Ch 9 of Bishop

ORIE 4742 - Info Theory and Bayesian ML

Chapter 11: Mixture Models

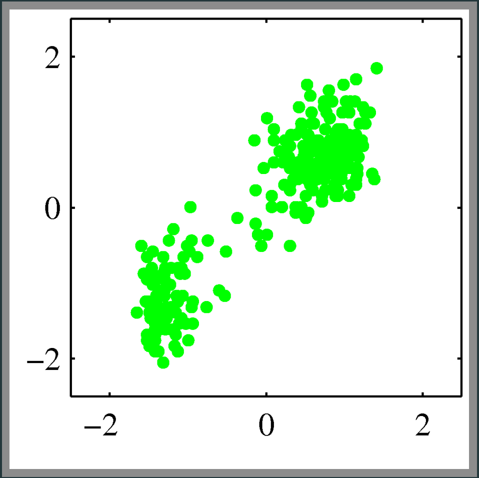
May 5, 2021

Sid Banerjee, ORIE, Cornell

Plan for next few classes

- More complex generative models / decision problems
  - latent variable models (mixture model)
  - optimizing complex fns
    - simulated annealing
      - random walks
      - gradient descent
      - stochastic gradient descent
    - fitting unknown fns with GPs (Bayesian optⁿ)
  - neural networks

example: clustering points in $\mathbb{R}^2$



- Note - No 'training' data (no ground truth)

Idea - 'K-means'

- pick K = number of clusters (Eg. $K=2$)
- pick cluster centers μ_1, μ_2
- For each point x_n, pick 'cluster membership' $r_n = \{r_{n1}, \ldots, r_{nK}\}$ s.t. $\sum_{i=1}^{K} r_{ni} = 1$, $r_{ni} \in \{0,1\}$

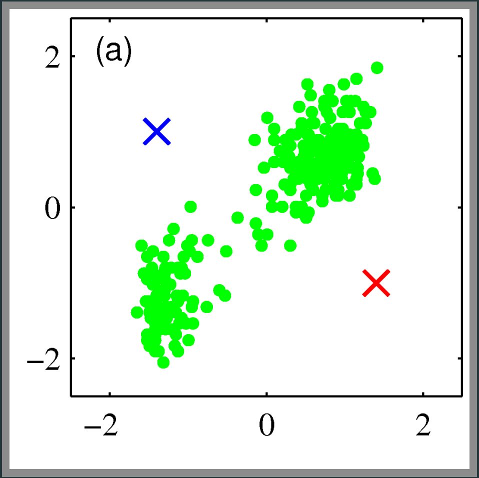
- Aim - minimize 'distortion'

$$\sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| x_n - \mu_k \|_2^2$$

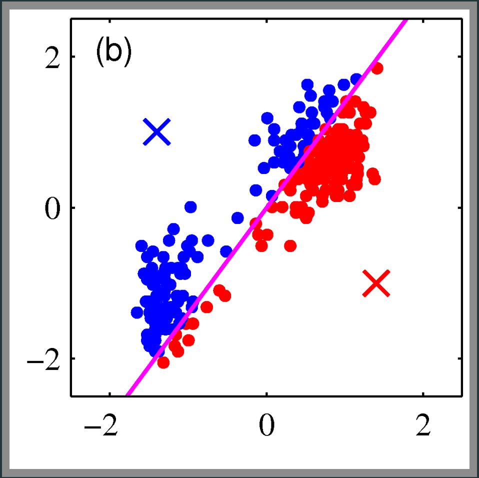
(ie, 'facility location')

