CSC345/M45: Big Data & Machine Learning (clustering: part two)

Dr. Xianghua Xie

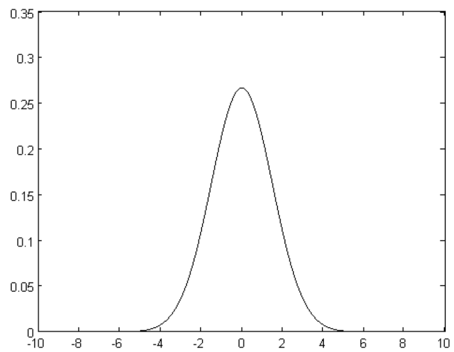
x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
510 Faraday Tower

Clustering

- Clustering is a process to find **similarity groups** in data, called **clusters**
  - **unsupervised learning**
  - K-means: minimise Sum of Squared Error (or sum of squared distances)

- Gaussian Mixture Model
  - Model based approach to data clustering
  - Model is described by several parameters (a set of parameters)
  - However, it is not so called parametric model
    - A parametric model often refers to a single Gaussian function

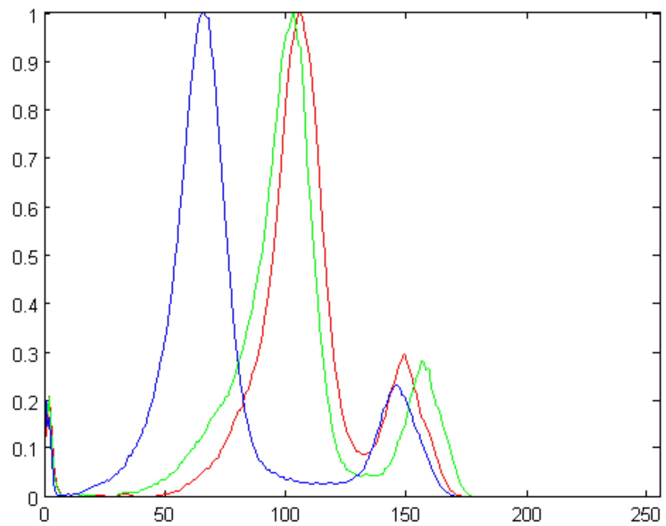
Gaussian mixture modelling



- Gaussian distribution
  - Also known as normal distribution

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}$$

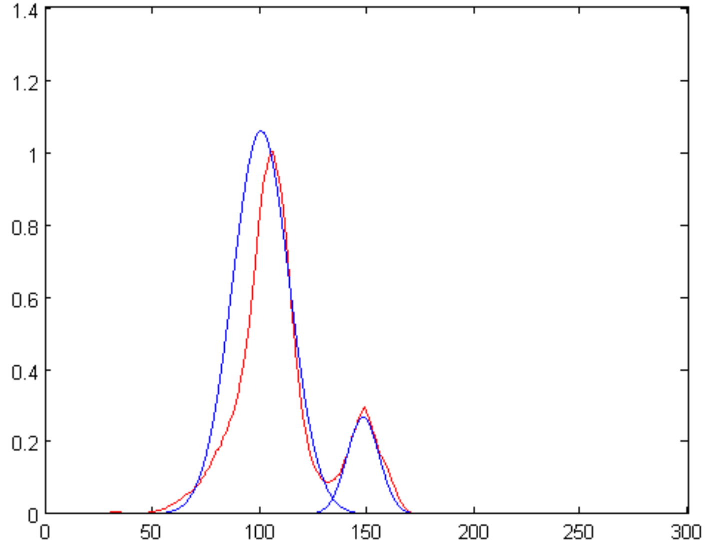
  - μ is the mean, σ is the standard deviation
- A single Gaussian function usually is not sufficient to model histogram distribution, for example
- Image data often is multimodal



