Statistical Pattern Recognition

- Review

Statistical Pattern Recognition – Review

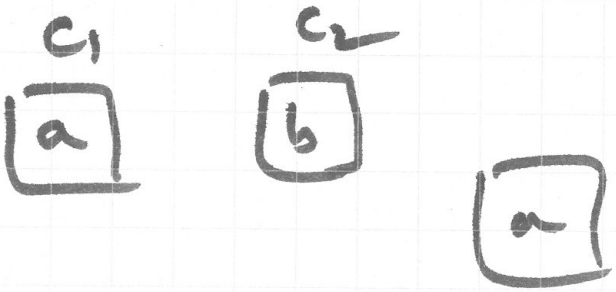
- Illustration of a classification problem
  [a] & [b]
- Classification Vs Regression
- Preprocessing & Feature extraction
- Curse of dimensionality
- Function approximation
  - Polynomial curve fitting
  - Degrees of Freedom (# weights)
  - Bias & Variance, Regularization

Bayes's Theorem : Probabilistic formulation for the classification task

Decision Making

Probability Density Estimation

- Parametric Models
  (Maximum likelihood & Bayesian learning)
- Non parametric
  * Histograms
  * Kernel - based Methods
  * k - nearest neighbour



$256 \times 256 \times 8 \longrightarrow$

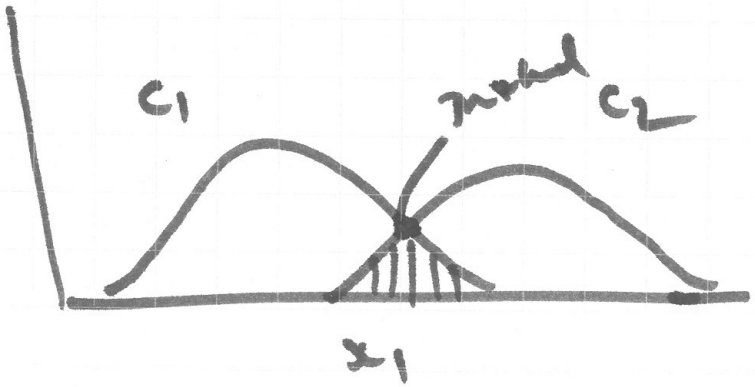
$256 \times 256 \times 8$

2

$15 \text{ } 000$

≈ 10