approach 1: K-means

start by 'guessing' μ_1, μ_2

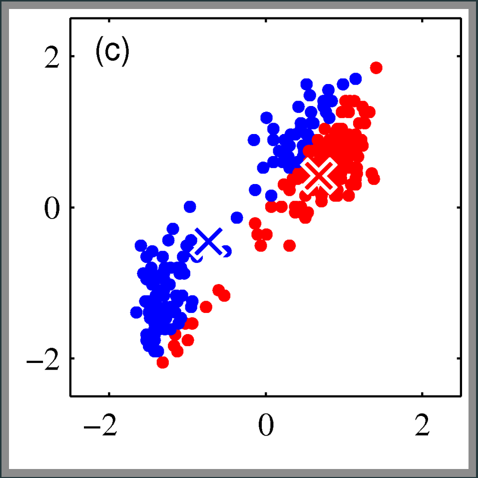


approach 1: K-means

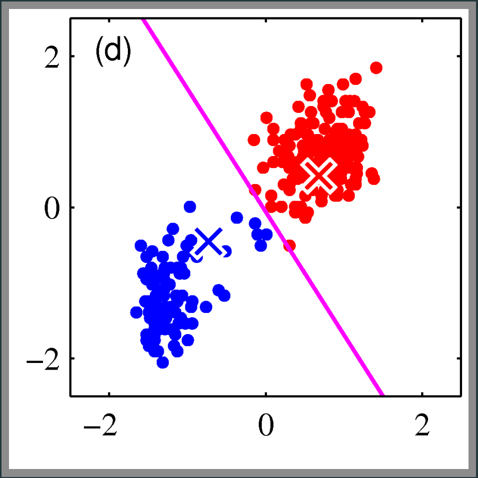


2 ~~1~~ next update cluster center μ_k to minimize $\sum \|x_n - \mu_k\|_2^2$
for x_n s.t. $r_{nk} = 1$

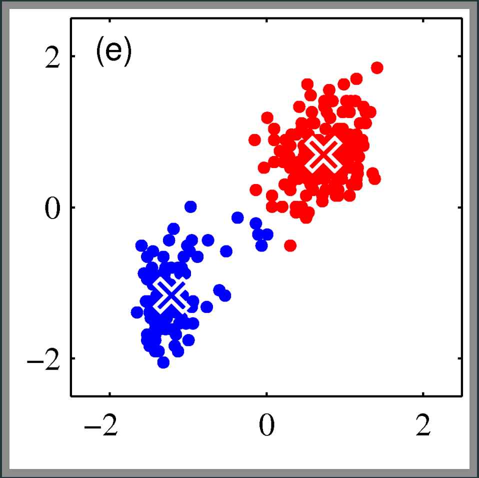
approach 1: K-means



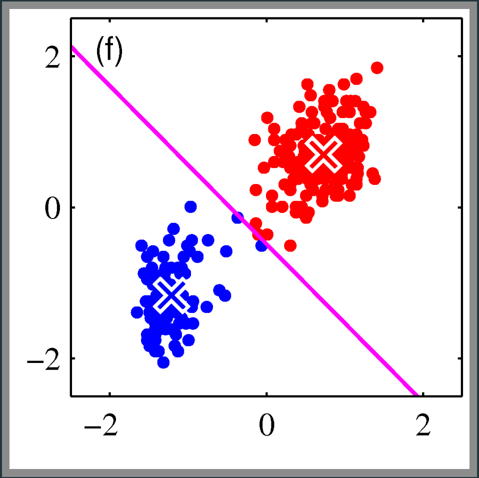
approach 1: K-means



approach 1: K-means



approach 1: K-means



i.e. - Iteratively set $\{r_{nk}\}$ and $\{\mu_k\}$

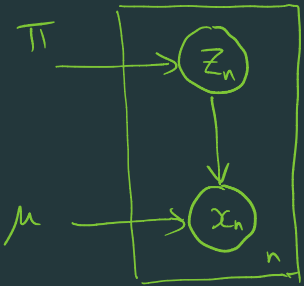
supervised vs. unsupervised learning

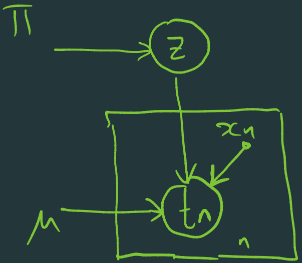
- Suppose we knew cluster labels ('supervised learning')
  - Can use 'standard' Bayesian ML classification models (Naive Bayes, Logistic regression, GP classification)

