Lecture 12: Generative Models

Shuai Li

John Hopcroft Center, Shanghai Jiao Tong University

https://shuaili8.github.io

https://shuaili8.github.io/Teaching/VE445/index.html

Outline

- Generative models
- GMM
- EM
- HMM

Generative Models (review)

Discriminative / Generative Models

- Discriminative models
  - Modeling the dependence of unobserved variables on observed ones
  - also called conditional models
  - Deterministic: $y = f_\theta(x)$
  - Probabilistic: $p_\theta(y|x)$
- Generative models
  - Modeling the joint probabilistic distribution of data
  - Given some hidden parameters or variables

$$p_\theta(x, y)$$

  - Then do the conditional inference

$$p_\theta(y|x) = \frac{p_\theta(x, y)}{p_\theta(x)} = \frac{p_\theta(x, y)}{\sum_{y'} p_\theta(x, y')}$$

Discriminative Models

- Discriminative models
  - Modeling the dependence of unobserved variables on observed ones
  - also called conditional models
  - Deterministic: $y = f_\theta(x)$
    - Linear regression
  - Probabilistic: $p_\theta(y|x)$
    - Logistic regression

  - Directly model the dependence for label prediction
  - Easy to define dependence on specific features and models
  - Practically yielding higher prediction performance
  - E.g. linear regression, logistic regression, k nearest neighbor, SVMs, (multi-layer) perceptrons, decision trees, random forest

Generative Models

- Generative models
  - Modeling the joint probabilistic distribution of data
  - Given some hidden parameters or variables

$$p_\theta(x, y)$$

  - Then do the conditional inference

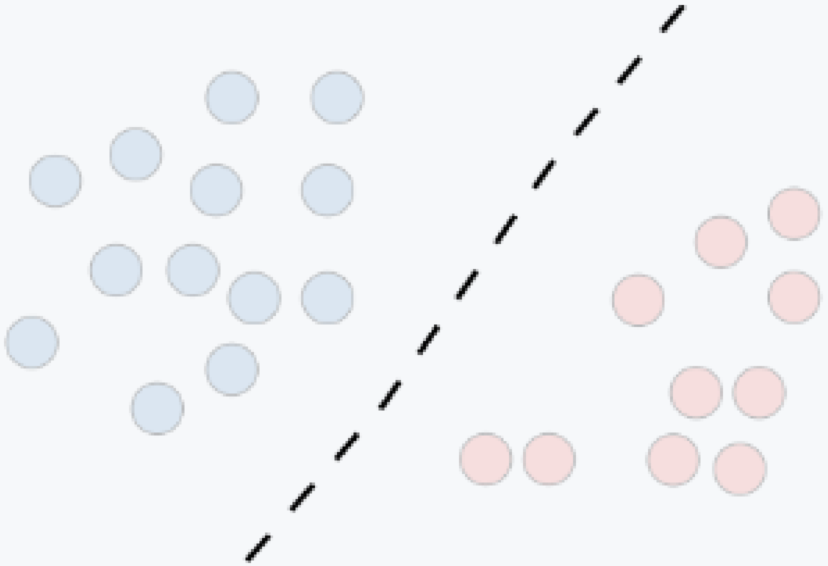
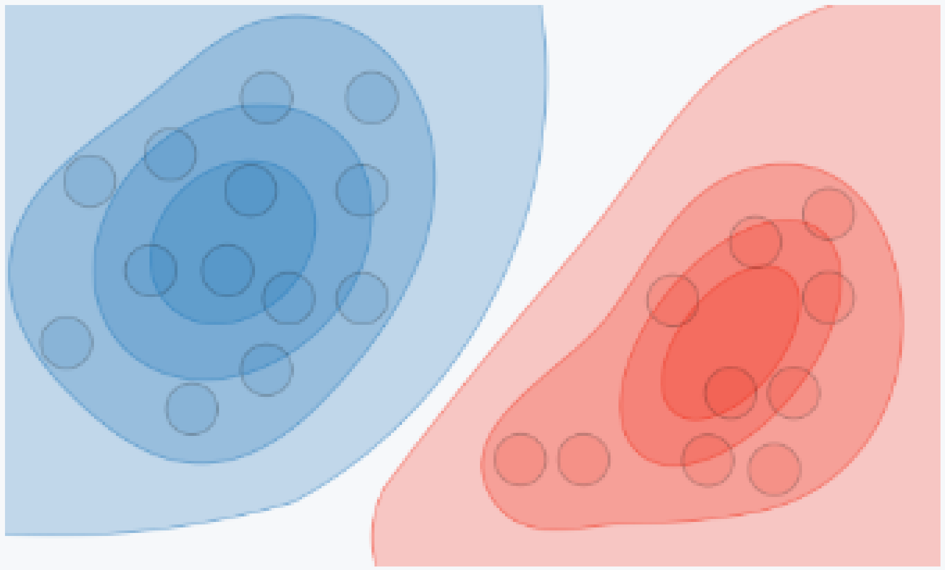
$$p_\theta(y|x) = \frac{p_\theta(x, y)}{p_\theta(x)} = \frac{p_\theta(x, y)}{\sum_{y'} p_\theta(x, y')}$$