$x_1 \rightarrow$ height / width $\rightarrow$



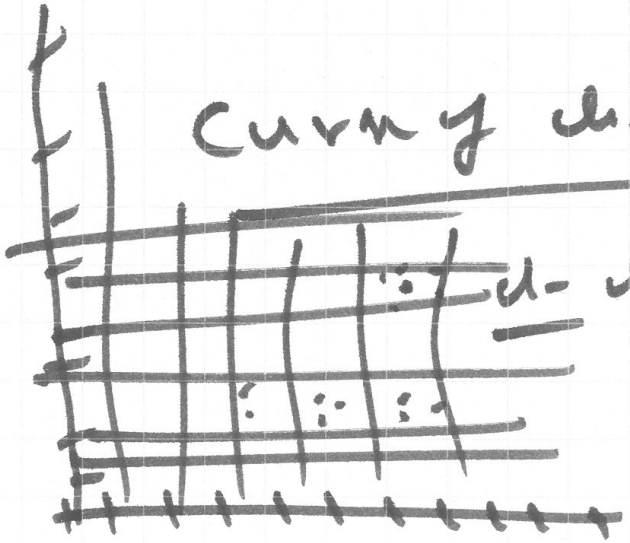
Classification vs Regression

Preprocessing ✓

a

b

Curse of dimensionality ✓



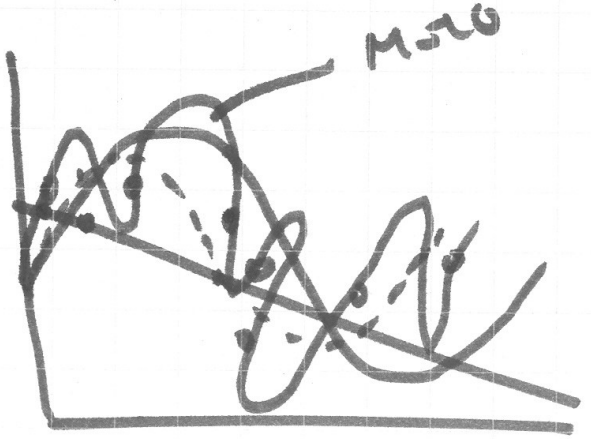
MP →

Curve fitting : Polynomial Model

$$y(x) = w_0 + w_1 x + \cdots + w_M x^M$$

$$= \sum_{j=0}^{M} w_j x^j$$

$y(x) = \sin(0.2\pi) + 0.5 \quad M = 0$



$$P(C_k/x_L) = \frac{P(x_L/C_k) \cdot P(C_k)}{P(x_L)} = 1$$

posterior

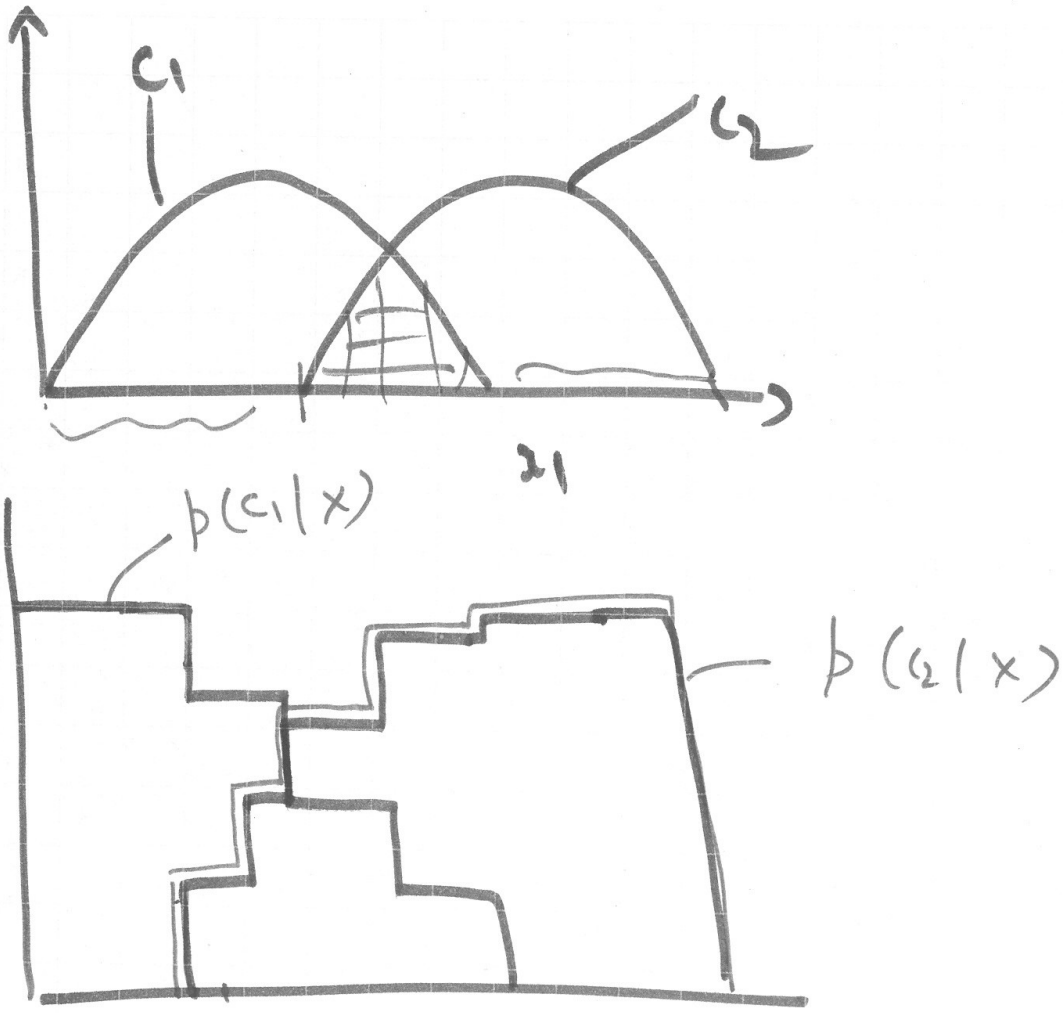
class conditional prob

unconditional prob

prior

$$\sum_k P(C_k/x_L) = 1$$

$$\sum_k P(x_L/C_k)\, P(C_k) = P(x_L)$$

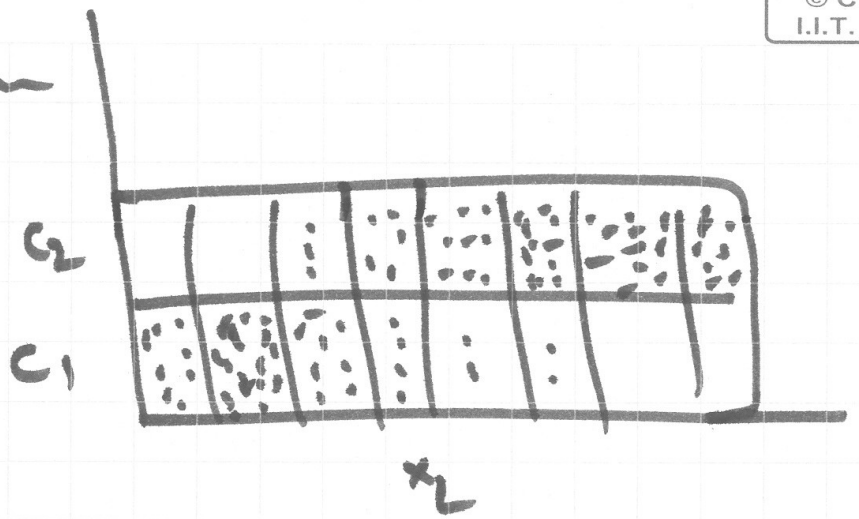


Buyer's Theorem

$N = 100$

$c_1 = 70$

$c_2 = 30$



$$P(c_k x_L) = \frac{4}{100}$$

$$P(c_k | x_L)\, P(x_L) \;;\quad P(x_L | c_k)\, P(c_k)$$

$$\frac{P(c_k | x_L)\, P(x_L)}{\frac{4}{9} \checkmark \quad \frac{1}{110} = \frac{4}{100}} \;\Bigg|\; \frac{4}{20} \cdot \frac{7}{10} = \frac{4}{100}$$

Decision Rules

$$P(C_k|x) > P(C_j|x) \quad k \neq j$$

$$P(x|C_k)\,P(C_k) > P(x|C_j)\,P(C_j)$$

$$P(e) = \oint P(x \in R_2, C_1) + P(x \in R_1, C_2)$$

$$= \int_{R_2} P(C_1|x)\,P(C_1) + \int_{R_1} P(C_2|x)\,P(C_2)$$

Probability Density Estimation

Parametric Methods

- Maximum likelyhood
- Bayesian inference

Non-parametric Methods

- Histogram
- Kernel based approaches
- K-nearest neighbour ✓

K-L distance (Kullback - Leibler distance) ✓

Mixture Models :

Discription of the Models

Training Methods

- ✓ — Nonlinear optimization
- ✓ — Re-estimation [EM-Algorithm]
- ✓ — Stochastic sequential estimation

Parametric Methods

Multivariate Gaussian

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-M)^T \Sigma^{-1} (x-M)\right]$$

