NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W14: 11/24/15

NC STATE UNIVERSITY

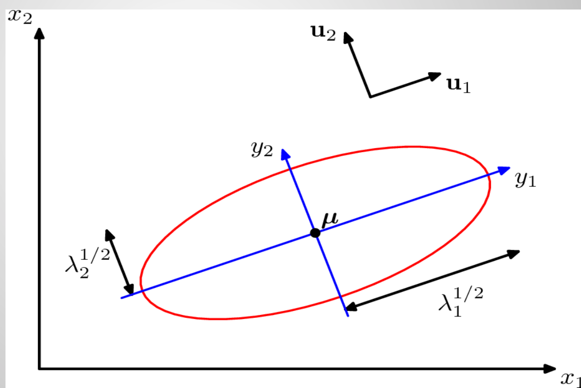
Today

- Mixture Models
- Expectation Maximization

NC STATE UNIVERSITY

Multivariate Gaussian

- Bivariate covariance matrix representation

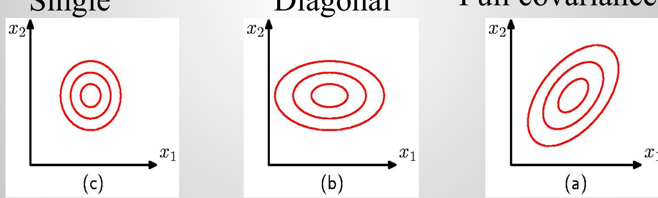


NC STATE UNIVERSITY

Multivariate Gaussian

- Σ plays key role, but accurate estimation requires large number samples (10xdim/class)
- However, Σ can be simplified

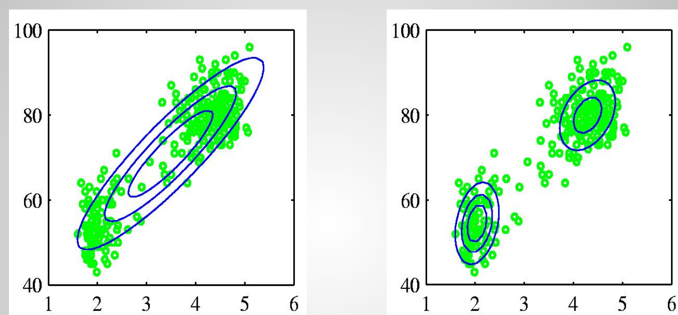
Single | Diagonal | Full covariance



$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^{\ 2} & 0 \\ 0 & \sigma_2^{\ 2} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11}^{\ 2} & \sigma_{12}^{\ 2} \\ \sigma_{12}^{\ 2} & \sigma_{22}^{\ 2} \end{pmatrix}$$

NC STATE UNIVERSITY

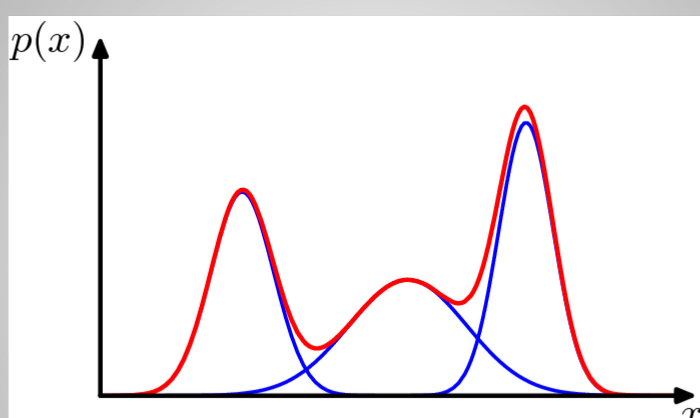
What If The Data Is Multimodal



- Real world data are rarely unimodal
- Many times we don't know the labels for all components

NC STATE UNIVERSITY

How to model such data?



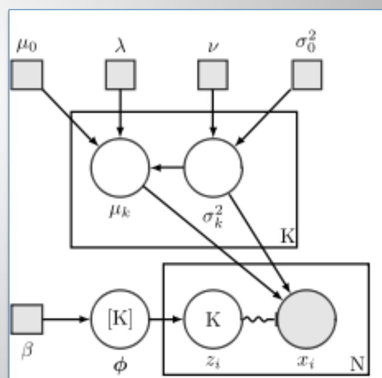
- Mixtures of Gaussians
- Mixture models in general

NC STATE UNIVERSITY

Mixture Models

- Mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs.