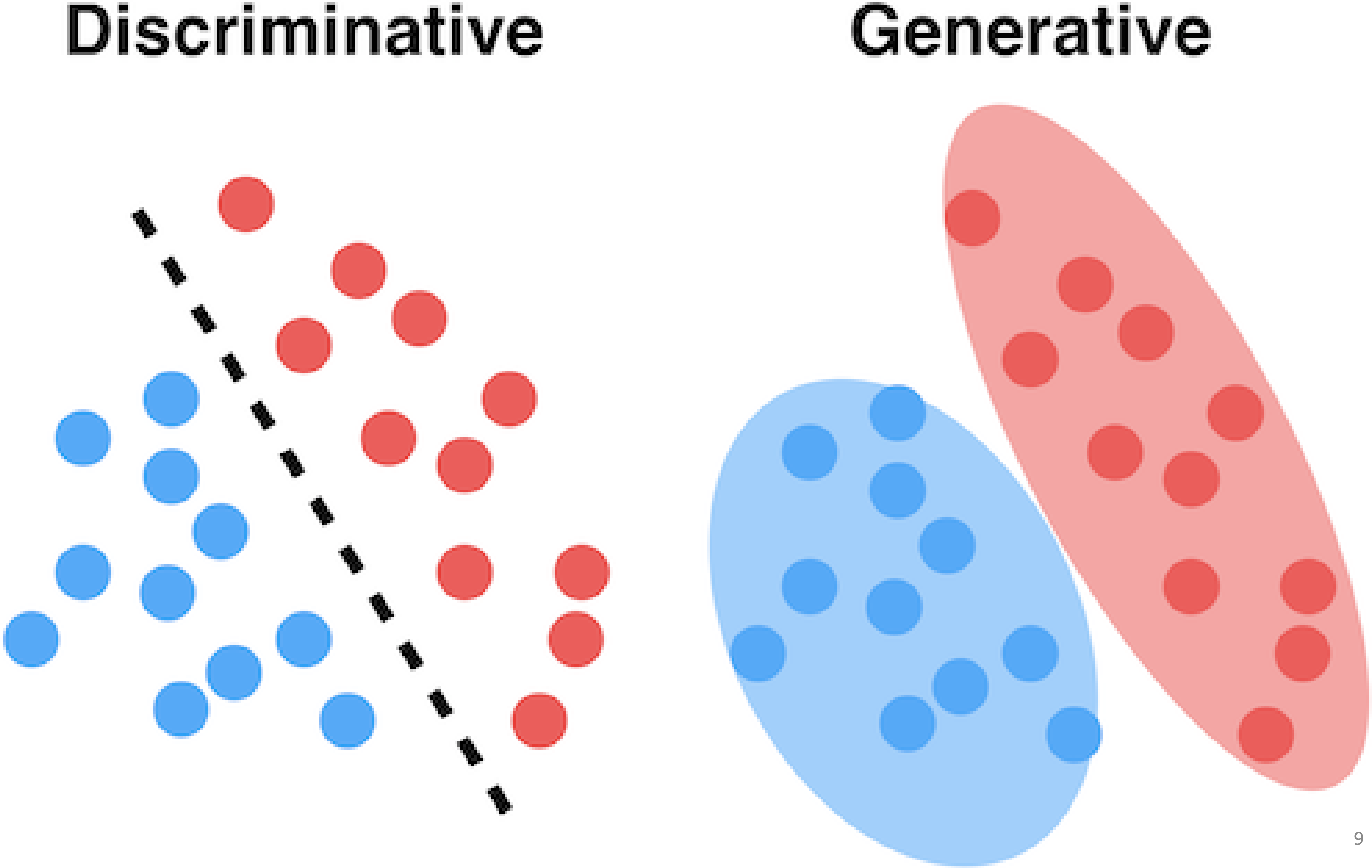
- Recover the data distribution [essence of data science]
- Benefit from hidden variables modeling
- E.g. Naive Bayes, Hidden Markov Model, Mixture Gaussian, Markov Random Fields, Latent Dirichlet Allocation