Discriminant function

$$g_k(x) = \ln p(x|C_k) + \ln p(C_k)$$

~~Maximum like~~

$$p(C_k|x) = \frac{p(x|C_k)\, p(C_k)}{p(x)}$$

Maximum Likelihood

$$p(x/\theta) ; \quad x = \{x^1, x^2, \cdots x^N\}$$

$$\theta = \{\theta_1, \theta_2 \cdots \theta_M\}$$

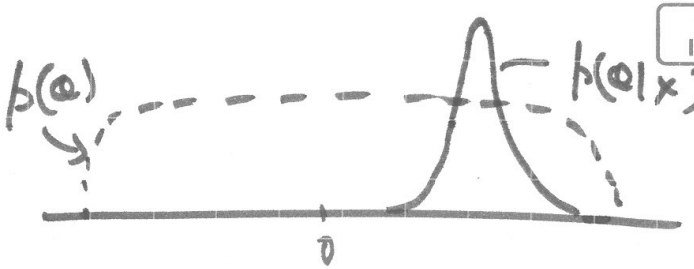
$$p(x/\theta) = \prod_{n=1}^{N} p(x^n/\theta) = L(\theta)$$

$$E = -\ln L(\theta) = -\sum_{n=1}^{N} \ln p(x^n/\theta)$$

$$\frac{\partial E}{\partial \theta} = 0 \Rightarrow \text{optimal parameters}$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x^n \; ; \quad \hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(x^n - \hat{\mu})(x^n - \hat{\mu})^T$$

Bayesian Inference



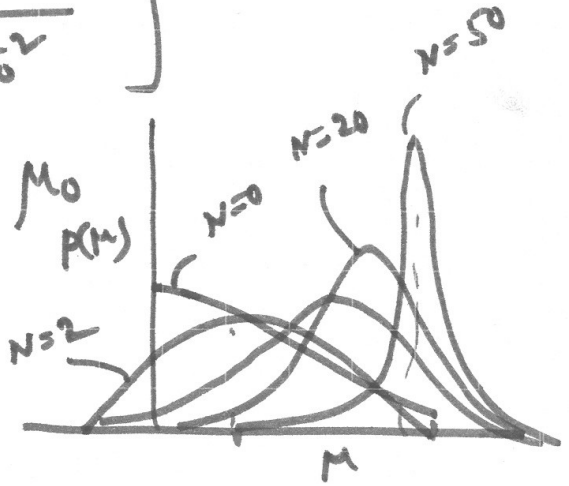
$$p(x|\theta) = \prod_{m=1}^{N} p(x^m|\theta)$$

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} = \frac{p(\theta)}{p(x)} \prod_{m=1}^{N} p(x^m|\theta)$$

$$p_0(\mu) = \frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

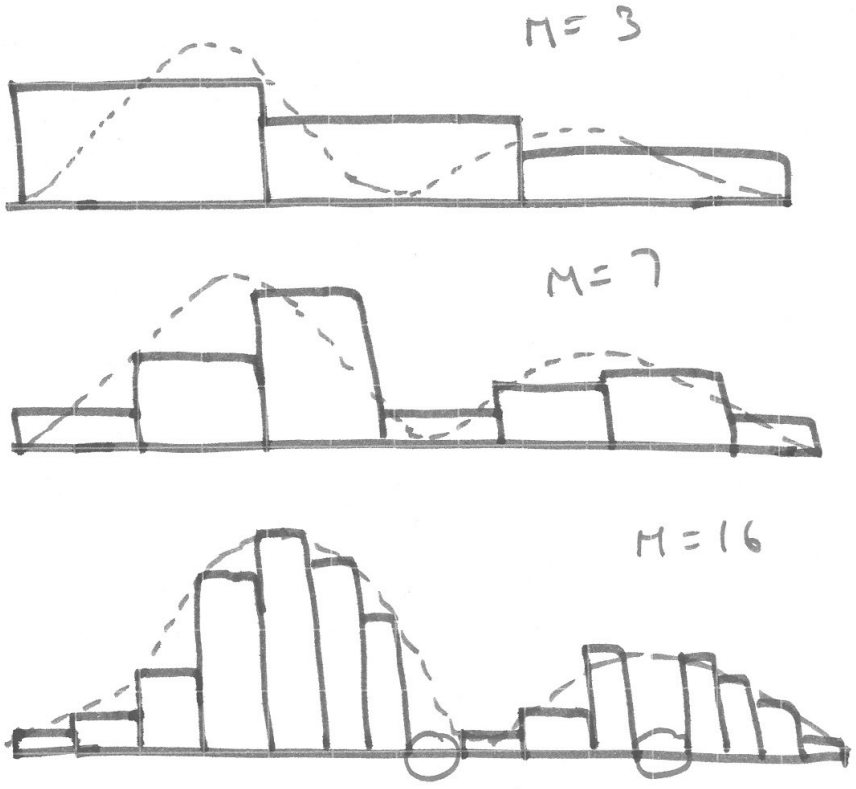
$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{x} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$



