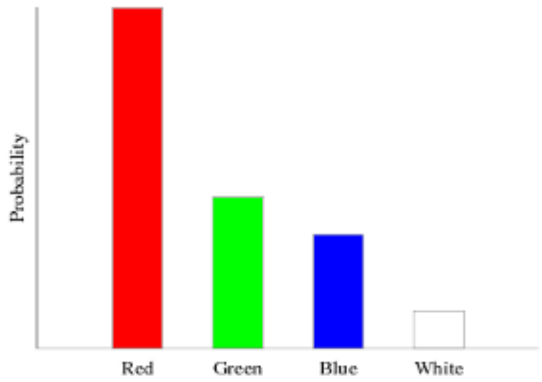
$$p(\boldsymbol{x}|\boldsymbol{z} = \mathbf{1}_k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\boldsymbol{z} = \mathbf{1}_k) = \pi_k$$

  - $\boldsymbol{z}$ can only be a one-hot vector, with $\mathbf{1}_k$ denoting the *k*-th element to be 1
  - $p(\boldsymbol{z} = \mathbf{1}_k) = \pi_k$, which is actually a categorical distribution, that is,

$$p(\boldsymbol{z}) = Cat(\boldsymbol{z}; \boldsymbol{\pi})$$

with $Cat(\boldsymbol{z} = \mathbf{1}_k; \boldsymbol{\pi}) = \pi_k$ and $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_K]$



- With $p(\boldsymbol{x}|\boldsymbol{z}=\mathbf{1}_k)=\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ and $p(\boldsymbol{z}=\mathbf{1}_k)=\pi_k$, we can easily see that

$$
\begin{aligned}
p(\boldsymbol{x},\boldsymbol{z}=\mathbf{1}_k) &= p(\boldsymbol{x}|\boldsymbol{z}=\mathbf{1}_k)p(\boldsymbol{z}=\mathbf{1}_k) \\
&= \pi_k\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)
\end{aligned}
$$

- Therefore, the joint distribution $p(\boldsymbol{x},\boldsymbol{z})$ can be written in a more compact form as

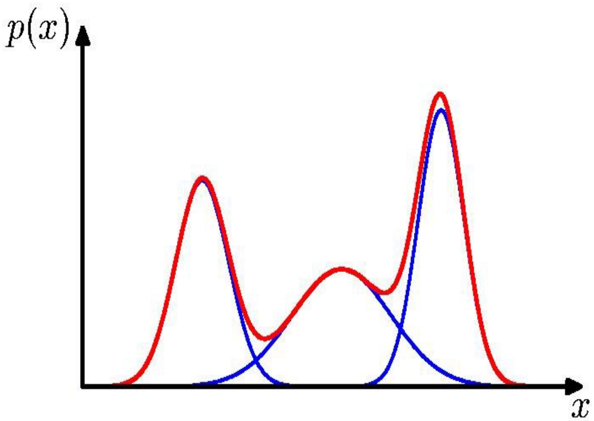
$$
p(\boldsymbol{x},\boldsymbol{z})=\prod_{k=1}^{K}[\pi_k\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)]^{z_k}
$$

where $\boldsymbol{z}=[z_1,z_2,\cdots,z_K]$ is a one-hot vector, that is, there is only one non-zero element (equal to 1) in $\boldsymbol{z}$

- Due to $p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z})$, we can easily see that

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

which is exactly the Gaussian mixture distribution



Gaussian mixture distributions can be equivalently represented by the latent-variable model

$$p(\boldsymbol{x}, \boldsymbol{z}) = \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_k}$$

Outline

- Gaussian Mixture Distribution
- Learning the Distribution Parameters
- Other Examples of LVMs

Training by Maximizing the Marginal

- Given a set of training data $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$, the goal is to learn the distribution parameters

$$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K} \triangleq \boldsymbol{\theta}$$

- The data points $\boldsymbol{x}^{(n)}$ are assumed *i.i.d*, thus we can write the joint distribution as

Not using the model with latent variable

$$p\left(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right) = \prod_{n=1}^{N} \underbrace{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}_{p(\boldsymbol{x}^n)}$$

- For probabilistic models, the training objective is *to maximize the log-likelihood function*, that is,

$$\boxed{\log p\left(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}$$

Maximizing $\log p\left(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right)$

- Substituting the expression of $\mathcal{N}\left(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ into it gives

$$\log p\left(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right)$$

$$= \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\right)\right\}\right)$$

- To optimize it, we require the *derivatives* of $\log p\left(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\right)$ *w.r.t.* the model parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

How to Use the Learned Model?

- After learning the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, that is, the distribution

$$p(\boldsymbol{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

is known, we can *use it to complete a lot of tasks*

- Example: Given a testing data point $\boldsymbol{x}$, can we use it to determine the probability that an $\boldsymbol{x}$ belongs to the k-th cluster?

$$p(\boldsymbol{x} \in k\text{-}th\ cluster) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^K \pi_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

- Can we explain the probability in a more principled way?

$$p(\boldsymbol{z} = \mathbf{1}_k | \boldsymbol{x}) = ?$$

$$p(\boldsymbol{z} = \mathbf{1}_k | \boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z}=\mathbf{1}_k)}{p(\boldsymbol{x})}$$

$$= \frac{p(\boldsymbol{x}, \boldsymbol{z}=\mathbf{1}_k)}{\sum_{i=1}^{K} p(\boldsymbol{x}, \mathbf{1}_i)}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