Discriminative Models vs Generative Models

- In General
  - A Discriminative model models the **decision boundary between the classes**
  - A Generative Model explicitly models the **actual distribution of each class**

- Example: Our training set is a bag of fruits. Only apples and oranges Each labeled. Imagine a post-it note stuck to the fruit
  - A generative model will model various attributes of fruits such as color, weight, shape, etc
  - A discriminative model might model color alone, **should that suffice** to distinguish apples from oranges

	Discriminative model	**Generative model**
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes



Discriminative

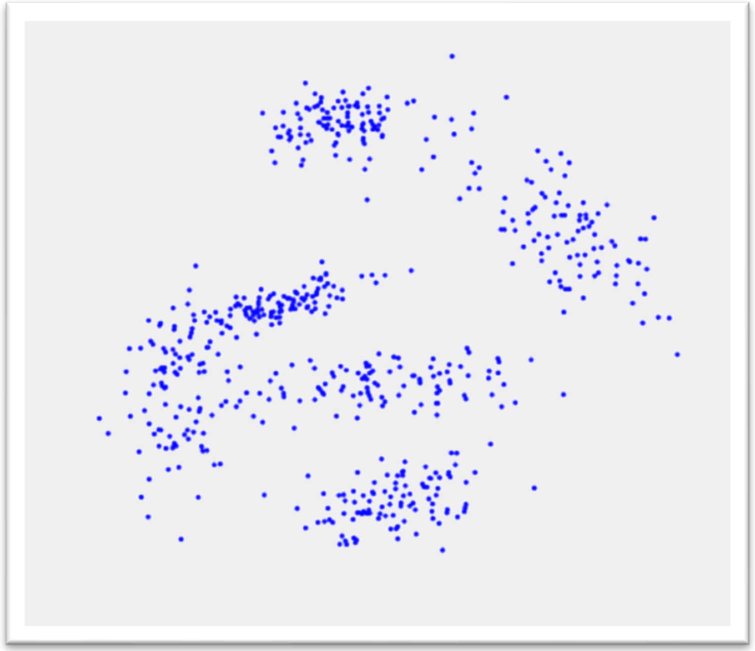
Generative

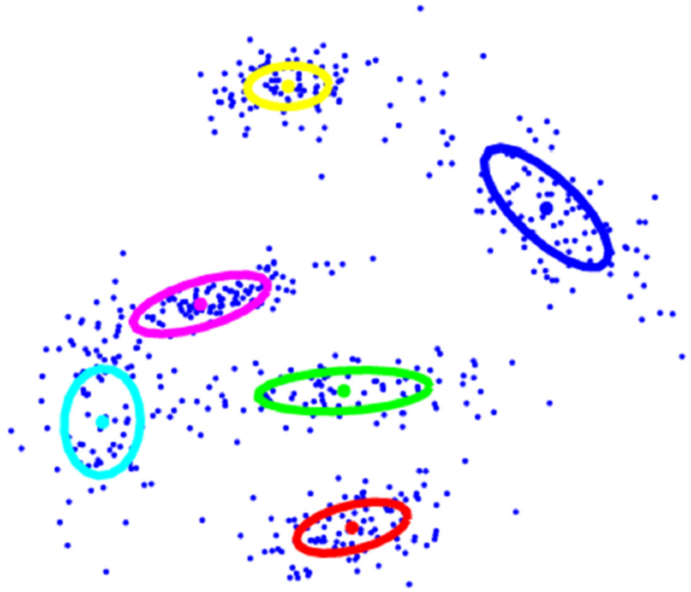
Gaussian Mixture Models

Gaussian Mixture Models

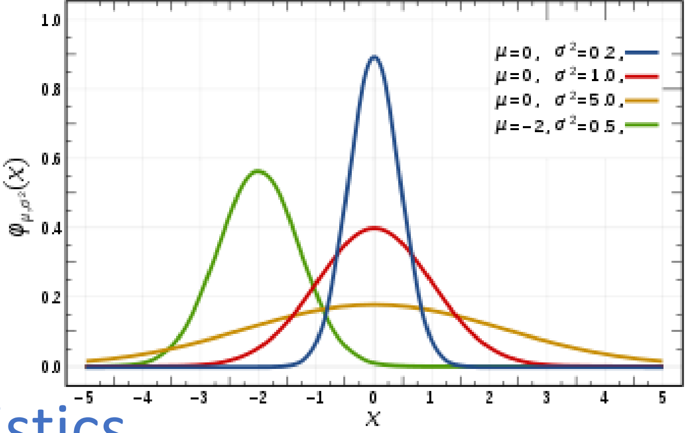
- Is a clustering algorithms

- Difference with K-means
  - K-means outputs the label of a sample
  - GMM outputs the probability that a sample belongs to a certain class