Nonparametric Methods

Histograms



Smoothing factor
= width of the bin

Disadv

- Discontinuous ✓
- Curse of dimensionality

Adv ... Can be constructed sequentially ✓

#bins < low (smooth out the crucial details)
high (spiky & noisy) ✓

Density Estimation (in General)

$$P = \int_R p(x)\,dx \;;\quad P_r(K) = \frac{N!}{K!\,(N-K)!}\,P^K (1-P)^{N-K}$$

$$\text{mean } P = \frac{K}{N} \;;\quad P = \int_R p(x)\,dx \simeq p(x)\,V$$

$$p(x) = \frac{K}{NV}$$

Kernel-based Methods

$$V = h^d$$

$$K = \sum_{n=1}^{N} H\left(\frac{x - x^n}{h}\right) \;;\quad \tilde{p}(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^d} H\left(\frac{x - x^n}{h}\right)$$

Multivariate Normal kernel

$$\tilde{p}(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{\frac{d}{2}}} \exp\left[\frac{-\|x - x^n\|^2}{2h^2}\right]$$

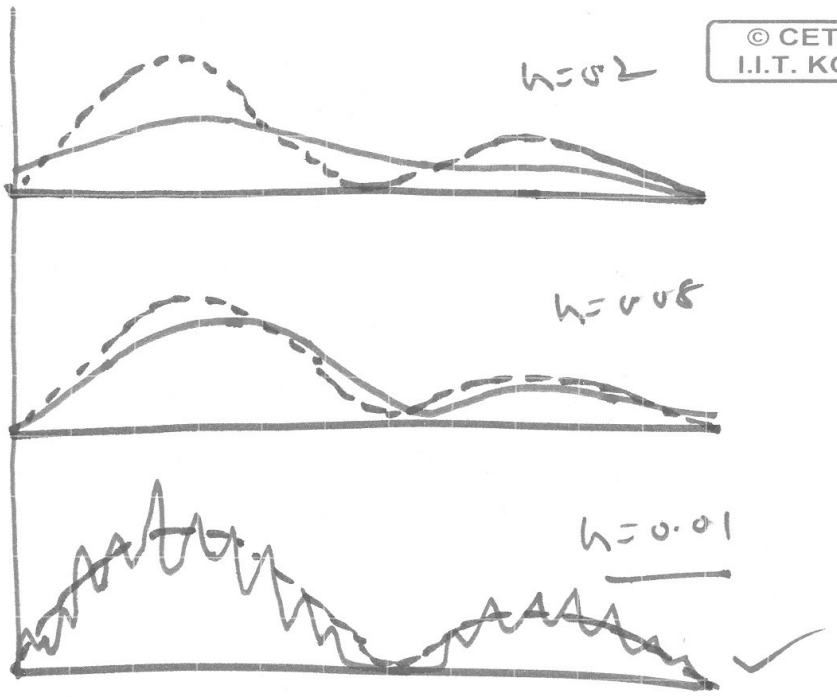
$$\tilde{p}(x) = h(x) * p(x)$$

$h(x)$ = impulse fn.