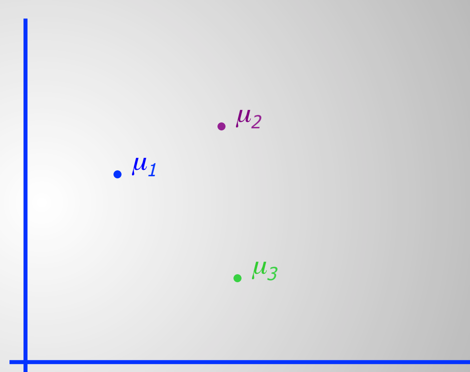
- K = # components
- N = # observations
- θ = parameter
- Φ = mixture weight
- z_i = component of observation i
- x_i = i^{th} observation



NC STATE UNIVERSITY

The GMM assumption

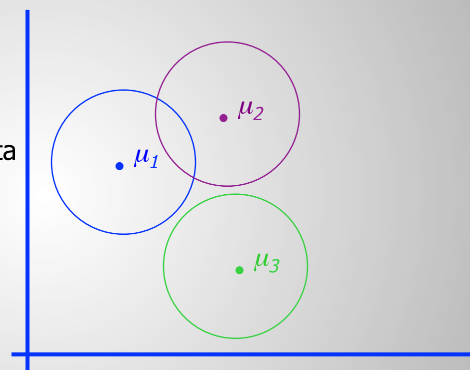
- There are k components. The i' th component is called ω_i
- Component ω_i has an associated mean vector μ_i



The GMM assumption

- There are k components. The i' th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 \boldsymbol{I}$

Assume that each datapoint is generated according to the following recipe:

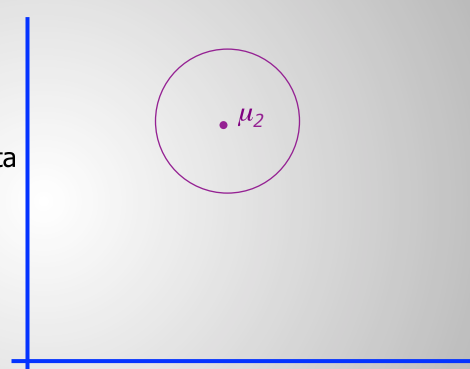


The GMM assumption

- There are k components. The i' th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 \boldsymbol{I}$

Assume that each datapoint is generated according to the following recipe:

1. Pick a component at random. Choose component i with probability $P(\omega_i)$.



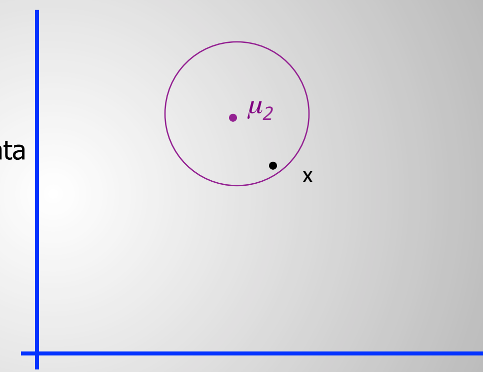
NC STATE UNIVERSITY

The GMM assumption

- There are k components. The i' th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2\mathbf{I}$

Assume that each datapoint is generated according to the following recipe:

1. Pick a component at random. Choose component i with probability $P(\omega_i)$.
2. Datapoint ~ N(μ_i, $\sigma^2\mathbf{I}$)



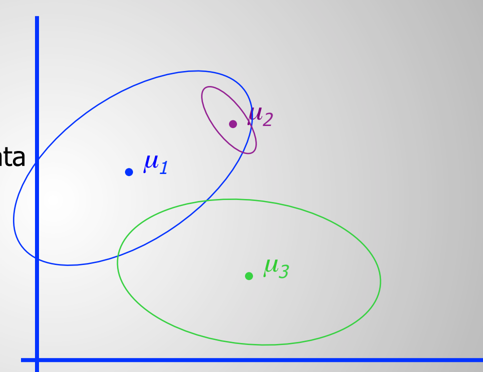
NC STATE UNIVERSITY

The General GMM assumption

- There are k components. The i' th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix Σ_i