- The clustering problem - no cluster labels (unsupervised learning) - no examples of 'correct' answers

latent variable generative models

exists, but is not in the data





'common' latent variable

Eg - regression

the Gaussian mixture model

- data $D = \{X_1, X_2 \ldots, X_N\} \in \mathbb{R}^d$
- each point X_n has a latent cluster label in $\{1, 2, \ldots, K\}$ indicator vectors
  denoted by $Z_n \in \{0,1\}^K, \sum_{k=1}^K z_{n,i} = 1$ (1-of-K encoding)
- latent variable: $Z_n \sim \mathrm{Mult}(\pi_1, \pi_2, \ldots, \pi_K)$ where $\sum_{i=1}^K \pi_i = 1$ i.e. $Z_n = e_i$ with prob $\frac{\pi_i}{\Sigma \pi_i}$
  data: if latent cluster is $k \in [K]$, then $X_n \sim \mathcal{N}(\mu_k, \Sigma_k)$
- joint likelihood:

$$p(X, Z|\mu, \Sigma, \pi) = \prod_{n=0}^{N-1} \prod_{k=0}^{K-1} [\pi_k \mathcal{N}(X_n|\mu_k, \Sigma_k)]^{z_{n,k}}$$

$(2\pi)^{-d/2} \Sigma_k^{-1/2} e^{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$

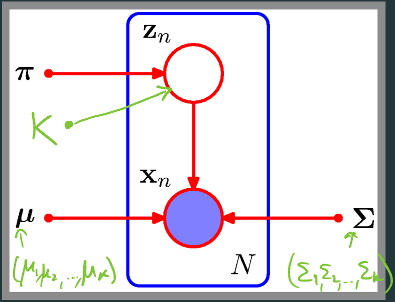
- log-likelihood of data:

$$\log p(X|\mu, \Sigma, \pi) = \sum_{n=0}^{N-1} \underbrace{\log \left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}(X_n|\mu_k, \Sigma_k) \right]}_{\text{not convex!}}$$

the Gaussian mixture model

log-likelihood of data:

$$\log p\left(X|\mu,\Sigma,\pi\right) = \sum_{n=0}^{N-1} \log\left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}\left(X_n|\mu_k,\Sigma_k\right)\right]$$



BayesNet for GMM

hyperparams

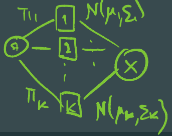
- K
- $\pi = \{\pi_1, \pi_2, \dots, \pi_K\}, \ \sum_{i=1}^{K} \pi_i = 1$
- For each $k \in [K]$: $\mu_k \in \mathbb{R}^d$

$\Sigma_k = d \times d$, pos. def

the responsibility function

given a Gaussian mixture model with known $\{\mu_k, \Sigma_k\}_{k\in[K]}$, and any data point X, we can associate a responsibility parameter to each cluster for the point to be the probability of the underlying latent cluster

responsibility



$$\gamma(z_k) = \mathbb{P}\left(z_k = 1 | X\right) = \frac{\pi_k \mathcal{N}\left(X|\mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(X|\mu_j, \Sigma_j\right)}$$

x

- Prior π, data X $\Rightarrow$ Posterior $=$ γ
- 'Responsibility' each cluster has for 'explaining' the data X

GMM: maximizing the likelihood

$$\log p\left(X|\mu,\Sigma,\pi\right)=\sum_{n=0}^{N-1}\log\left[\sum_{k=0}^{K-1}\pi_k\overbrace{\mathcal{N}\left(X_n|\mu_k,\Sigma_k\right)}^{\propto\ \Sigma_k^{-1/2}\exp\left(-\frac{1}{2}(X_n-\mu_k)^T\Sigma_k^{-1}(X_n-\mu_k)\right)}\right]$$

s.t. $\sum_{i=1}^{K}\pi_i=1$

- Set $\dfrac{\partial \log p(X|\mu,\Sigma,\pi)}{\partial\Theta}=0$ for (first order conditions)

$$\Theta\in\{\mu_1,\mu_2,\dots,\mu_K,\Sigma_1,\Sigma_2,\dots,\Sigma_K,\pi_1,\pi_2,\dots,\pi_K\}$$

GMM: maximizing the likelihood (for μ_k)

$$\log p\left(X|\mu,\Sigma,\pi\right) = \sum_{n=0}^{N-1} \log\left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}\left(X_n|\mu_k,\Sigma_k\right)\right]$$