Histograms for R, G, & B channels

Gaussian mixture modelling

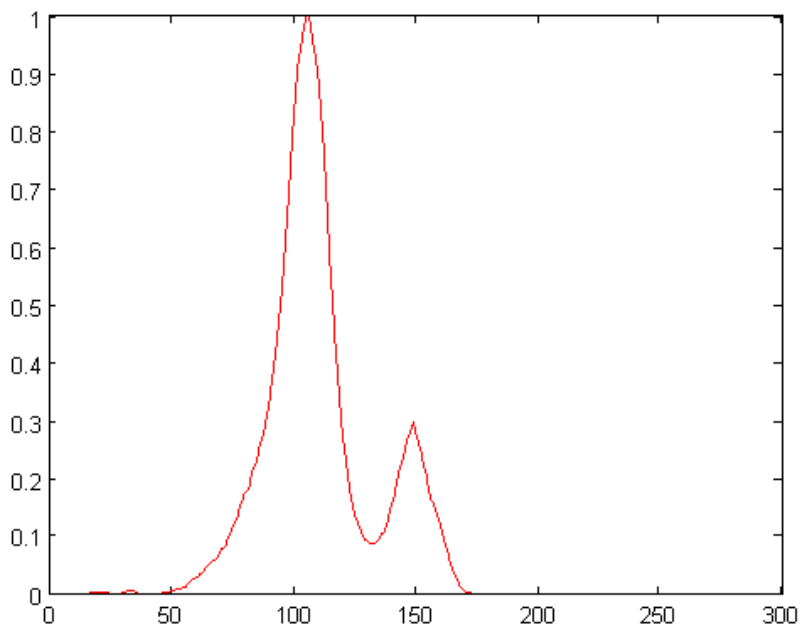
- We can use a mixture of Gaussian functions to learn the distribution of data
- The distribution is modelled by a set of parameters
  - k, the number of Gaussian functions
  - μ, the mean for each Gaussian function
  - σ, the standard deviation for each Gaussian function; more generally: covariance matrix Σ for each Gaussian
  - P, mixing coefficients: weight for each Gaussian function
- Particularly useful in modelling multimodal distribution

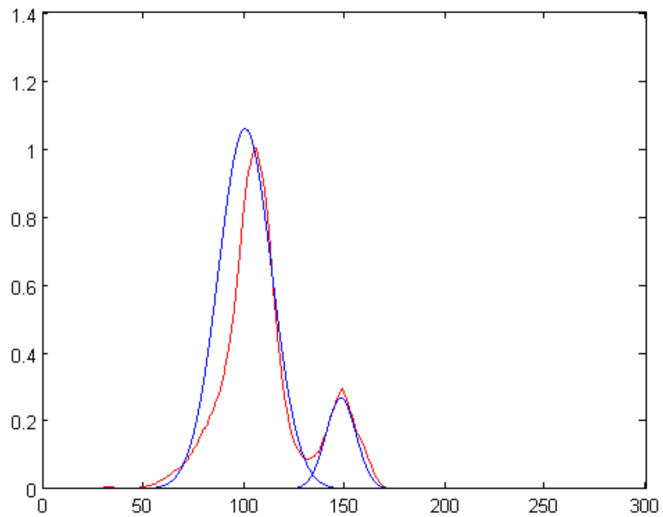


- Red curve: R channel histogram
- Blue curves: 2 Gaussian functions

Gaussian mixture modelling

- The distribution is modelled by a set of parameters
  - k, the number Gaussians
  - μ, the mean for each Gaussian function
  - σ, the standard deviation for each Gaussian function; more generally: covariance matrix Σ for each Gaussian
  - P, mixing coefficients: weight for each Gaussian function





component	μ	σ	P
1	100	13.2	0.83
2	148	7.5	0.17

Gaussian mixture modelling

- GMM is described as a weighted sum of single Gaussian functions:

$$p(x) = \sum_{j=1}^{k} p(x|j)P(j)$$

  - $P(j)$ are the mixing coefficient: known as the prior probability
  - $\sum_{j=1}^{k} P(j) = 1$
  - j indicates jth Gaussian function in GMM
  - Each Gaussian function is given as:

$$p(x|j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} e^{\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\}}$$

