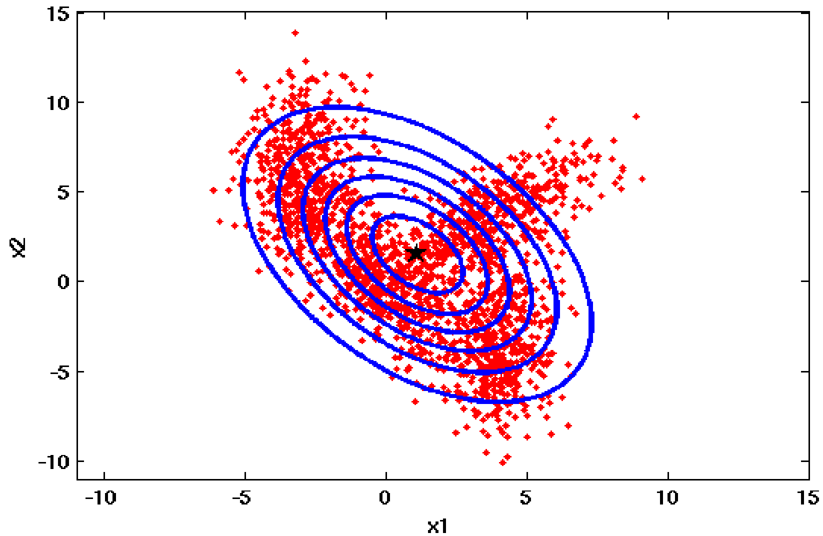
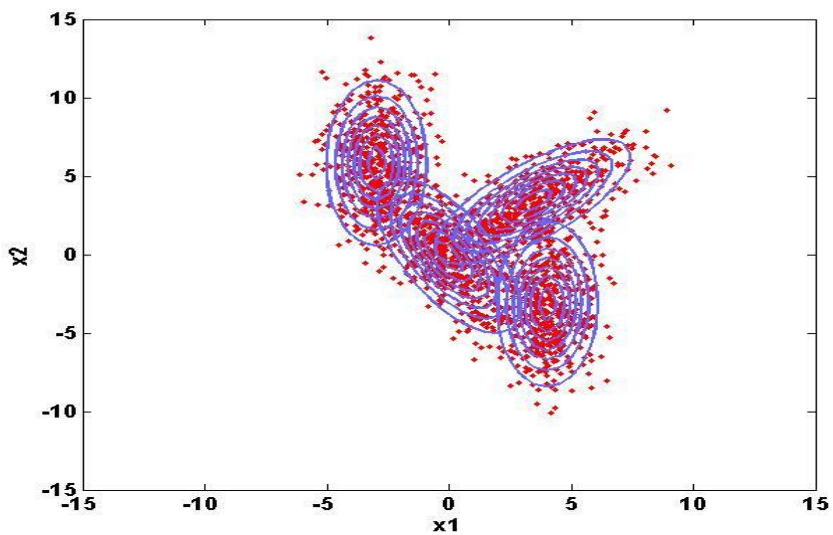
Gaussian Mixture Models

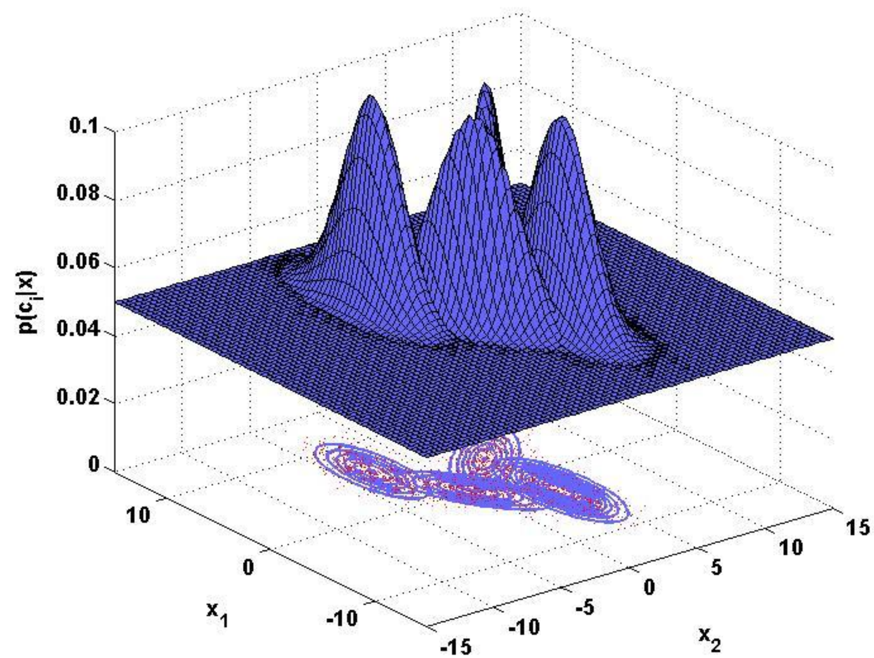
Multimodal Distribution

- For a class whose data is considered to have multiple clusters, the probability distribution is multimodal





Bivariate multimodal distribution



Gaussian Mixture Model

- Gaussian mixture model (GMM) can be used to represent a multimodal distribution

- GMM is a linear superposition of multiple Gaussians:

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

- For a d-dimensional feature vector representation of data, the parameters of a component in a GMM are
  - Mixture coefficient, π_k
  - d-dimension mean vector, $\boldsymbol{\mu}_k$
  - dxd size covariance matrix, $\boldsymbol{\Sigma}_k$

- Maximum likelihood method for training a GMM: Expectation-Maximization (EM) method

GMM – Continued …

- A quantity that plays an important role is the responsibility term, $\gamma(z_{nk})$
- It is given by

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{Q} \pi_j \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

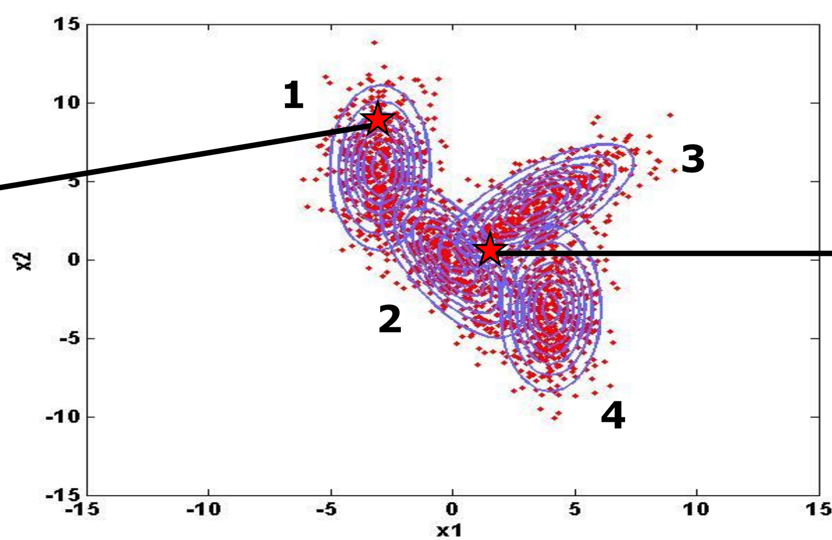
- π_k : prior probability of component q,
- $\gamma(z_{nk})$ gives the posterior probability of the component k for the observation $\mathbf{x}$