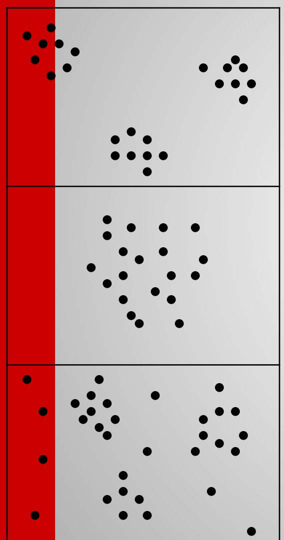
Assume that each datapoint is generated according to the following recipe:

1. Pick a component at random. Choose component i with probability $P(\omega_i)$.
2. Datapoint ~ N(μ_i, Σ_i)



NC STATE UNIVERSITY

Unsupervised Learning: not as hard as it looks



Sometimes easy

Sometimes impossible

and sometimes in between

IN CASE YOU'RE WONDERING WHAT THESE DIAGRAMS ARE, THEY SHOW 2-d UNLABELED DATA (X VECTORS) DISTRIBUTED IN 2-d SPACE. THE TOP ONE HAS THREE VERY CLEAR GAUSSIAN CENTERS

NC STATE UNIVERSITY

GMM Clustering With Given K

- GMM (K=M)

$$p(x \mid \theta) = \sum_{i=1}^{M} \alpha_i p_i(x \mid \theta_i)$$

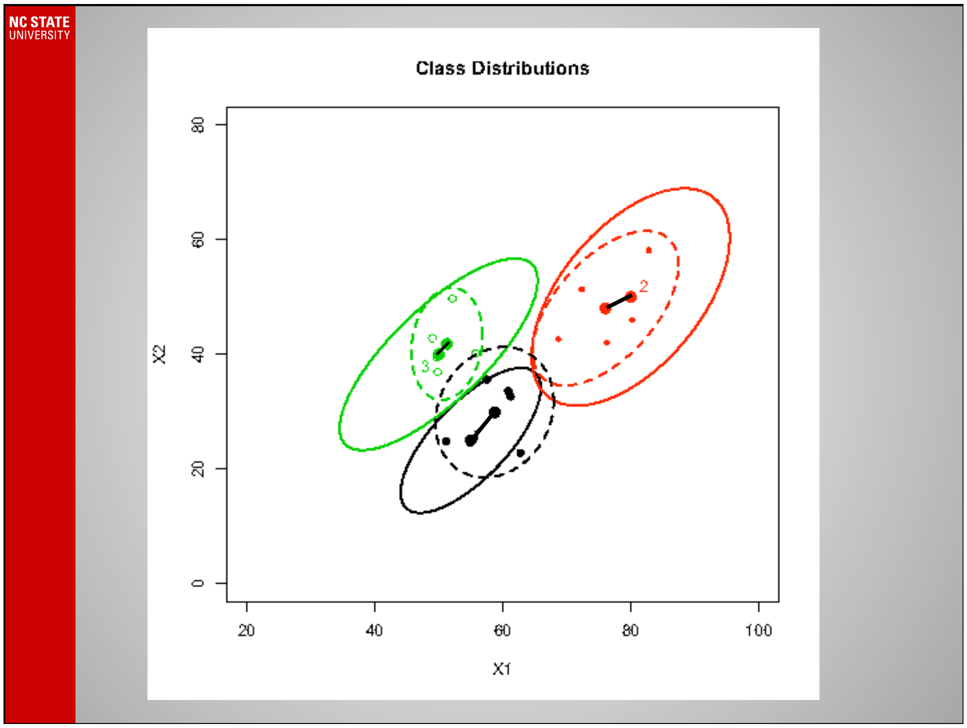
where $\theta = (\alpha_1, \ldots, \alpha_M; \theta_1, \ldots \theta_M)$

