Latent-Variable Models & Its Gaussian Case

Qinliang Su （苏勤亮）

Sun Yat-sen University

suqliang@mail.sysu.edu.cn

Outline

- Latent-Variable Models
- Gaussian Latent-Variable Model
- Relation to the PCA

What are LVMs & Why They are Needed?

- In supervised learning, both regression and classification can be understood as learning *conditional probability distributions*

$$p(y|\boldsymbol{x};\boldsymbol{w})$$

  - In regression, the conditional pdf is assumed of the form

$$p(y|\boldsymbol{x};\boldsymbol{w}) = \mathcal{N}(y;\boldsymbol{w}^T\boldsymbol{x},\sigma^2)$$

  - For classification, the conditional pdf is assumed of the form

$$p(y|\boldsymbol{x}) = \big(\sigma(\boldsymbol{x}\boldsymbol{w})\big)^y \cdot \big(1-\sigma(\boldsymbol{x}\boldsymbol{w})\big)^{1-y}$$

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k=1}^{K}[softmax_k(\boldsymbol{W}\boldsymbol{x})]^{y_k}$$

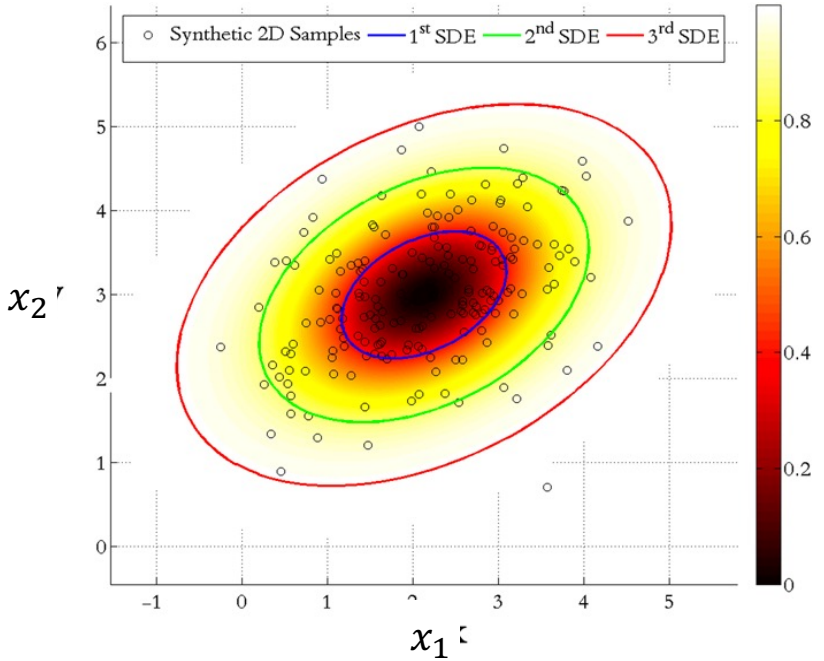
- Analogously, unsupervised learning can also be understood as learning a probability distribution, but it only concerns *input data* $\boldsymbol{x}$

$$p(\boldsymbol{x}; \boldsymbol{w})$$

- Modeling $\boldsymbol{x}$ is much difficult than modeling the label y. The simplest way is to restrict $p(\boldsymbol{x}; \boldsymbol{w})$ to the Gaussian form

$$p(\boldsymbol{x};\ \boldsymbol{w}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

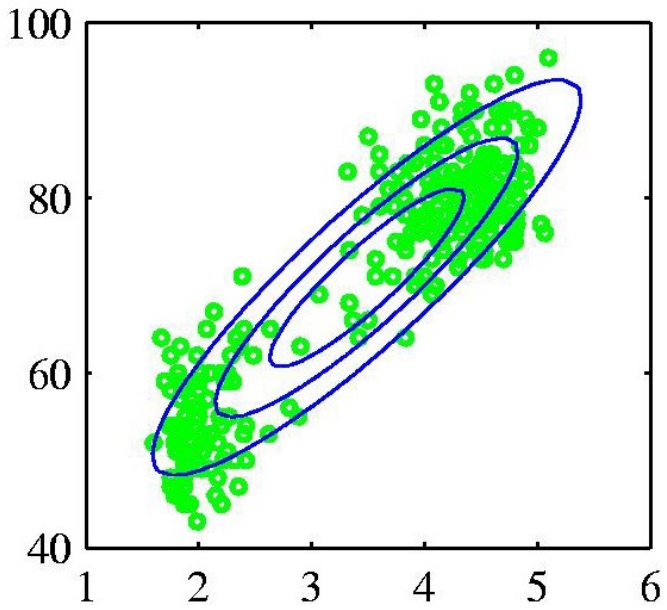
  - $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are optimized to describe the data points $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ best



If x_1 is observed to be 3, what value x_2 would most likely be?