$\gamma_1(\mathbf{x}_i) = 0.99$

$\gamma_2(\mathbf{x}_i) = 0.01$

$\gamma_3(\mathbf{x}_i) = 0.00$

$\gamma_4(\mathbf{x}_i) = 0.00$



$\gamma_1(\mathbf{x}_j) = 0.08$

$\gamma_2(\mathbf{x}_j) = 0.42$

$\gamma_3(\mathbf{x}_j) = 0.34$

$\gamma_4(\mathbf{x}_j) = 0.16$

Parameter Estimation for GMMs

- Expectation-Maximization (EM) method: An elegant and powerful method for finding the maximum likelihood solution for a model with latent variables
- Total data log-likelihood:

$$\mathcal{L} = \ln p(\mathcal{D} \mid \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Setting the derivatives of $\mathcal{L}$ with respect to the means $\boldsymbol{\mu}_k$ to zero, we obtain:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

  - N_k: Effective number of points assigned to the component k

Parameter Estimation for GMMs – Continued …

- Setting the derivative of $\mathcal{L}$ with respect to the covariance matrices $\mathbf{\Sigma}_k$, we obtain:

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}$$

- Finally, maximize $\mathcal{L}$ with respect to the mixing coefficients π_k subject to the constraint:

$$\sum_{k=1}^{K} \pi_k = 1$$

- Maximization can be achieved using the Lagrange multiplier method to obtain

$$\pi_k = \frac{N_k}{N}$$

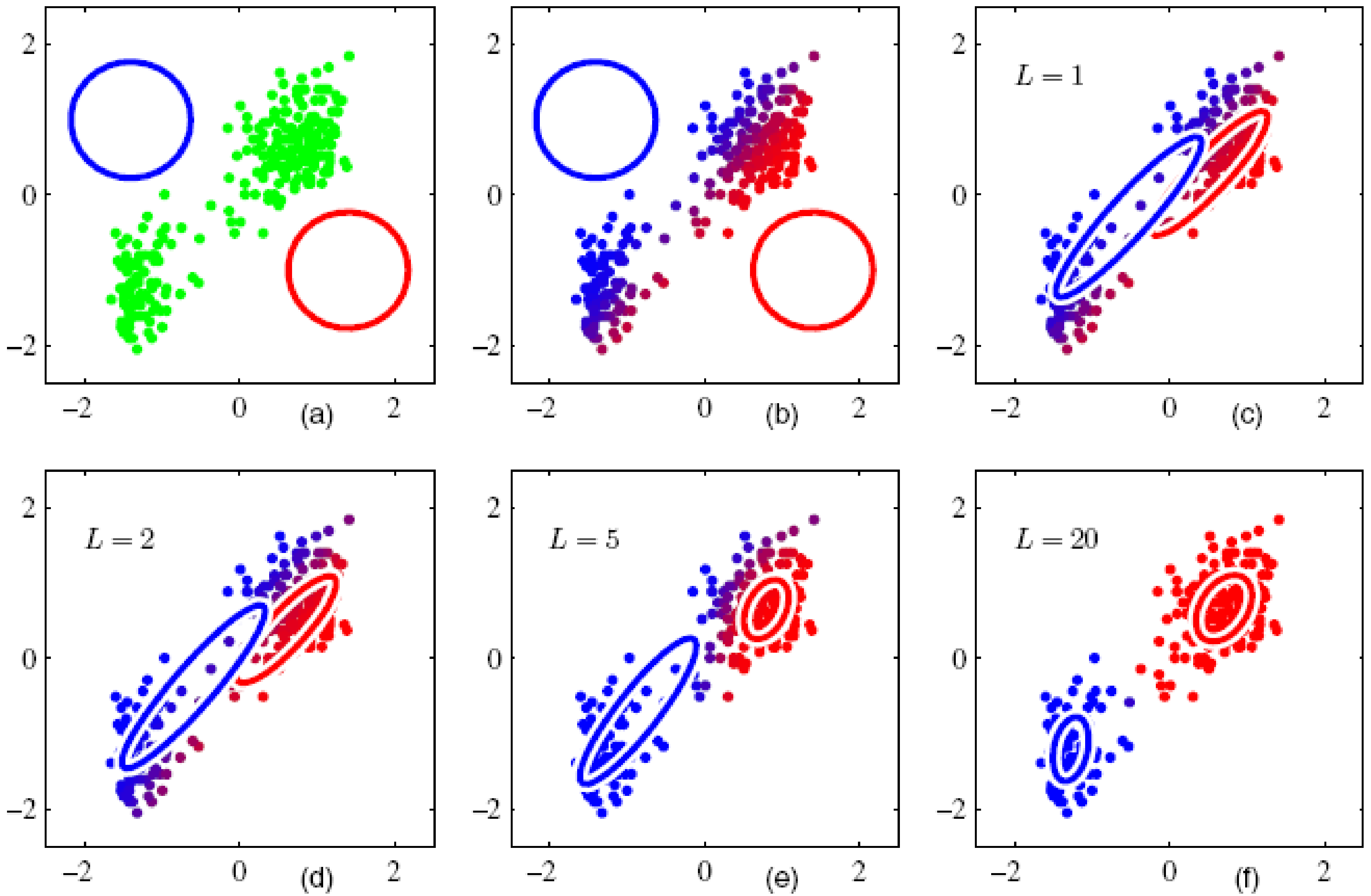
Expectation-Maximization for GMMs

- Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood

2. **E step**: Evaluate the responsibilities $\gamma(z_{nk})$ using the current parameter values
3. **M step**: Re-estimate the parameters $\boldsymbol{\mu}_k^{new}$, $\boldsymbol{\Sigma}_k^{new}$ and π_k^{new} using the current responsibilities
4. Evaluate the log likelihood and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