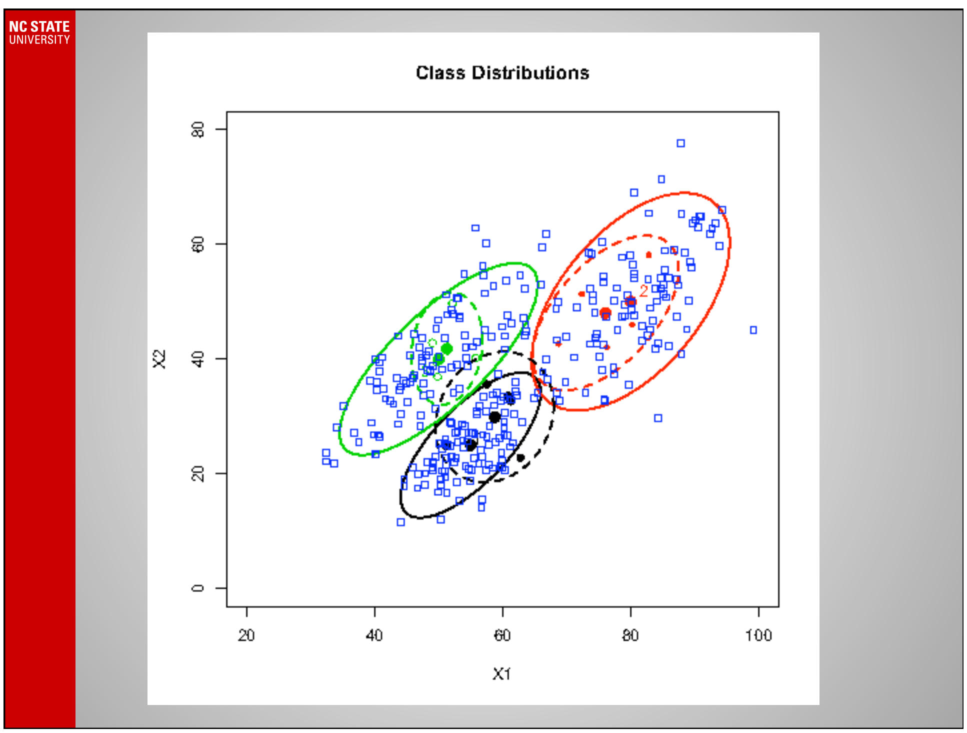
such that $\sum_1^M \alpha_i = 1, \quad 0 < \alpha_i < 1$ and