- The distribution of data $\boldsymbol{x}$ is much more complex than what a simple Gaussian distribution can represent

For example, data points below cannot be fit well by a Gaussian distribution



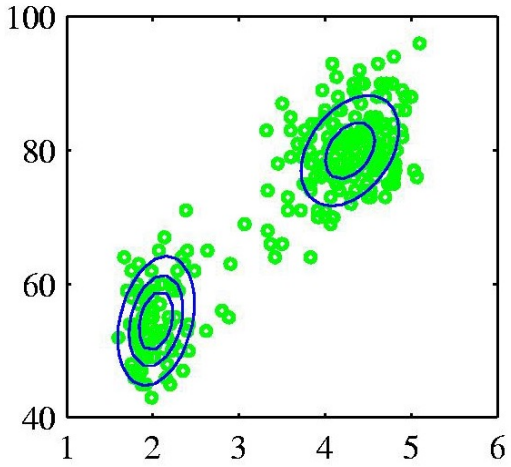
Obviously, to model the data well, the distribution $p(\boldsymbol{x})$ is required to be sufficiently expressive

- But how to obtain $p(\boldsymbol{x})$ that is able to *model complex data?*
  - A possible way is to composite 'many simple distributions' to constitute a expressive one, that is,

$$p(\boldsymbol{x}) = \pi_1 p_1(\boldsymbol{x}) + \cdots + \pi_K p_K(\boldsymbol{x})$$

where $p_k(\boldsymbol{x})$ are simple distributions; to ensure $p(\boldsymbol{x})$ is valid, $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$

For example, the previous data can be modelled well by two Gaussian distributions



- The distribution $p(\boldsymbol{x}) = \pi_1 p_1(\boldsymbol{x}) + \cdots + \pi_K p_K(\boldsymbol{x})$ can be derived from the following joint distribution

$$p(\boldsymbol{x}, z) = p(\boldsymbol{x}|z)p(z)$$

where $p(\boldsymbol{x}|z = k) = p_k(\boldsymbol{x})$ and $p(z = k) = \pi_k$

- Using the marginal-joint relation $p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}, k)$, we know that

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p_k(\boldsymbol{x})\pi_k$$

$$= \pi_1 p_1(\boldsymbol{x}) + \cdots + \pi_K p_K(\boldsymbol{x})$$

Thus, by defining a simple joint distribution $p(\boldsymbol{x}, z)$, it is possible to obtain a flexible marginal distribution $p(\boldsymbol{x})$

LVMs in General Form

- LVMs: a probabilistic model with latent variables

$$p(\boldsymbol{x}, \boldsymbol{z})$$

  - $\boldsymbol{x}$ is the random variable of interest
  - $\boldsymbol{z}$ is the latent variable (nuisance variable)

  - There sometimes exist multiple latent variables, *i.e.* $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_K$

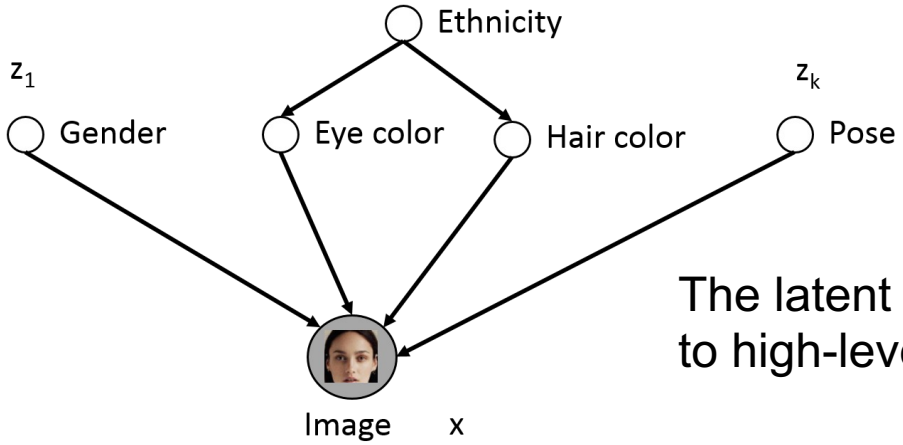
$$p(\boldsymbol{x}, \boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_K)$$