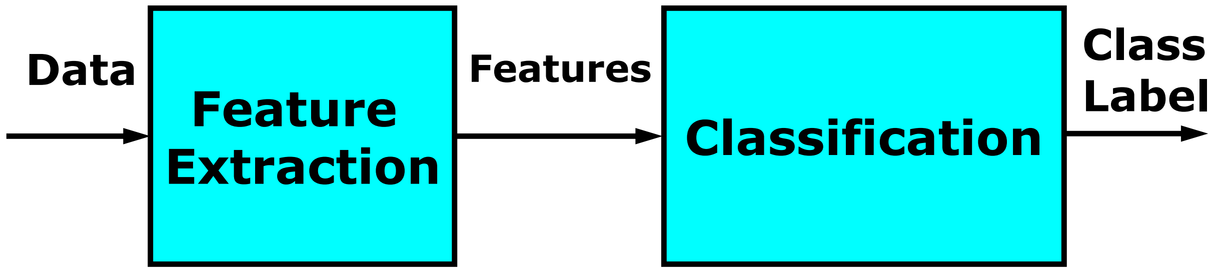
Illustration of Parameter Estimation



C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

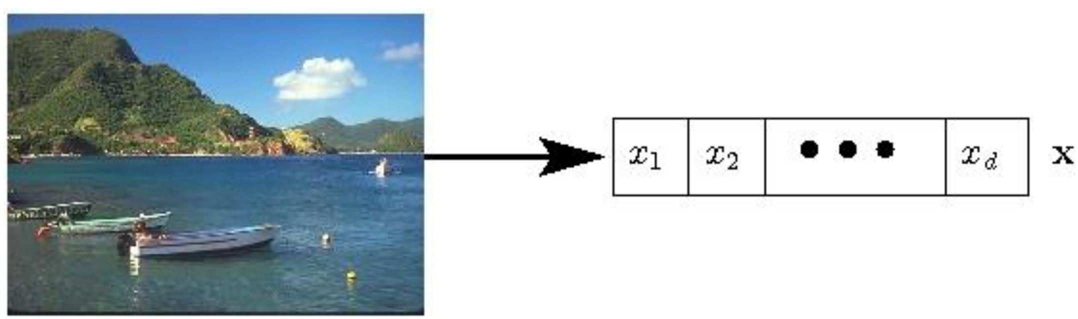
Pattern Classification

- Pattern: Any regularity, relation or structure in data or source of data

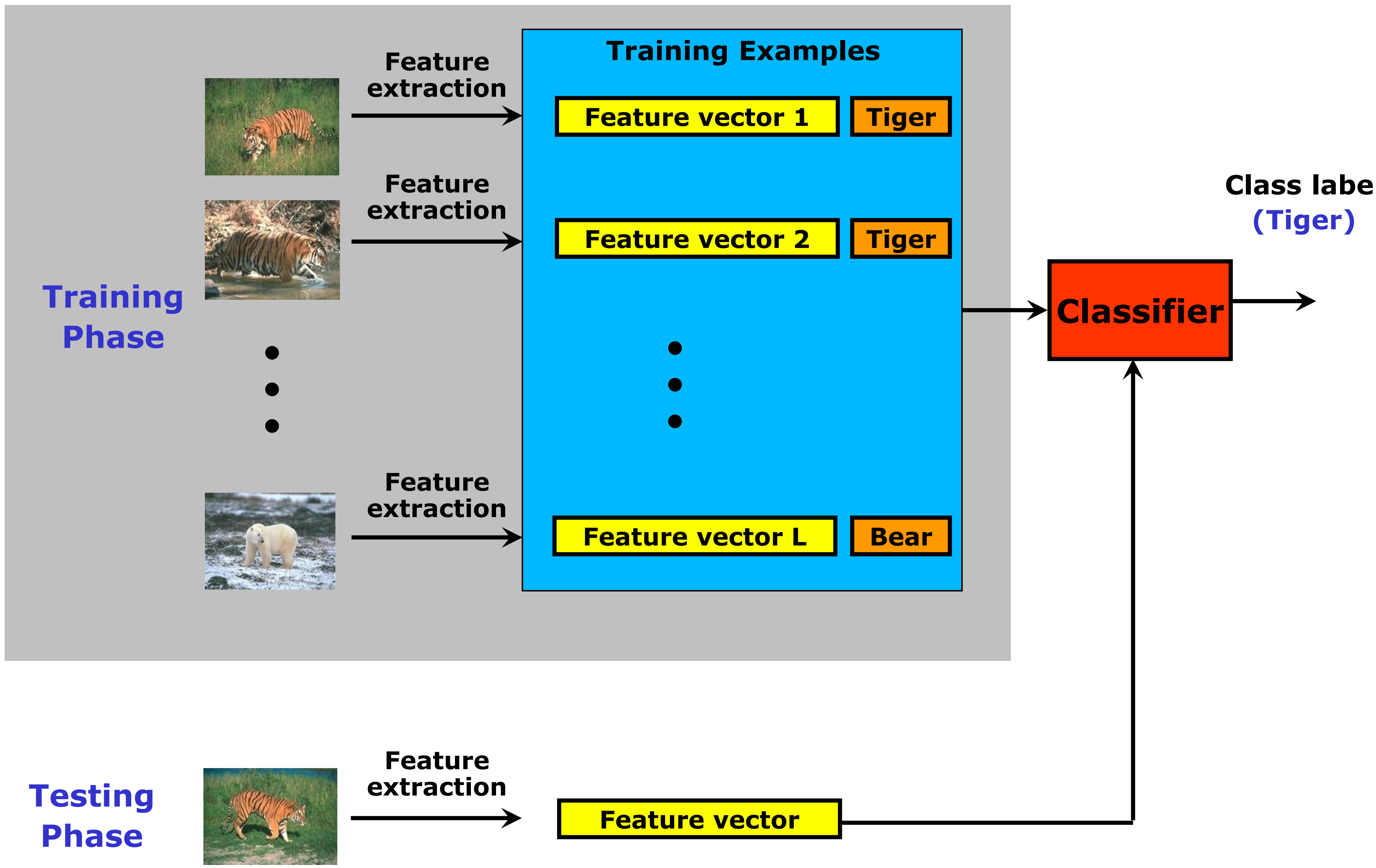


Static Patterns

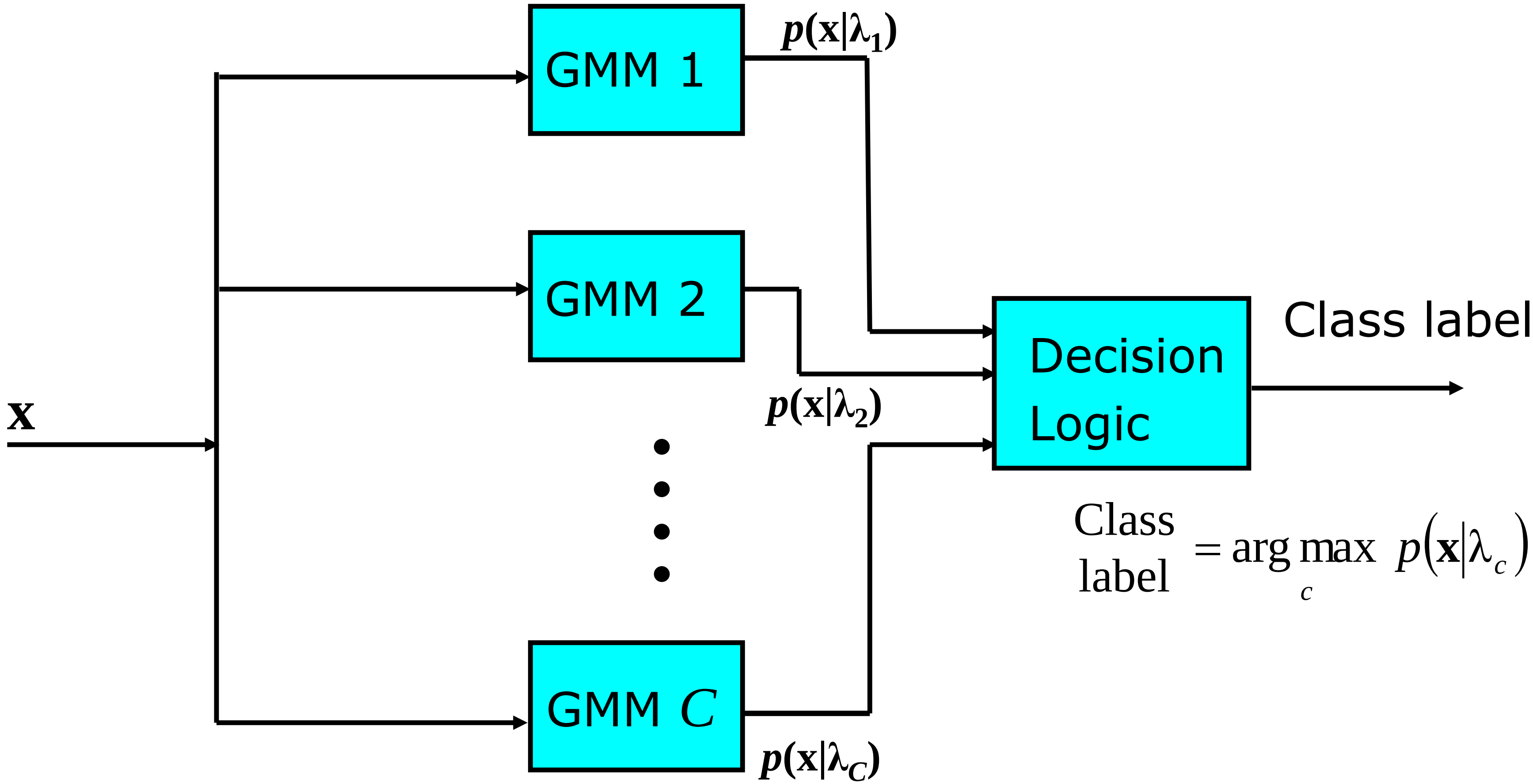
- Static pattern: An example is represented by a **vector of features**