  - GMM can also be used to generate new samples!





Gaussian distribution



- Very common in probability theory and important in statistics
- often used in the natural and social sciences to represent real-valued random variables whose distributions are not known
- is useful because of the central limit theorem
  - averages of samples independently drawn from the same distribution converge in distribution to the normal with the true mean and variance, that is, they become normally distributed when the number of observations is sufficiently large
- Physical quantities that are expected to be the sum of many independent processes often have distributions that are nearly normal
- The probability density of the Gaussian distribution is

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

High-dimensional Gaussian distribution

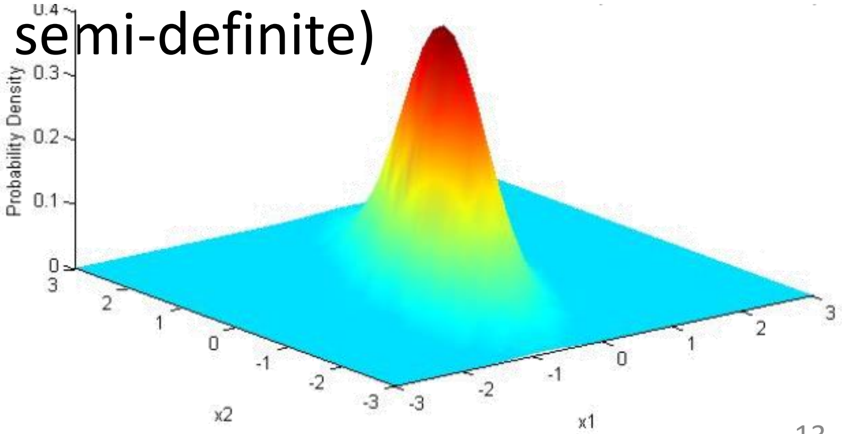
- The probability density of Gaussian distribution on $x = (x_1, \dots, x_d)^\top$ is

$$\mathcal{N}(x|\mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^d |\Sigma|}}$$