    - d is the number of dimensions, T denotes transpose
    - Σ is the d x d covariance matrix; for 1D (d=1), $\Sigma = \sigma^2$

$$\frac{1}{(2\pi\sigma^2)^{1/2}} e^{\{-\frac{(x-\mu)^2}{2\sigma^2}\}}$$

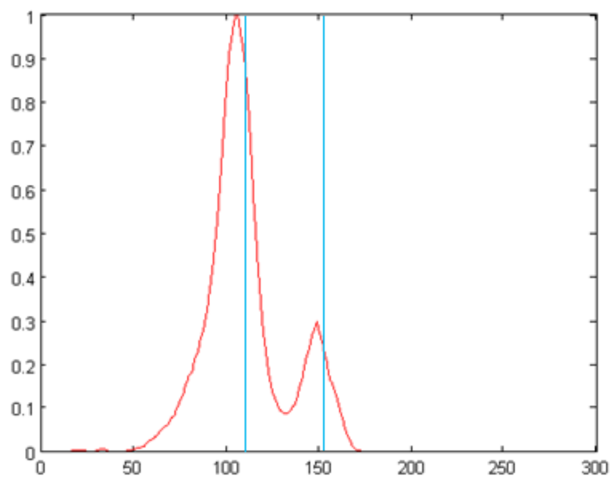
1D case

Gaussian mixture modelling

- Learning GMM parameters: initialisation (step 1)
  – μ, σ or Σ , P; k is given
  – Start from an initial guess of the parameters usually using k-means
    - k-means clustering directly gives mean values
    - Standard deviation or covariance matrix for each Gaussian function can be conveniently computed from clustered data

      $\sigma = \sqrt{\mathrm{E}[(X-\mu)^2]}$ $\mathrm{E}[X] = \mu$ E denotes expectation
    - P is computed by counting number of data belong to each Gaussian component

  – These are our initial expectations



Gaussian mixture modelling

- Learning GMM parameters: posterior probability (step 2)
  - Based on initial expectations, we can now compute the posterior probability – the responsibility of a Gaussian component for explaining the data (or observation)
  - Given by Bayes' theorem:

$$p(j|x) = \frac{p(x|j)P(j)}{p(x)}$$

    - (in other words) the probability the given data x belongs to component j

Recap:

$$p(x) = \sum_{j=1}^{k} p(x|j)P(j)$$

$$p(x|j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} e^{\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\}}$$

Gaussian mixture modelling

- Learning GMM parameters: updating parameters (step 3)
  - We can now update our parameters (x^n is the index of the data)

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

The mixing coefficient is simply the normalised summation of posterior probability

$$\mu_j^{new} = \frac{\sum_n p^{old}(j|x^n)x^n}{\sum_n p^{old}(j|x^n)}$$

Posterior probability weighted mean

$$\Sigma_j^{new} = \frac{\sum_n p^{old}(j|x^n)(x^n - \mu_j^{new})(x^n - \mu_j^{new})^T}{\sum_n p^{old}(j|x^n)}$$