Static Pattern Classification



Static Pattern Classification using GMMs



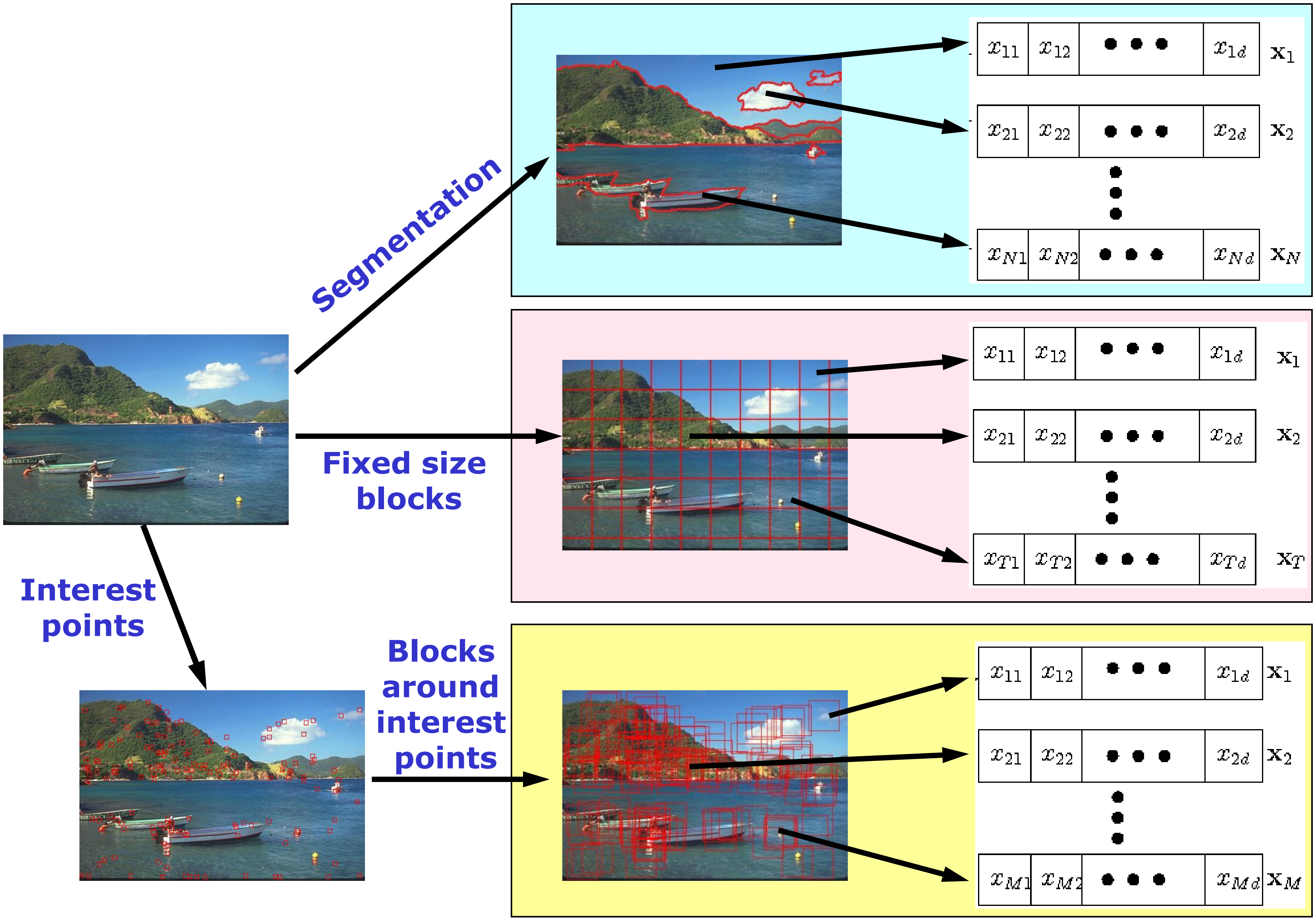
$$\text{Class label} = \arg\max_c \; p(\mathbf{x}|\lambda_c)$$

- $\mathbf{x}$ is a feature vector derived from a test example (an image)
- N_c is the number of components in GMM for the class c
- λ_c is the GMM for the class c

$$p(\mathbf{x}\,|\,\lambda_c) = \sum_{k=1}^{K_c} \pi_{ck} \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck})$$