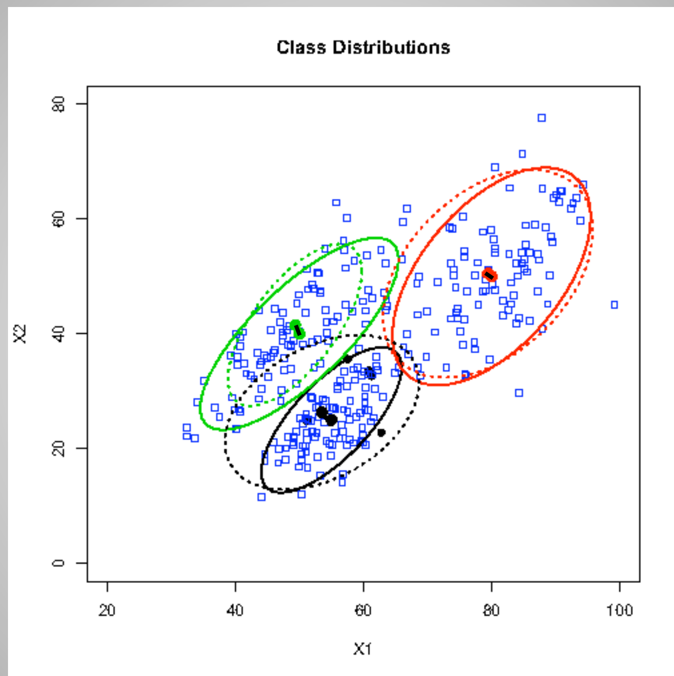
p_i pdf parameterized by θ_i

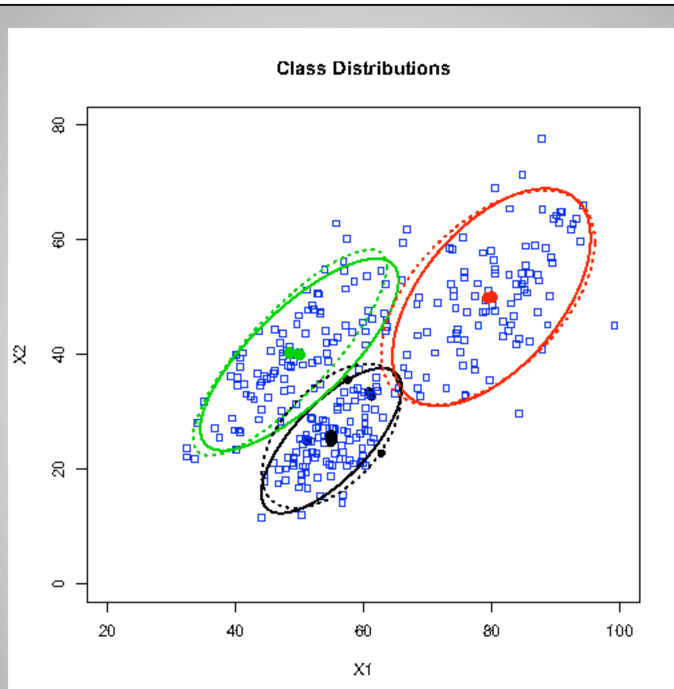
- Maximize

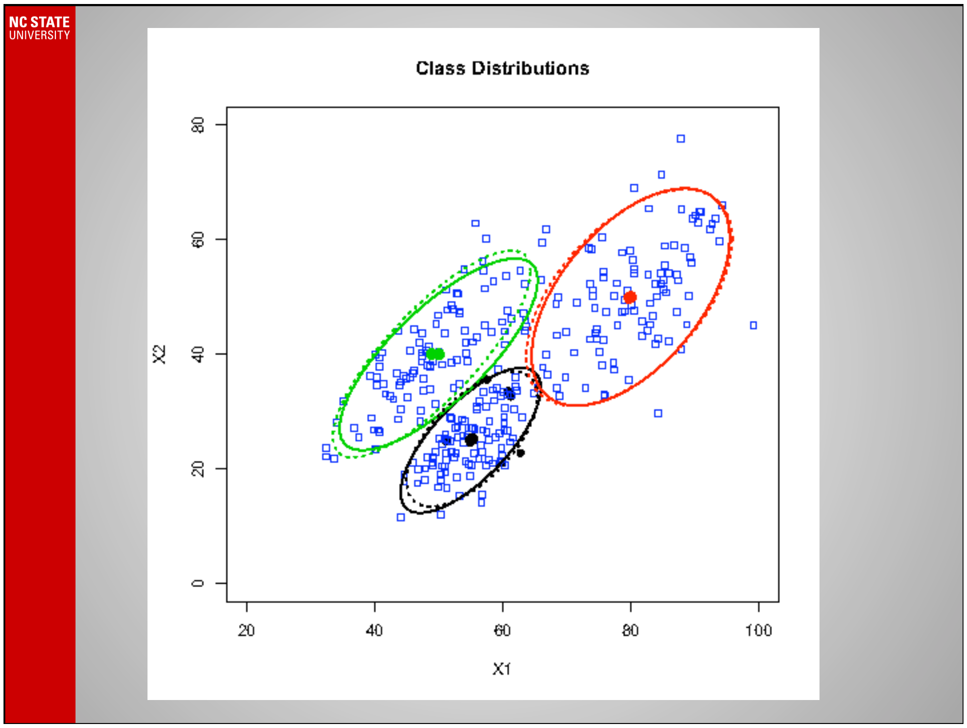
$$L(\theta) - L(\theta_i) = \ln \frac{\sum_z p(x \mid z, \theta) p(z \mid \theta)}{p(x \mid \theta_i)}$$