- The probabilistic model *w.r.t.* the interested variable $\boldsymbol{x}$

$$p(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \quad \text{or} \quad p(\boldsymbol{x}) = \int_{\boldsymbol{z}_1 \cdots \boldsymbol{z}_K} p(\boldsymbol{x}, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_K) d\boldsymbol{z}_1 \cdots d\boldsymbol{z}_K$$

Advantages of LVMs

- Representation ability: yielding expressive distributions by compositing simple ones
- Interpretability: latent variables can sometimes be associated with some physical meanings



The latent variables $\boldsymbol{z}$ often correspond to high-level features

- Integrating prior knowledges: injecting our prior knowledges about a task or data into the model
- Low-dimensional feature: the latent vector can often be used as the feature of a data instance

Outline

- Latent-Variable Models
- Gaussian Latent-Variable Model
- Relation to the PCA

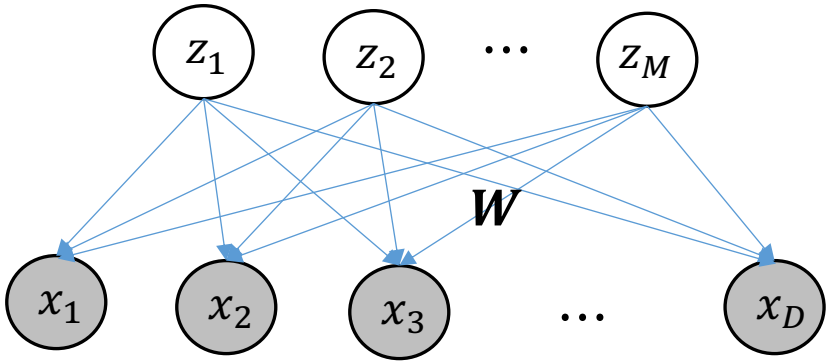
Modelling

- Assuming both of the prior and conditional pdfs are independent Gaussian

Prior distribution: $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$

Likelihood function: $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$

- Actually, the model describes how data samples $\boldsymbol{x}$ are generated



$\boldsymbol{z} = [z_1, \cdots, z_M]$ & $\boldsymbol{x} = [x_1, \cdots, x_D]$

Training Objective

- Given a set of samples $\{\boldsymbol{x}_n\}_{n=1}^N$, the question now becomes how to train the model $p(\boldsymbol{x}, \boldsymbol{z})$ so that it can describe the data best

- The model parameter $\boldsymbol{W}$ can be learned by maximizing the log-likelihood

$$\max_{\boldsymbol{W}} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n)$$

- In LVMs, what we have is the joint pdf

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n) = p(\boldsymbol{x}_n|\boldsymbol{z}_n)p(\boldsymbol{z}_n)$$

$$= \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{z}_n; \boldsymbol{0}, \boldsymbol{I}),$$

But what we need is to optimize $p(\boldsymbol{x}_n)$

Marginal Distribution $p(x)$

- The most direct method is to compute the marginal pdf first

$$p(\boldsymbol{x}_n) = \int_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n) d\boldsymbol{z}_n$$

- Deriving the analytical expression for $p(\boldsymbol{x}_n)$ is impossible in most scenarios due to existence of the integration

- But for the Gaussian case, we can easily obtain it as

$$p(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I})$$

A simple method to derive the marginal distribution

- From the model

$$\mathcal{N}(\boldsymbol{x}_n; \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})\mathcal{N}(\boldsymbol{z}_n; \boldsymbol{0}, \boldsymbol{I}),$$

the data point $\boldsymbol{x}_n$ can be understood as generated as

$$\boldsymbol{x}_n = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\epsilon}_n$$

where $\boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{z}_n; \boldsymbol{0}, \boldsymbol{I})$ *and* $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\boldsymbol{z}_n; \boldsymbol{0}, \sigma^2 \boldsymbol{I})$

- That is, data $\boldsymbol{x}_n$ can be understood as generated from $\boldsymbol{z}_n$ and $\boldsymbol{\epsilon}_n$ as $\boldsymbol{x}_n = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\epsilon}_n$

Lemma: Any linear combination of Gaussian random variables is also Gaussian

- Therefore, $\boldsymbol{x}_n$ also follows a Gaussian distribution

How can a Gaussian distribution be determined?