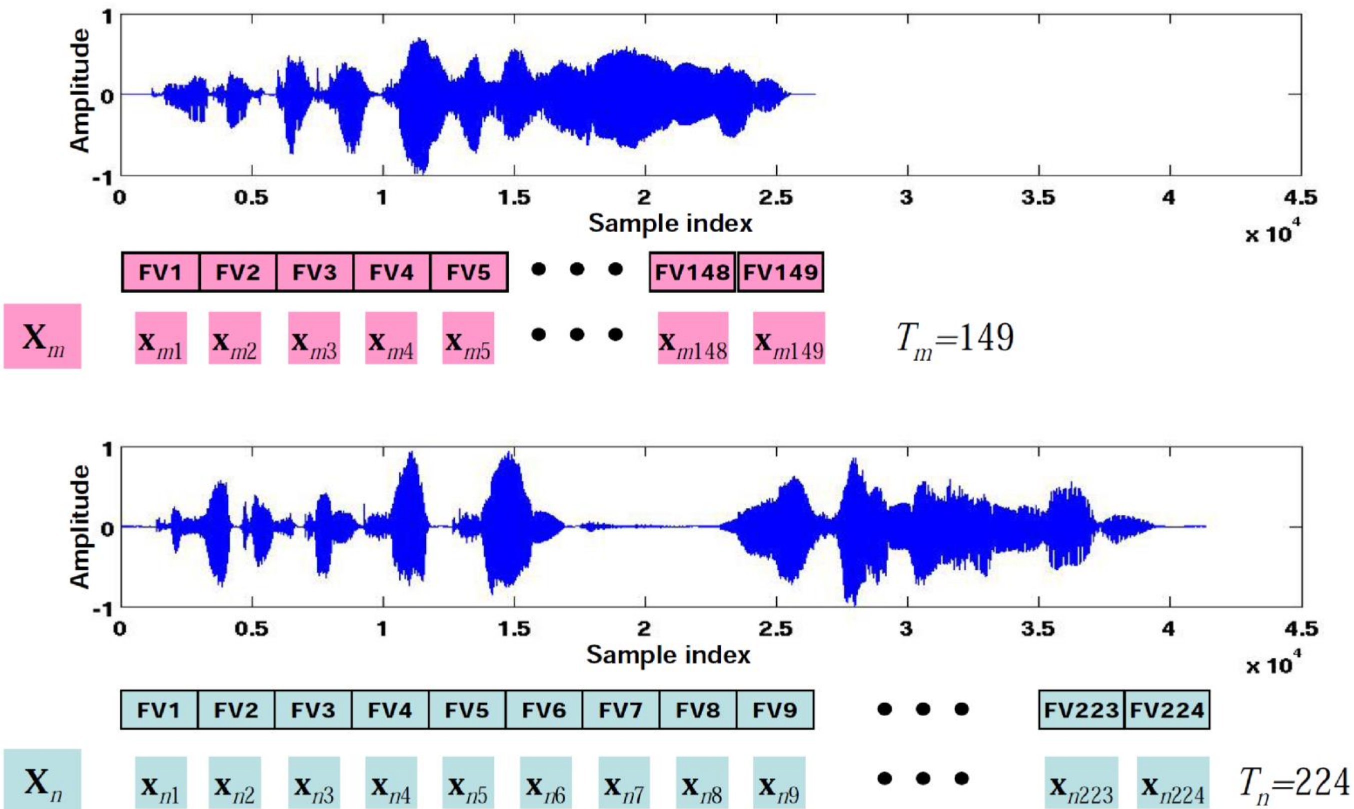
Varying Length Patterns

- Varying length pattern: An example is represented by a **set** or **sequence of feature vectors**

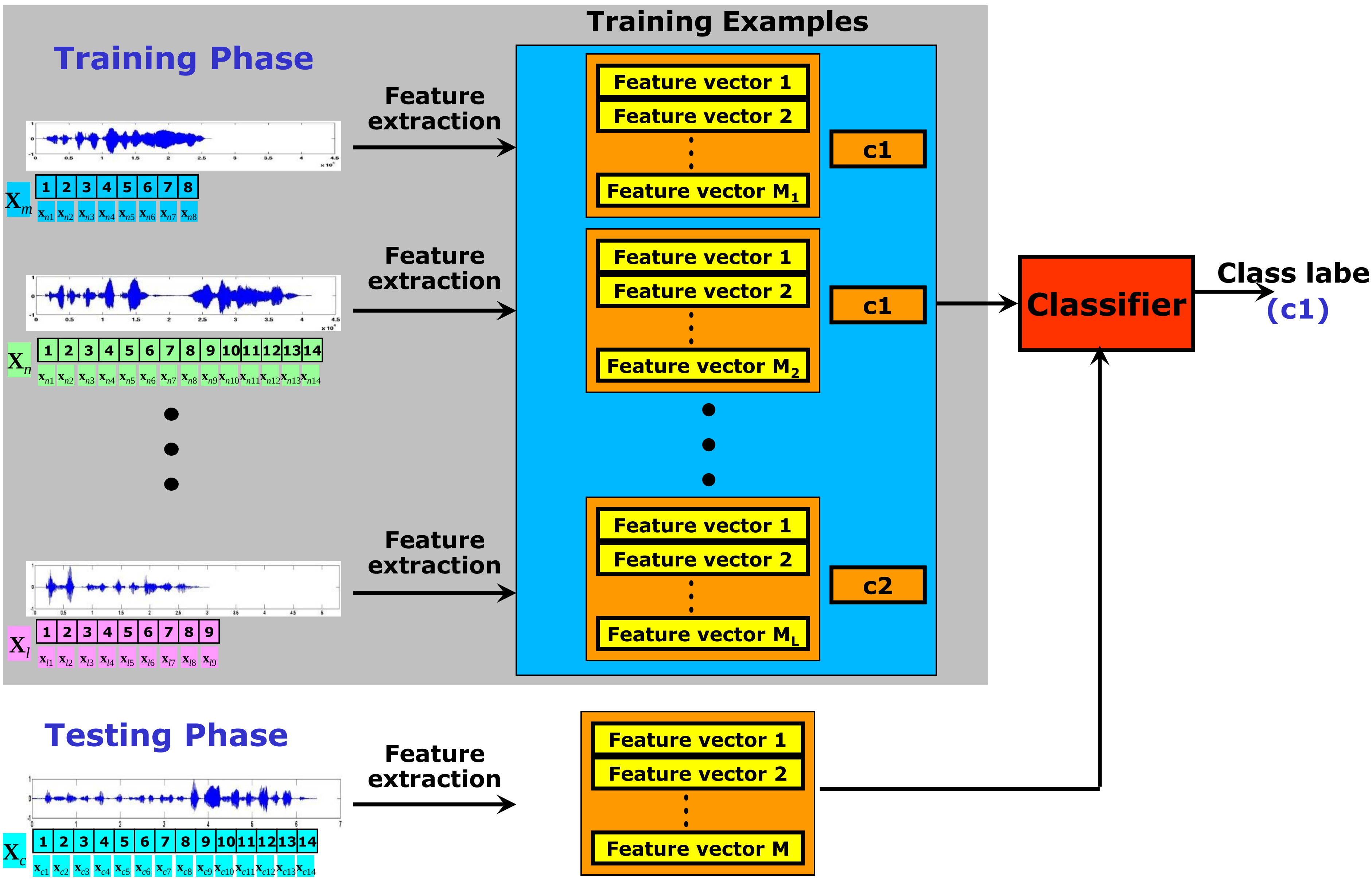


Varying Length Patterns: Sets of Feature Vectors

- **Tasks**: **Speaker recognition**, **speech emotion recognition**, and **spoken language identification**
- Duration of the data is long
- Preserving sequence information is not critical
  - Speech signal of a sentence with anger as emotion