$\tilde{p}(x) \approx p(x)$

$h \to 0 \quad \tilde{p}(x) = p(x)$

∵ due to finite points $\tilde{p}(x)$ is noisy



K - nearest neighbour

Total data points $= N$

points of class $k = N_k$

$$\sum_k N_k = N$$

$$p(x|c_k) = \frac{k_k}{N_k V}$$

$$p(x) = \frac{k}{NV}$$

$$p(c_k) = \frac{N_k}{N}$$

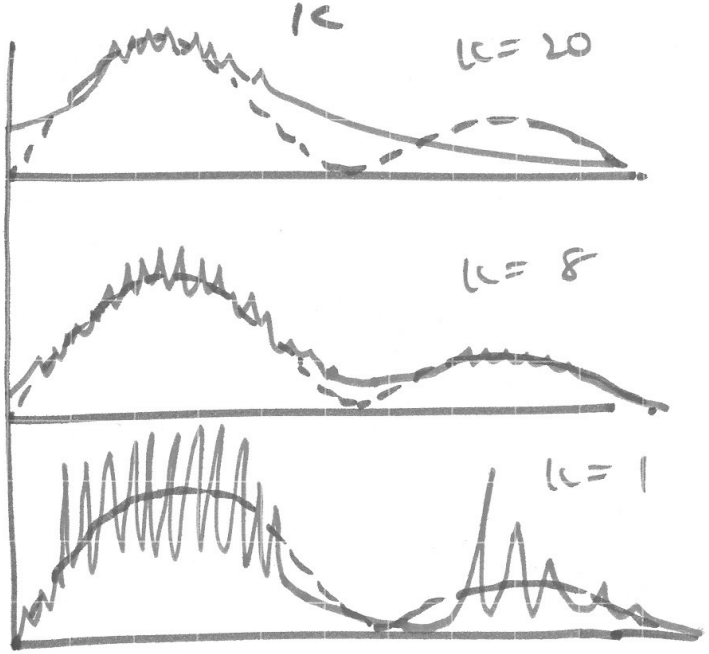
$$P(C_k|x) = \frac{p(x|C_k)\,p(C_k)}{p(x)} = \frac{|C_k|}{k}$$

Smoothing parameters

Histogram — width of the bins

Kernel — width of the kernel 'h'

k-nearest neighbours — value of 'k'



$$L = -\int \tilde{p}(x) \ln \frac{\tilde{p}(x)}{p(x)}\, dx$$

k-L distance

Kullback-Leibler distance

$\Rightarrow$ Distance between true and the estimated distance

GMM : EM Algorithm

Mixture Models

Parametric <
- Specific functional form is chosen ✓
- Evaluation with new data is faster ✓

Non-parametric <
- No specific functional form ✓
- # variables grow with # training points
- very slow for evaluation of new data

✓ **Semi-parametric** <
- No specific functional form
- # variables grows with complexity
- Computationally easier case ✓

Training methods —
- Non-linear optimization
- Re-estimation [EM algorithm]
- Stochastic sequential estimation