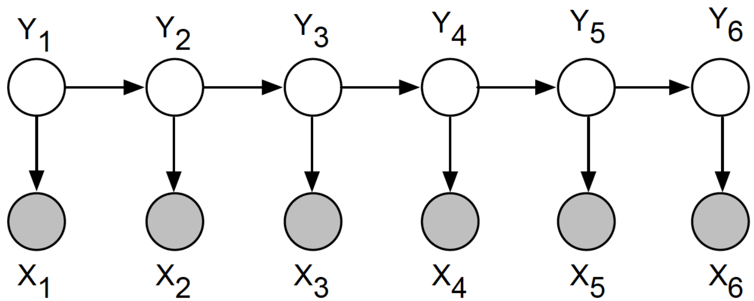
Thus, in the latent-variable model, the posteriori $p(\boldsymbol{z}|\boldsymbol{x})$ indicates the probability that a data instance belongs to different clusters

Outline

- Gaussian Mixture Distribution
- Learning the Distribution Parameters
- Other Examples of LVMs

Applications: Hidden Markov Model

- Hidden Markov Model (HMM)



- It is widely used in speech recognition, part-of-speech tagging, localization *etc.*

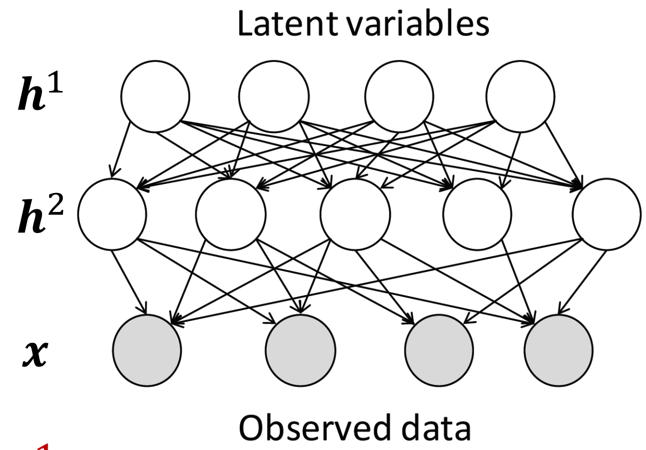
- Joint distribution

$$p(\boldsymbol{y}, \boldsymbol{x}) = p(y_1)p(x_1|y_1)\prod_{t=2}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$

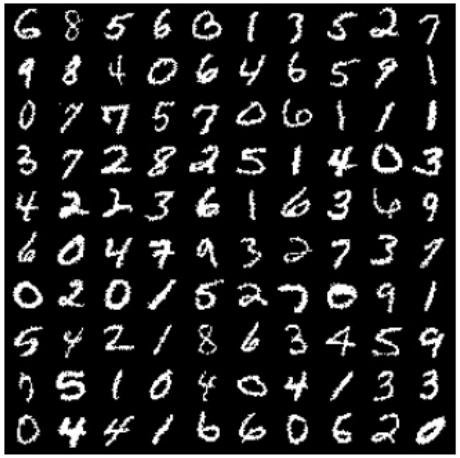
where $p(y_t|y_{t-1})$ is the transition probability; $p(x_t|y_t)$ is the emission probability

Applications: Image Modeling

- Sigmoid belief networks (SBN)
  - $h_i^1 \sim Bernoulli(0.5)$
  - $h_j^2 \sim Bernoulli\left(\sigma\left([\boldsymbol{W}_1\boldsymbol{h}^1 + \mathbf{b}_1]_j\right)\right)$
  - $x_k \sim Bernoulli\left(\sigma([\boldsymbol{W}_2\boldsymbol{h}^2 + \mathbf{b}_2]_k)\right)$

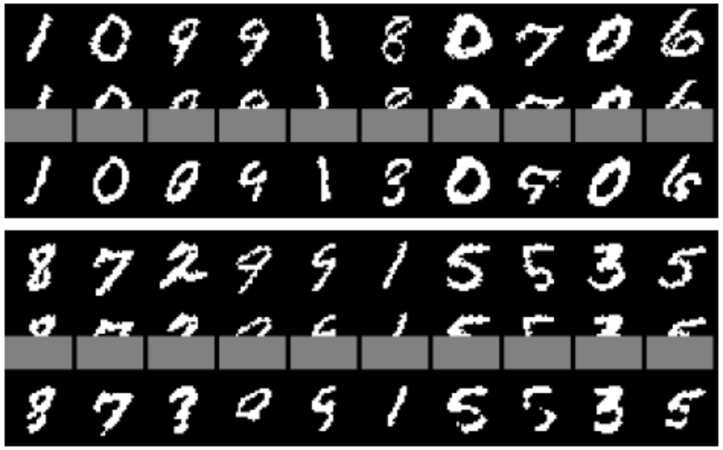


Joint pdf: $p(\boldsymbol{x}, \boldsymbol{h}^2, \boldsymbol{h}^1) = p(\boldsymbol{x}|\boldsymbol{h}^2)p(\boldsymbol{h}^2|\boldsymbol{h}^1)p(\boldsymbol{h}^1)$



Original

Generating



In-painting

Applications: Text Modeling

- Topic Model: Latent Dirichlet Allocation (LDA)
  - $\theta \sim Dir(\alpha)$: the distribution of different topics
  - $\varphi_k \sim Dir(\beta)$: the distribution of words for topic
  - $z_n \sim Multinomial(\theta)$: the topic of n-th word
  - $w_n \sim Multinomial(\varphi_{z_n})$: the n-th word