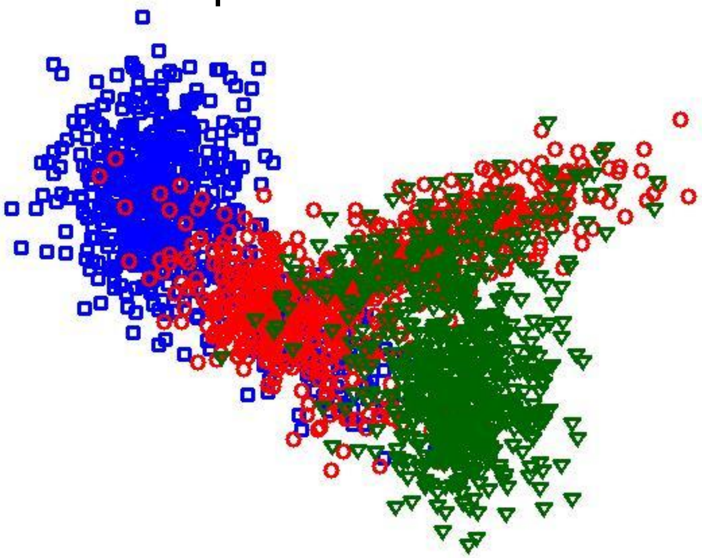
Varying Length Pattern Classification



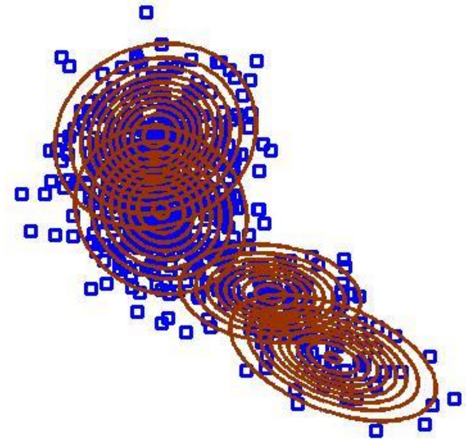
Varying Length Pattern Classification using GMMs - Illustration

- Build GMMs for different classes

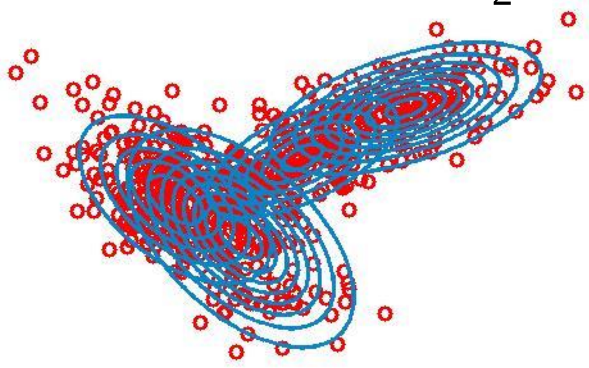
Local feature vectors from examples of all classes