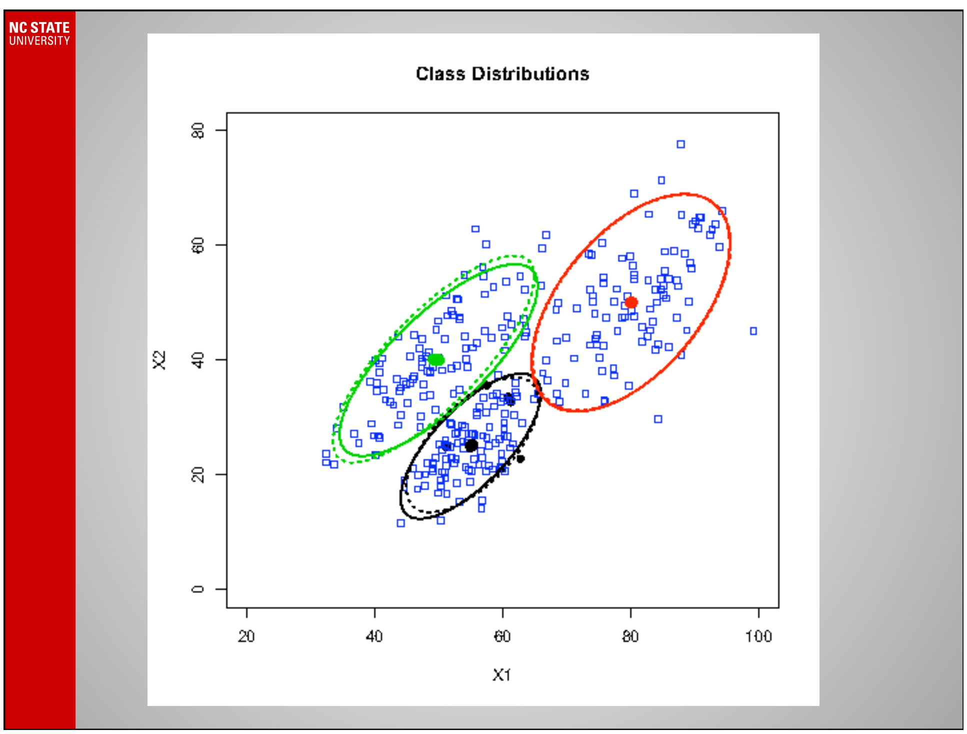

NC STATE UNIVERSITY
Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100
2
3


NC STATE UNIVERSITY
Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100
2


Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100


Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100


NC STATE UNIVERSITY
Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100


NC STATE UNIVERSITY
Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100

NC STATE UNIVERSITY

Semi-supervised Learning

EM to estimate GMM parameters

- E-Step

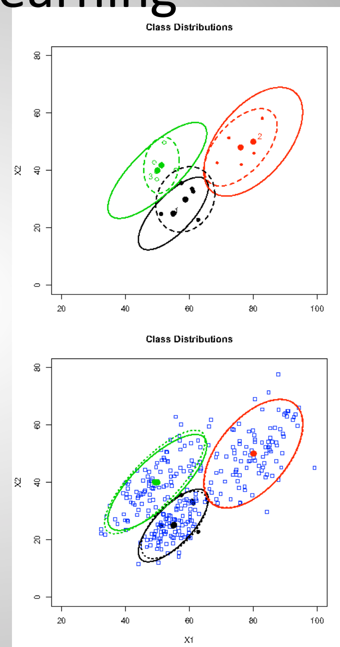
$$e_{ij} = \frac{\left|\hat{\Sigma}_j^k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_j^k\right)^T \hat{\Sigma}_j^{-1,k}\left(x_i - \hat{\mu}_j^k\right)\right\}}{\sum_{l=1}^{M}\left|\hat{\Sigma}_l^k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_l^k\right)^T \hat{\Sigma}_l^{-1,k}\left(x_i - \hat{\mu}_l^k\right)\right\}}$$

- M-Step

$$\alpha_j = \frac{\sum_{i=1}^{N} e_{ij}}{N}, \qquad \hat{\mu}_j^{k+1} = \frac{\sum_{i=1}^{N} e_{ij} x_i}{\sum_{i=1}^{N} e_{ij}},$$

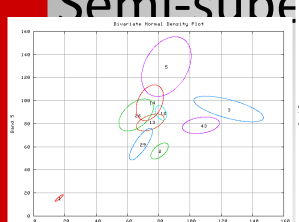
and $$\hat{\Sigma}_j^{k+1} = \frac{\sum_{i=1}^{N} e_{ij}\left(x_i - \hat{\mu}_j^{k+1}\right)\left(x_i - \hat{\mu}_j^{k+1}\right)^T}{\sum_{i=1}^{N} e_{ij}}$$