- FOC :

$$-\sum_{n=1}^{N} \underbrace{\frac{\pi_k N(X|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x|\mu_j,\Sigma_j)}}_{\gamma_{x_n}(z_k)} \underbrace{\Sigma_k}\,(X_n-\mu_k) = 0$$

assuming invertible, multiply by Σ^{-1}

$$\Rightarrow \quad \mu_k^* = \frac{\sum_{n=1}^{N} \gamma_{x_n}(z_k) X_n}{\underbrace{\sum_{n=1}^{N} \gamma(z_k)}} = \underbrace{\frac{1}{N_k} \sum_{n=1}^{N} \gamma_{x_n}(z_k) X_n}$$

weighted sum of X_n

'effective' # of pts in cluster k

GMM: maximizing the likelihood (for σ_k)

$$\log p\left(X|\mu,\Sigma,\pi\right) = \sum_{n=0}^{N-1} \log\left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}\left(X_n|\mu_k,\Sigma_k\right)\right]$$

Similarly (after some algebra) $\left(N_k = \sum_{n=1}^{N} \gamma_{x_n}(z_k)\right)$

$$\Sigma_k^{*} = \underbrace{\frac{1}{N_k} \sum_{n=1}^{N} \gamma_{x_n}(z_k^{*})}_{\text{weighted sum}} \underbrace{(x_n - \mu_k^{*})^T (x_n - \mu_k^{*})}_{\text{empirical cov mat}}$$

GMM: maximizing the likelihood (for π_k)

$$\log p\left(X|\mu,\Sigma,\pi\right) = \sum_{n=\cancel{0}\,1}^{N-\cancel{1}} \log\left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}\left(X_n|\mu_k,\Sigma_k\right)\right]$$

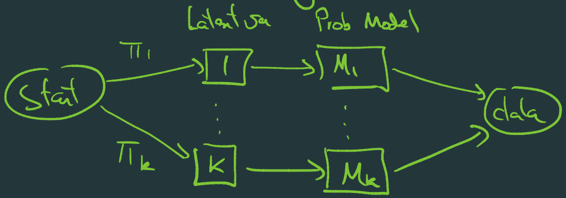
s.t. $\sum_{k=1}^{K} \pi_k = 1$

- $\min\limits_{\lambda} \max\limits_{\pi_k} \ \ln\left(p(X|\mu,\Sigma,\pi)\right) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$
- Inner problem: wrt π_k $\quad \sum_{n=1}^{N} \frac{1}{\pi_k}\gamma_{x_n}(z_k) + \lambda = 0, \quad \sum \pi_k^* = 1$

$$\Rightarrow \quad \pi_k^* = \frac{N_k}{\sum_{k=1}^{K} N_k}, \quad N_k = \sum_{n=1}^{N} \gamma_{x_n}(z_k^*)$$

Notes

1) Works for any mixture model



2) Problem - The FOCs are 'circular'

'expectation'
E-step

'maximization'
M-step

heuristic - alternate max^n
'EM algorithm'

problems with MLE for GMMs

log-likelihood of data:

$$\log p(X|\mu, \Sigma, \pi) = \sum_{n=0}^{N-1} \log \left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}(X_n|\mu_k, \Sigma_k) \right]$$

1) log-likelihood non-convex ⇒ unclear if unique maxima

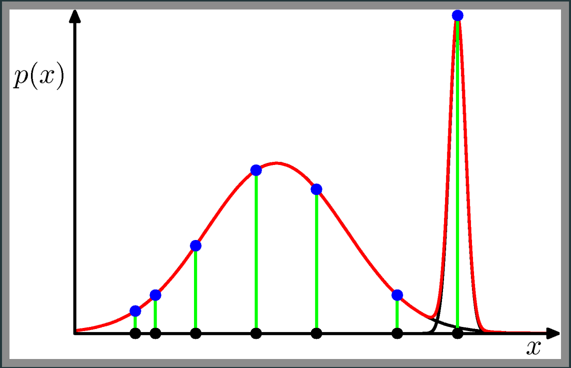
2) What if we change 'labels' of clusters? likelihood remains same! ⇒ K! alternate maxima ('benign' alternate maxima)

problems with MLE for GMMs

(singularity problem) 'kaboom'

log-likelihood of data:

$$\log p\,(X|\mu,\Sigma,\pi) = \sum_{n=0}^{N-1} \log \left[\sum_{k=0}^{K-1} \pi_k \mathcal{N}\,(X_n|\mu_k,\Sigma_k) \right]$$



- Can partition points into 2 clusters $X \backslash x_1$, x_1
- Fit x_1 with a very sharp distⁿ