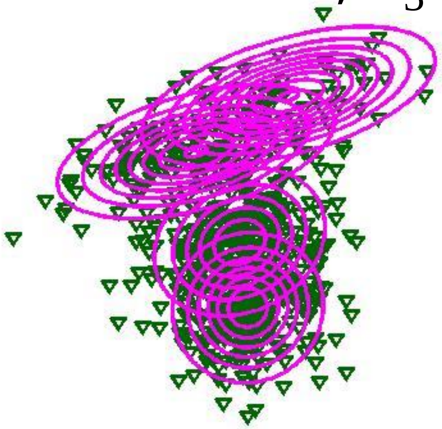
GMM for class 1, λ_1



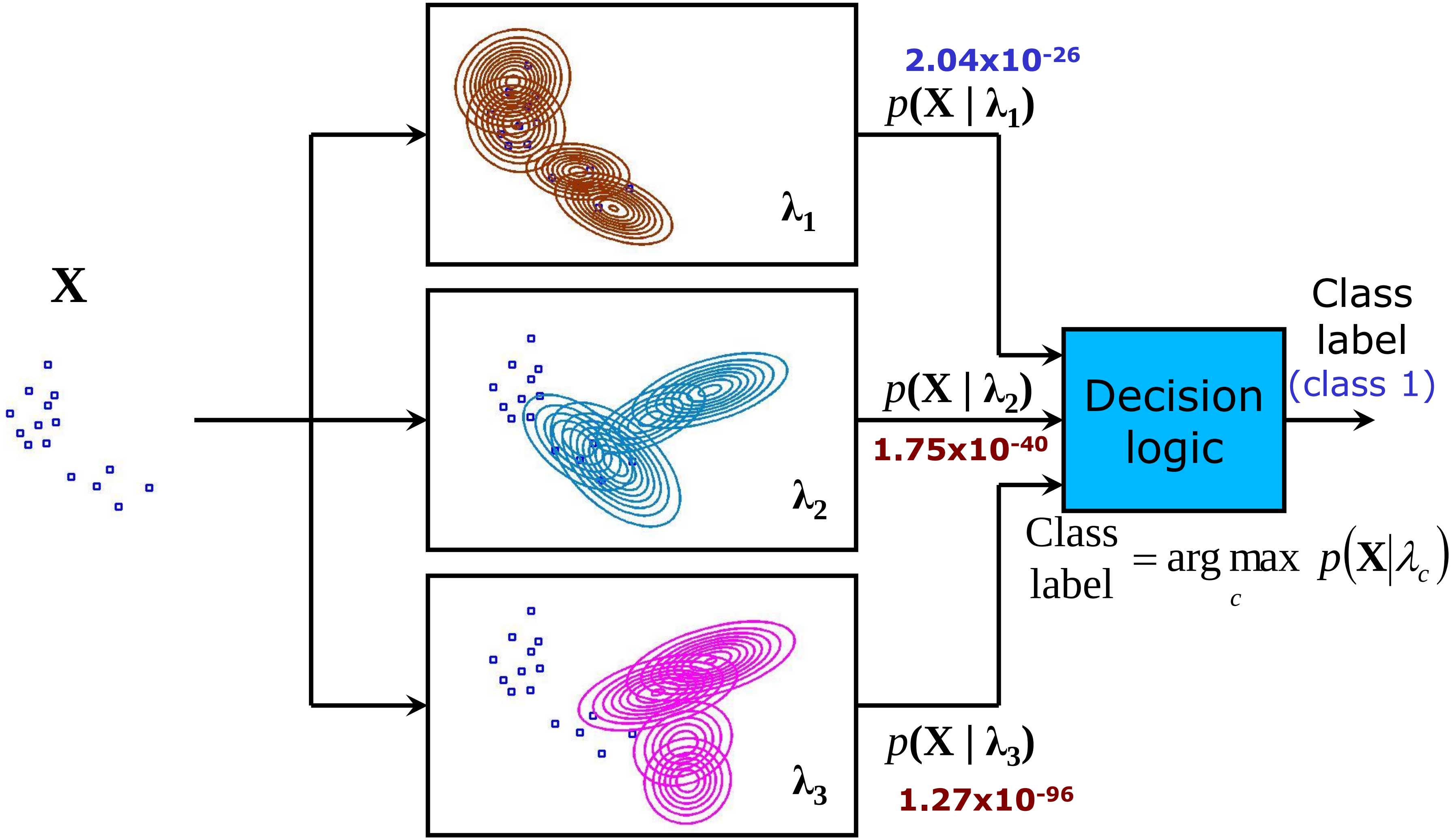
GMM for class 2, λ_2



GMM for class 3, λ_3



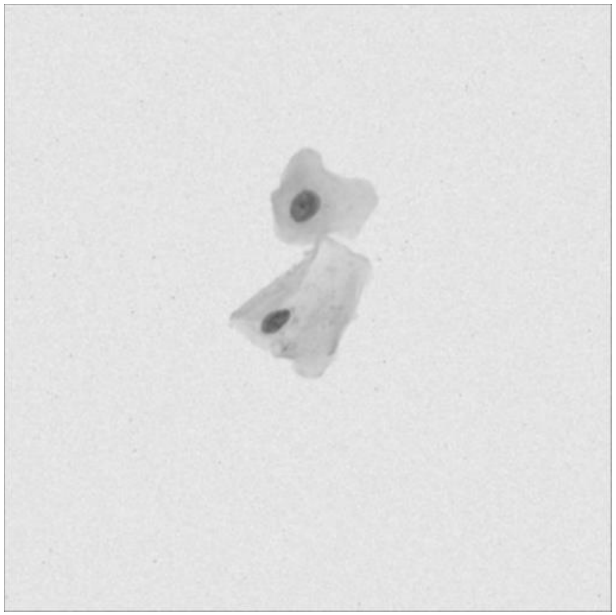
Varying Length Pattern Classification using GMMs – Illustration - Continued …

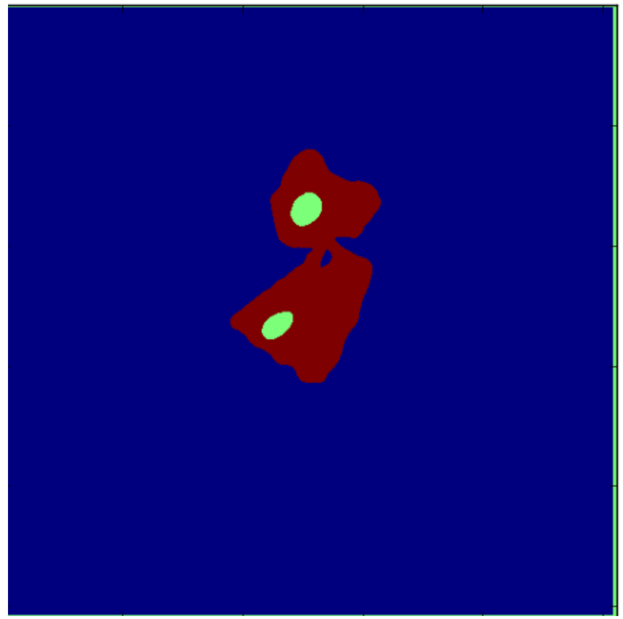


- Set of feature vectors for the test example: $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$

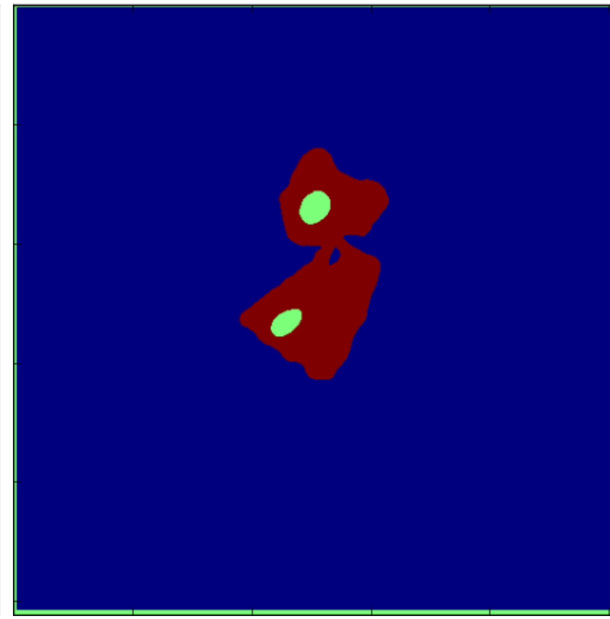
$$p(\mathbf{X} \mid \lambda_c) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \lambda_c) = \prod_{t=1}^{T} \sum_{k=1}^{K_c} \pi_{ck} \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck})$$