⇒ Mean & Covariance

- Mean & Covariance

Mean: $\mathbb{E}[\boldsymbol{x}_n] = \boldsymbol{\mu} + \boldsymbol{W}\mathbb{E}[\boldsymbol{z}_n] + \mathbb{E}[\boldsymbol{\epsilon}_n] = \boldsymbol{\mu}$

Covariance: $\mathbb{E}[(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T] = \boldsymbol{W}\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^T]\boldsymbol{W}^T + \mathbb{E}[\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T]$

$= \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$

- Thus, the marginal distribution of $\boldsymbol{x}_n$ is

$$p(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$$

Training by Optimizing $\log p(\boldsymbol{x})$

- Given the training dataset $\{\boldsymbol{x}_n\}_{n=1}^N$, to learn $\boldsymbol{W}, \boldsymbol{\mu}$ and σ^2, what we need to do is to optimize the log-probability

$$\log p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$$

- Due to $p(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$, we have

$$\log p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = \sum_{n=1}^N \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$$

- It can be further written as

$$\log p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = -\frac{ND}{2}\log 2\pi - \frac{N}{2}\log\det(\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$$

$$-\frac{1}{2}\sum_{n=1}^N (\boldsymbol{x}_n - \boldsymbol{\mu})^T (\boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

- By setting $\frac{\partial \log p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}{\partial \boldsymbol{\mu}} = 0$, we obtain

$$\boldsymbol{\mu} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n}{N}$$

$$\frac{\partial \ln \det(\boldsymbol{X})}{\partial \boldsymbol{X}} = (\boldsymbol{X}^{-1})^T$$
$$\frac{\partial \ln trace(\boldsymbol{X}^{-1}\boldsymbol{B})}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1}\boldsymbol{B}\boldsymbol{X}^{-1})^T$$

- By denoting $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$, we have

$$\frac{\partial \log p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)}{\partial \boldsymbol{\Sigma}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_n)(\boldsymbol{x}_n - \boldsymbol{\mu}_n)^T\boldsymbol{\Sigma}^{-1}$$

$$= -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{N}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}$$

  ⇒ Thus, it can be derived that $\boldsymbol{\Sigma} = \boldsymbol{S}$

- When $\boldsymbol{\Sigma}$ is restricted to the form $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$, it can be derived that

$$\boldsymbol{W} = \boldsymbol{U}\left(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}}$$

  - $\boldsymbol{U}$ consists of the top-M eigenvectors of $\boldsymbol{S}$
  - $\boldsymbol{\Lambda}$ is a diagonal matrix with the top-M eigenvalues of $\boldsymbol{S}$

Outline

- Latent-Variable Models
- Gaussian Latent-Variable Model
- Relation to the PCA

- Comparing the expression

$$\boldsymbol{W} = \boldsymbol{U}\left(\boldsymbol{\Lambda} - \sigma^{\mathbf{2}}\boldsymbol{I}\right)^{\frac{\mathbf{1}}{\mathbf{2}}}$$

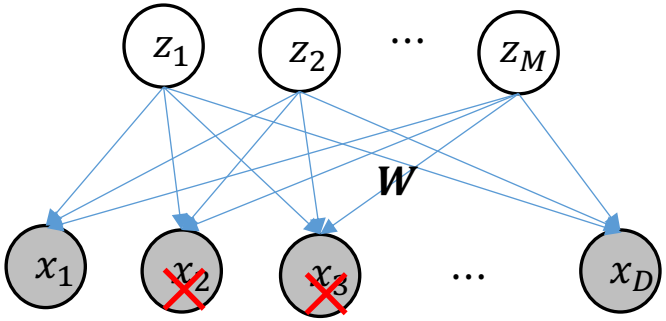
to the principle components of PCA, which are the matrix $\boldsymbol{U}$, we can see that

- $\boldsymbol{W}$ can be viewed as un-normalized principle components of data $\boldsymbol{x}_n$, *with the i-th component scaled by a coefficient* $\sqrt{\lambda_i - \sigma^2}$

Gaussian latent-variable models are called *probabilistic PCA*

Advantages of Probabilistic PCA over PCA

1) Being able to deal with incomplete data observation



2) Easily extend to more complex models, *e.g.*, let $p(\boldsymbol{z})$ be a mixture distribution, or each x_i is assigned a distinctive σ_i^2

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, diag(\boldsymbol{\sigma}^2)\right)$$

3) More computationally efficient: the optimal $\boldsymbol{W}$ can be learned by a mini-batch based SGD, rather than working on the whole dataset to compute the similarity matrix $\boldsymbol{S}$