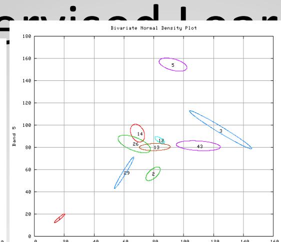
i^{th} data vector, j^{th} class

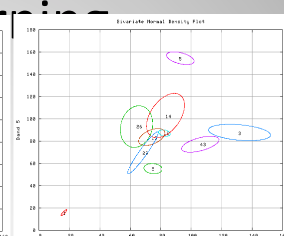


NC STATE UNIVERSITY

Semi-supervised Learning



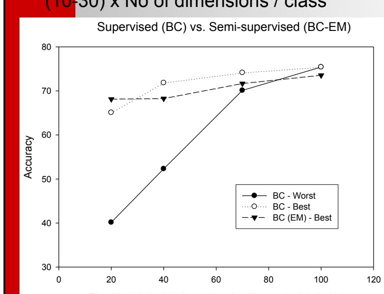




10 Classes, 100 Training Samples (10-30) x No of dimensions / class

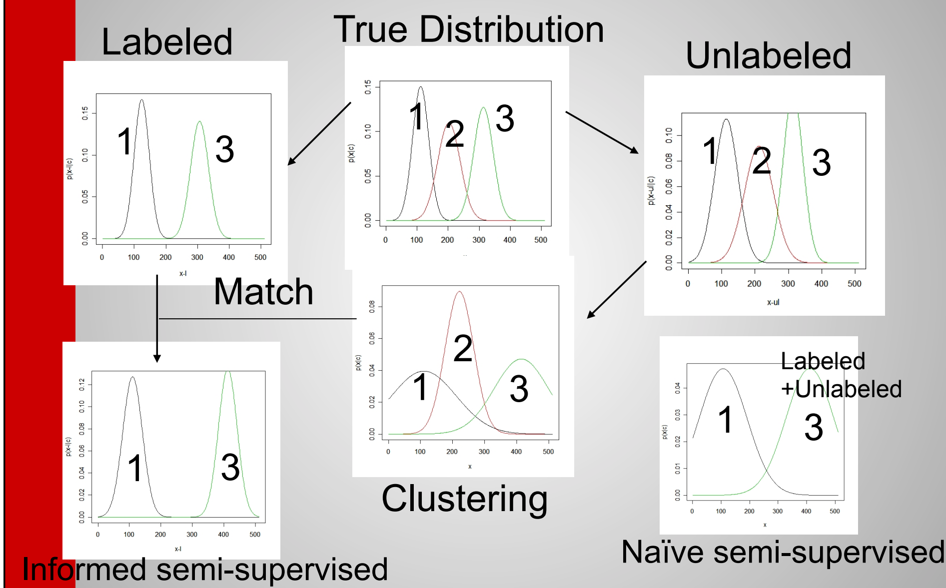
Small Subset of 20 Training Samples

20 labeled + 80 unlabeled samples



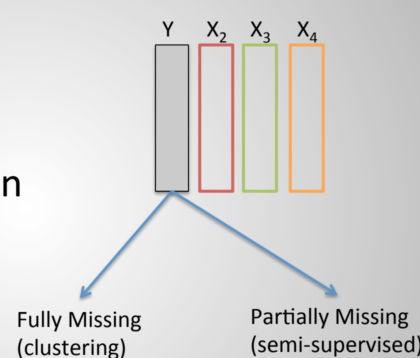
Ranga Raju Vatsavai, Shashi Shekhar, Thomas E. Burk: A Semi-Supervised Learning Method for Remote Sensing Data Mining. ICTAI 2005: 207-211

Informed Semi-supervised Learning



EM/GMM & Missing Data

- EM/GMM can be used to address Latent variable missing data pattern



NC STATE UNIVERSITY

Summary

	Categorical inputs only	Real-valued inputs only	Mixed Real / Cat okay
Inputs → Classifier → Predict category	Joint BC Naïve BC	Gauss BC	Dec Tree
Inputs → Density Estimator → Prob-ability	Joint DE Naïve DE	Gauss DE	
Inputs → Regressor → Predict real no.			

NC STATE UNIVERSITY

Acknowledgements

- Andrew Moore, CMU
- C. Fraley, A.E. Raftery Model-based clustering, discriminant analysis, and density estimation
- GX-Means: A model-based divide and merge algorithm for geospatial image clustering. R. Vatsavai, et. al.
- G. Hamerly, C. Elkan Learning the k in k-means, in: In Neural Information Processing Systems

11/25/15 © Raju Vatsavai CSC-591. 26