  - where μ is the mean vector
  - Σ is the symmetric covariance matrix (positive semi-definite)

- E.g. the Gaussian distribution with

$$\mu = (0,0)^T \qquad \Sigma = \begin{pmatrix} 0.25 & 0.30 \\ 0.30 & 1.00 \end{pmatrix}$$



Mixture of Gaussian

- The probability given in a mixture of *K* Gaussians is:

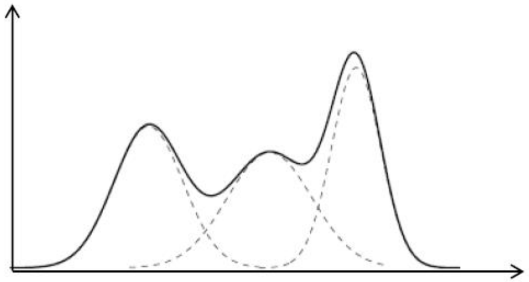
$$p(x) = \sum_{j=1}^{K} w_j \cdot N(x \mid \mu_j, \Sigma_j)$$

where w_j is the prior probability of the j-th Gaussian

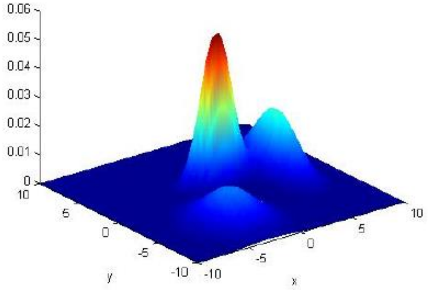
$$\sum_{j=1}^{K} w_j = 1 \qquad \text{and} \qquad 0 \le w_j \le 1$$