Cell and Nucleus Clustering



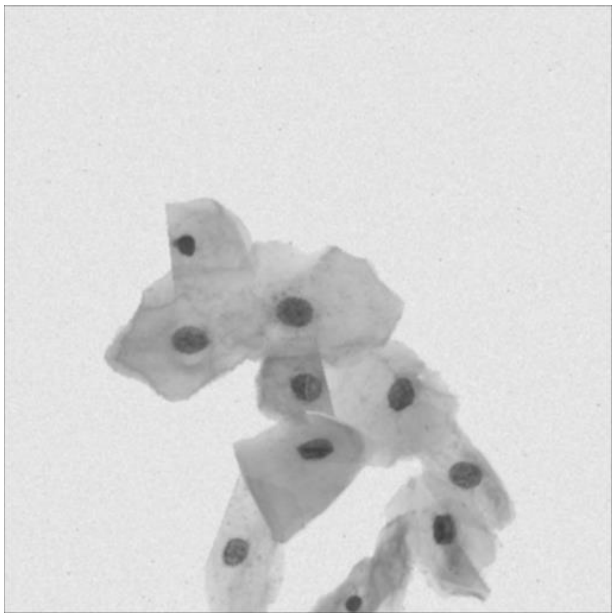


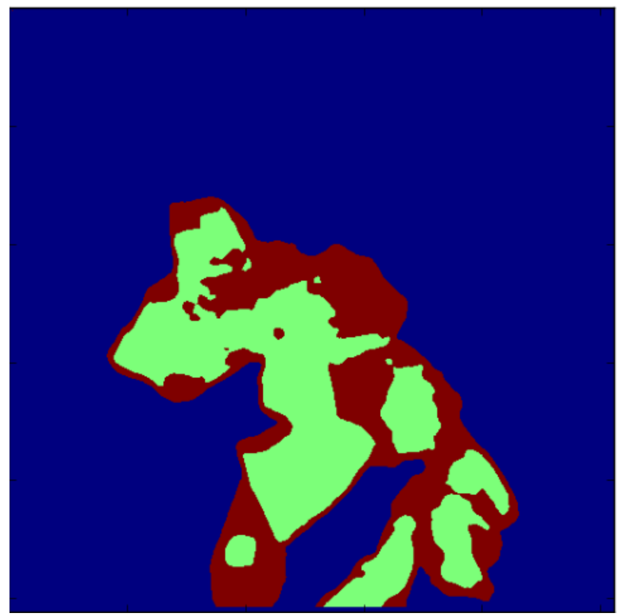
K-Means



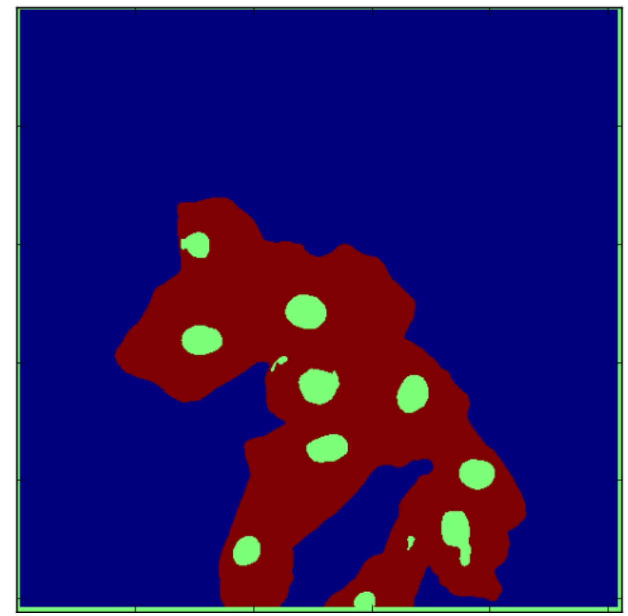
GMM

Cell and Nucleus Clustering



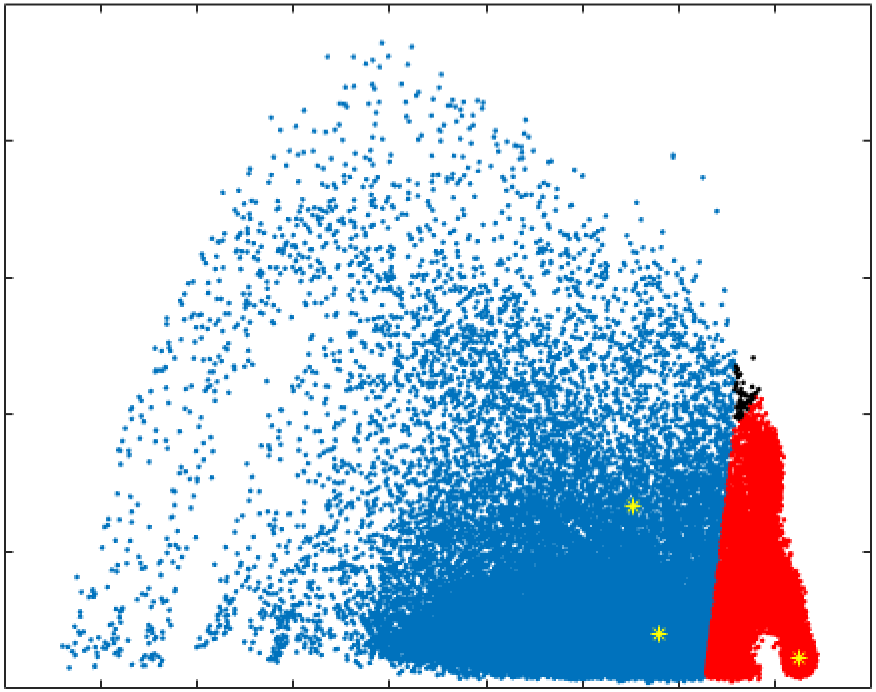


K-Means

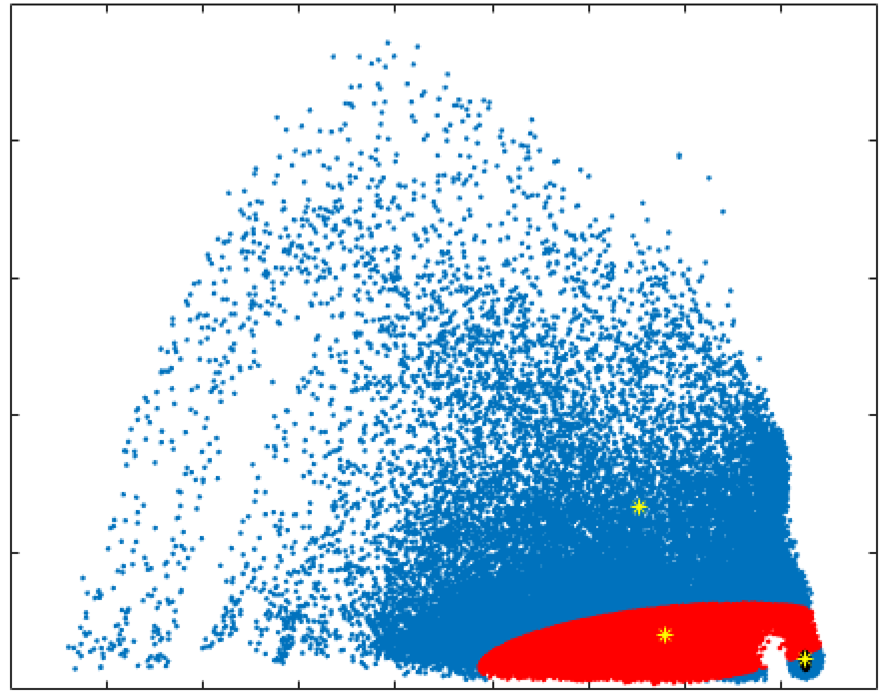


GMM

K-Means vs GMM for Clustering



K-Means



GMM

Gaussian Mixture Models – Summary

- Multimodal probability distribution for each class is represented by a Gaussian mixture model.
- Number of parameters to be estimated for each class is dependent on:
  - Dimensionality of the data space d
  - Number of Gaussian mixtures K

  $K \text{x} d + K \text{x} (d(d+1))/2$
- For large values of d and K, the number of examples required to estimate the parameters properly will be large.
  - Diagonal covariance matrices
- When the estimated class-conditional densities are the same as the true densities, Bayes classifier gives minimum classification error
- An example is represented by a static pattern (a feature vector) or by a set of feature vectors.
- Conventional methods for training the statistical models are non-discriminative learning based methods.

Text Books

1. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley, 2001.

2. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 2009.

3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

Thank You