Posterior probability weighted covariance matrix

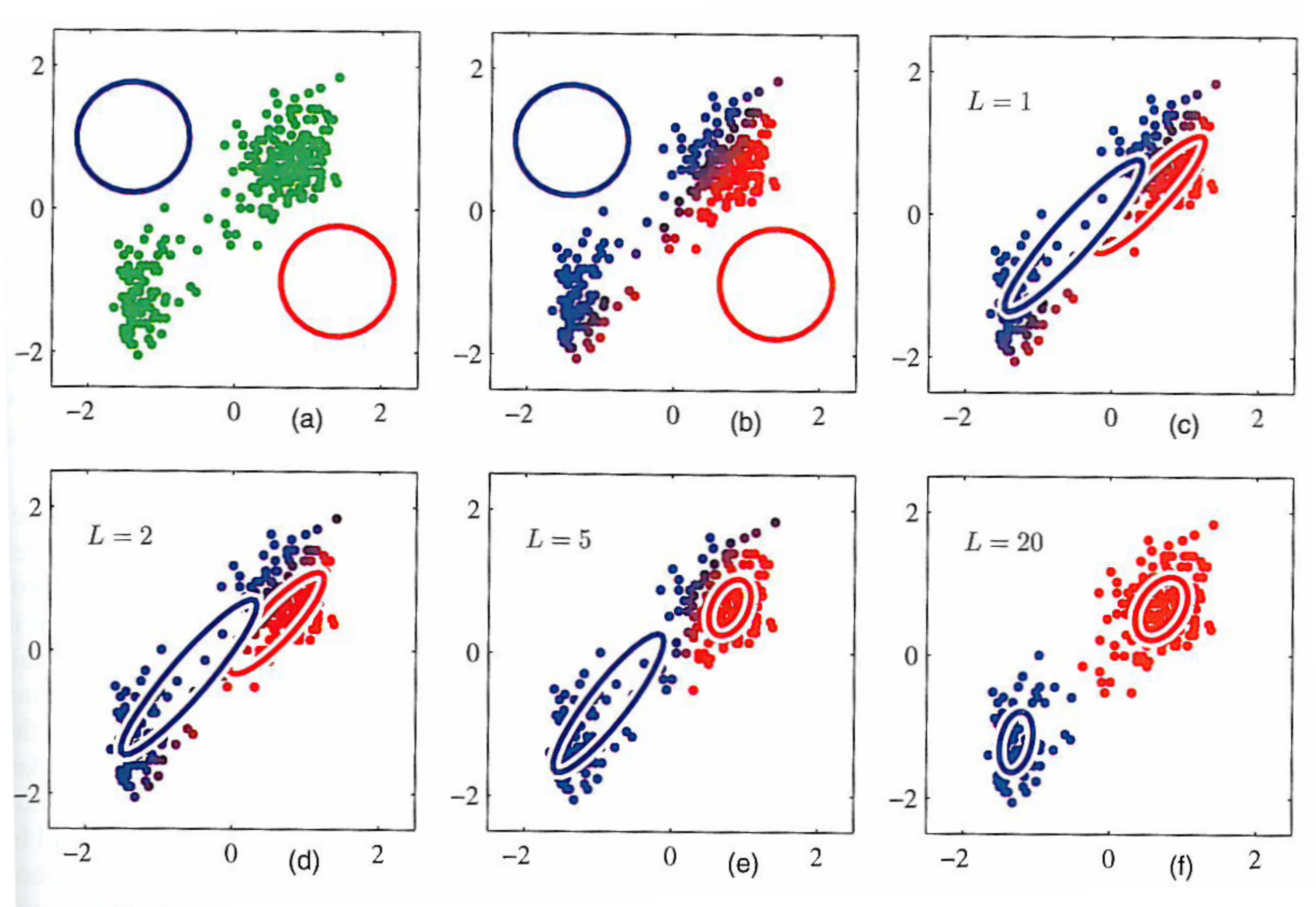
Gaussian mixture modelling

- Learning GMM parameters
  - Iterate steps 2 and 3 until stabilise
  - This process effectively is maximising the log posterior probability
  - This iterative process is a class of Expectation and Maximisation (EM)

- To summarise:
  - Parameters to estimate: μ, Σ and P (k sets of them!)
  - 1. k-means clustering to have initial values
  - 2. compute posterior probability for each data point
  - 3. update parameters
  - 4. repeat 2 and 3 until converges

Gaussian mixture modelling

- Illustration of GMM parameter estimation process
  - Here, no k-means initialisation



Bishop, PRML