- Example

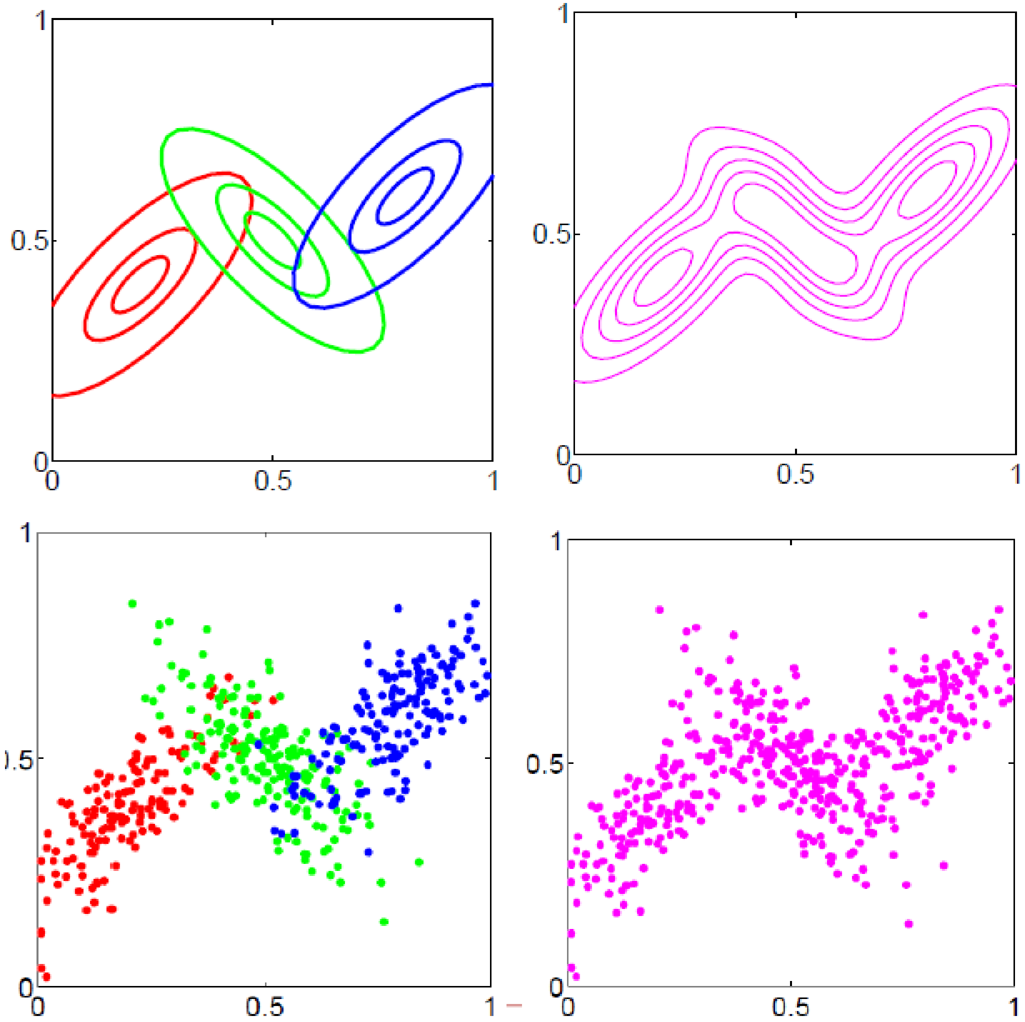
d=1:

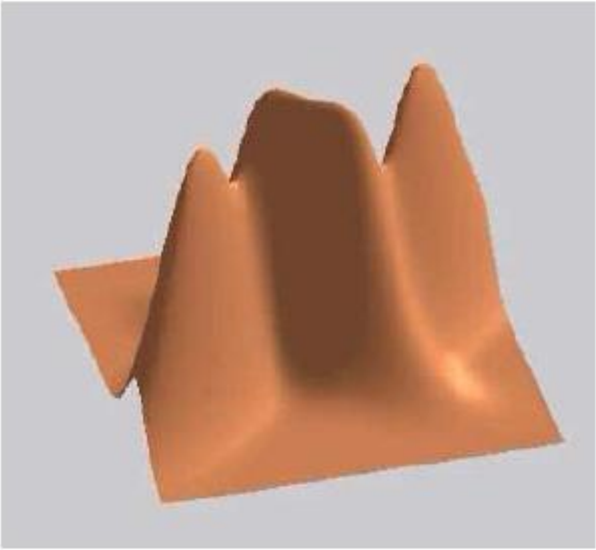


d=2:



Examples





Data generation

- Let the parameter set $\theta = \{w_j, \mu_j, \Sigma_j\}$, then the probability density of mixture Gaussian can be written as

$$p(x|\theta) = \sum_{j=1}^{K} w_j \cdot \mathcal{N}(x|\mu_j, \Sigma_j)$$

- Equivalent to generate data points in two steps
  - Select which component j the data point belongs to according to the categorical (or multinoulli) distribution of $(w_1, \ldots, w_K)$
  - Generate the data point according to the PMF of j-th component

Learning task

- Given a dataset $X = \{x_1, x_2, \cdots, x_N\}$ to train the GMM model

- Find the best θ that the maximize the probability $p(X|\theta)$
- Maximal likelihood estimator (MLE)

$$\boxed{\theta^* = \arg\max_\theta p(X \mid \theta) = \arg\max_\theta \prod_{i=1}^{N} p(x_i \mid \theta)}$$