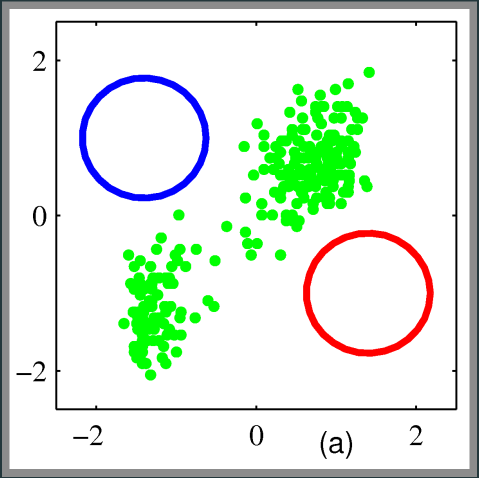
$\Rightarrow$ likelihood $\nearrow \infty$

(bad local minima)

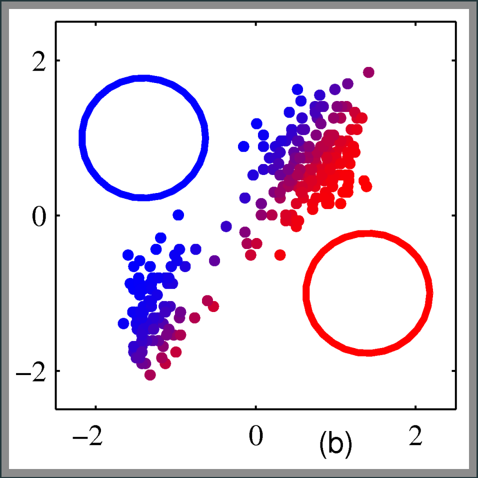
MLE for GMM: an alternate viewpoint

the EM algorithm

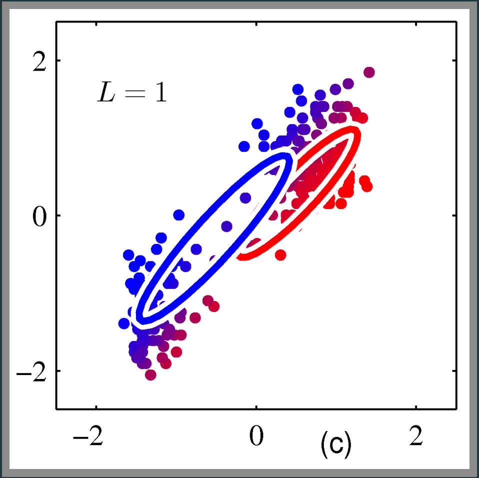
EM algorithm in action



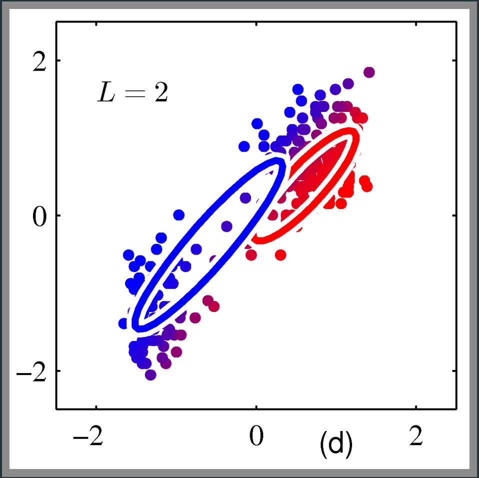
EM algorithm in action



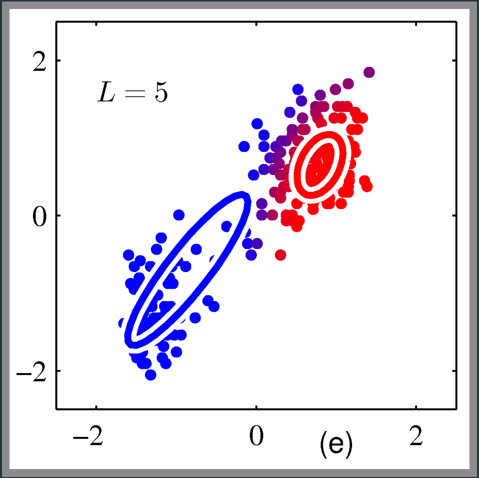
EM algorithm in action



EM algorithm in action



EM algorithm in action



EM algorithm in action