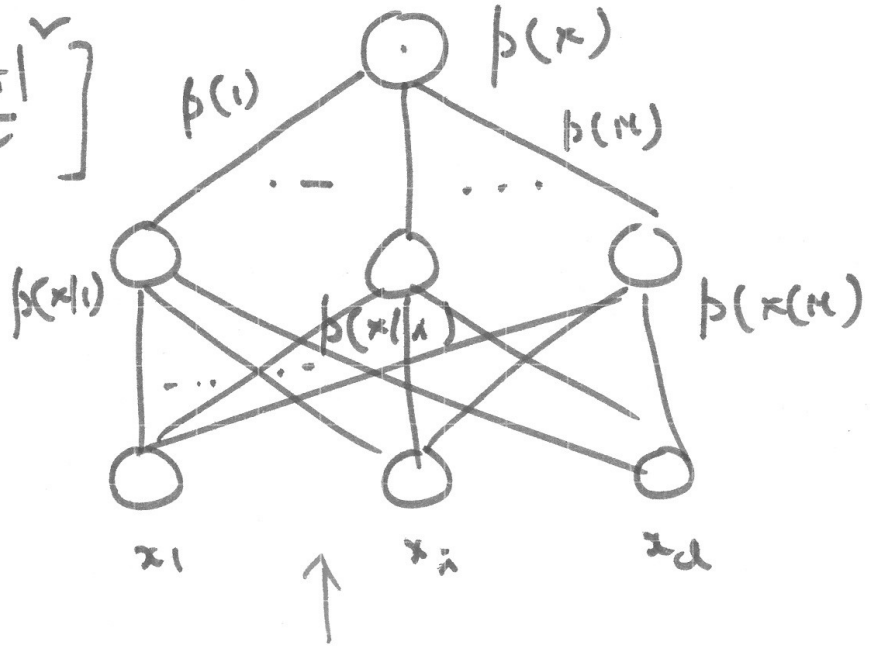
Mixture Models (cont..)

$$p(x) = \sum_{j=1}^{M} p(x|j)\, p(j)$$

- M: Mixture diff models
- $p(x|j)$: Component density
- $p(j)$: Mixing parameters
- Linear combination of component densities

$$p(x|j) = \frac{1}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left[-\frac{|x-\mu_j|^2}{2\sigma_j^2}\right]$$

$p(x|c_k)$ — Class specific Mix Model



Maximum likelihood formulation (Mixture model)

$$E = -\ln L = -\sum_{n=1}^{N} \ln p(x^n) = -\sum_{n=1}^{N} \ln \left[\sum_{J=1}^{M} p(x^n/J)\, p(J) \right]$$

$$= -\sum_{n=1}^{N} \ln \left[\sum_{J=1}^{M} p(x^n/J)\; p(J) \right]$$

Complexity in minimizing the error function (E)

- Singularity problems
- Avoiding local minima

M ✓

$$\frac{\partial E}{\partial \mu_j} = 0 \; ; \quad \frac{\partial E}{\partial \sigma_j} = 0 \; ; \quad \frac{\partial E}{\partial p_j} = 0 \; ;$$

$$\hat{\mu}_j = \sum_n P(j|x^n)\, x^n \Big/ \sum_n P(j|x^n)$$

$$\hat{\sigma}_j^2 = \frac{1}{d} \frac{\sum_n P(j|x^n) \, \| x^n - \hat{\mu}_j \|^2}{\sum_n P(j|x^n)}$$

$$\hat{P}(j) = \frac{1}{N} \sum_n P(j|x^n)$$

The EM Algorithm

$\hat{M}_T$, $\hat{\sigma}_T^2$, $\hat{P}(T)$ → do not provide direct soln

Nonlinear coupled eqs

Soln :- Iterative scheme to find min of E'

1. Initial guess of the parameters ⇒ old parameters
2. Evaluate RHS of $\hat{M}_T$, $\hat{\sigma}_T^2$, $\hat{P}(T)$
3. Re estimate of the parameters ⇒ new parameters
4. Verify the new E' with new parameters

$$E^{new} - E^{old} = -\sum_n \ln\left[\frac{p^{new}(x^n)}{p^{old}(x^n)}\right]$$

$$= -\sum_n \ln\left[\frac{\sum_J p^{new}(J)\, p^{new}(x^n|J)}{p^{old}(x^n)} \cdot \frac{p^{old}(J(x^n))}{p^{old}(J(x^n))}\right]$$

$$\leq -\sum_n \sum_J p^{old}(J(x^n)) \ln\left[\frac{p^{new}(J)\, p^{new}(x^n|J)}{p^{old}(x^n)\, p^{old}(x^n|J)}\right]$$

$$E^{new} \leq E^{old} + Q$$

$$Q = -\sum_n \sum_J p^{old}(J(x^n) \ln\left[p^{new}(J)\, p^{new}(J(x^n))\right]$$

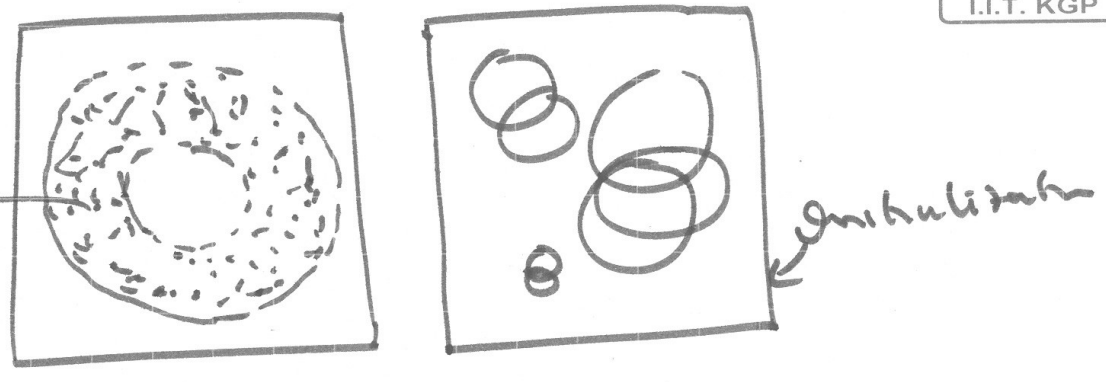
$$Q = -\sum_n \sum_J p^{old}(J(x^n)\left[\ln p^{new}(J) - d \ln \sigma_J^{new} - \frac{\|x^n - M_J^{new}\|^2}{2(\sigma_J^{new})^2}\right] + \text{Gaussian}$$

$$\mu_j^{new} = \frac{\sum_x p^{old}(j|x^x)\, x^x}{\sum_x p^{old}(j|x^x)} \checkmark$$

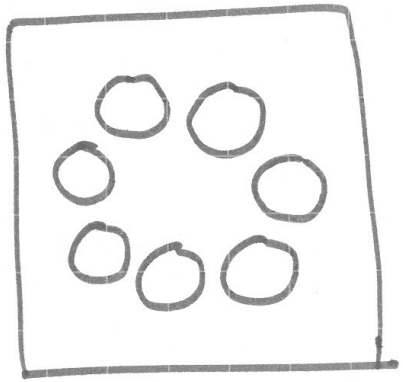
$$\left(\sigma_j^{new}\right)^2 = \frac{1}{d} \frac{\sum_x p^{old}(j|x^x)\, \|x^x - \mu_j^{new}\|^2}{\sum_x p^{old}(j|x^x)} \checkmark$$

$$P^{new}(j) = \frac{1}{N} \sum_x p^{old}(j|x^x) \checkmark$$

Example



After 20 Iterations



GMM VS clustering

Incomplete data

hypothetical complete data, x^m

$z^m \longrightarrow (1, M)$ ✓

$$E^{Comp} = \cancel{\ln} - \sum_n \ln p^{new}(x^n, t^n)$$

$$= - \sum_n \ln \left[p^{new}(t^n)\, p^{new}(x^n | t^n) \right] \checkmark$$

1. Given some values to the parameters of GMM ✓
2. Find PD of t^n ✓ using Bayes's theorem & old parameters
3. Compute $E[E^{Comp}]$ wrt PD of t^n ✓ — E step
4. Compute New parameters by minimizing E ✓

$$E\left[E^{Comp}\right] = \sum_{t^1=1}^{M} \cdots \sum_{t^N=1}^{M} E^{Comp} \prod_{n=1}^{N} p^{old}(t^n | x^n)$$

$$= - \sum_n \sum_J p^{old}(J | x^n) \ln \left[p^{new}(J)\, p^{new}(x^n | J) \right]$$