Expectation Maximization:

Inferring model parameters and class labels

CS 229: Machine Learning

Emily Fox
Stanford University

March 4, 2024

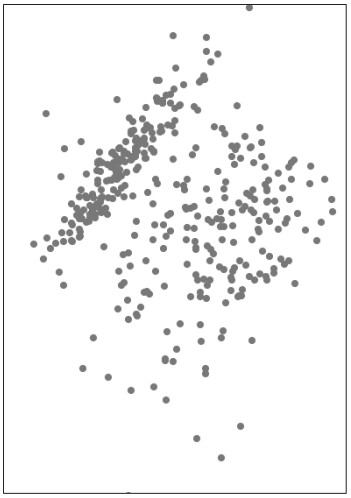
1

Inferring soft assignments with expectation maximization (EM)

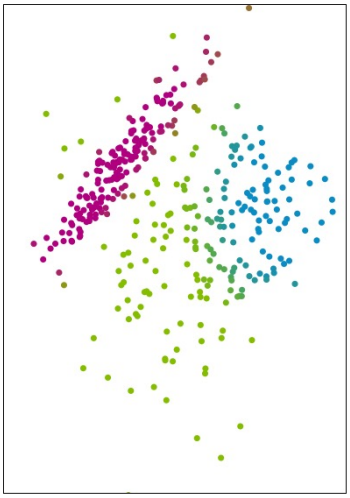
2

Inferring cluster labels

Data



Desired soft assignments

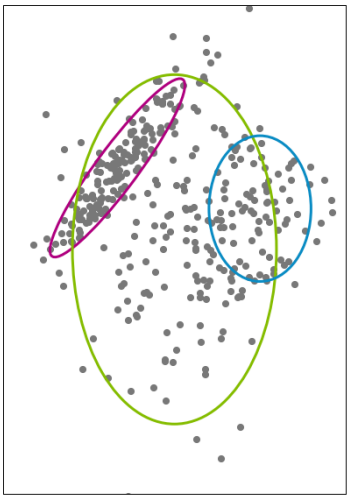


3

Part 1:
What if we knew the cluster parameters $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$?

4

Compute responsibilities



Responsibility cluster k takes for observation i

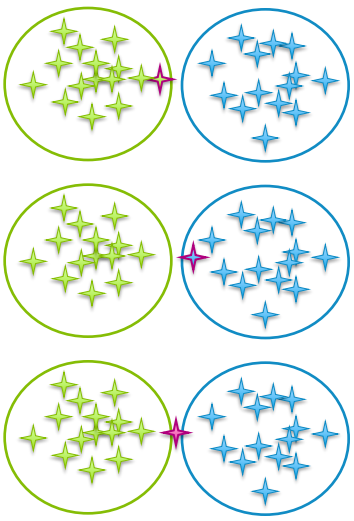
$$r_{ik} = p(z_i = k \mid \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}, x_i)$$

probability of assignment to cluster k

given model parameters and observed value

5

Responsibilities in pictures



Green cluster takes more responsibility

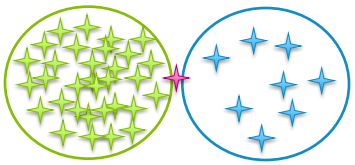
Blue cluster takes more responsibility

Uncertain... split responsibility

6

Responsibilities in pictures

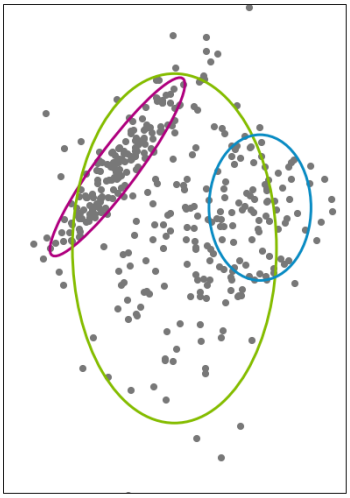
Need to weight by cluster probabilities, not just cluster shapes



Still uncertain, but green cluster seems more probable…
takes more responsibility

7

Responsibilities in equations

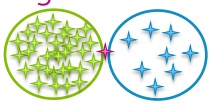


Responsibility cluster k takes for observation i

$$r_{ik} = \pi_k \; N(x_i \mid \mu_k, \Sigma_k)$$

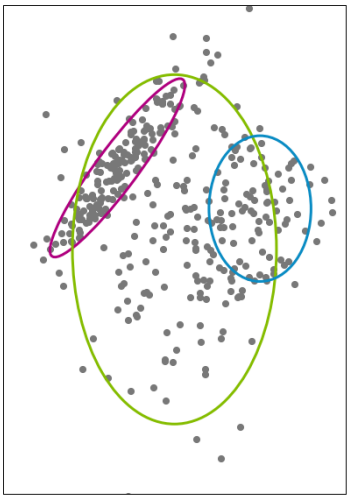
Initial probability of being from cluster k

How likely is the observed value $\mathbf{x}_i$ under this cluster assignment?



8

Responsibilities in equations



Responsibility cluster k takes for observation i

$$r_{ik} = \frac{\pi_k \; N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

Normalized over all possible cluster assignments

9

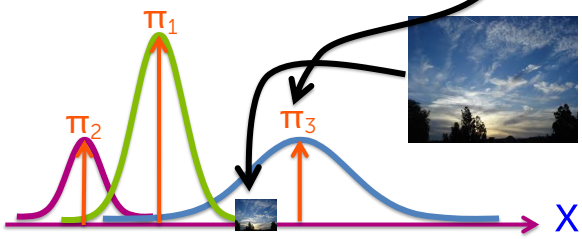
Recall: According to the model…

Without observing the image content, what's the probability it's from cluster k? (e.g., prob. of seeing "clouds" image)

$$p(z_i = k) = \pi_k$$

Given observation $\mathbf{x}_i$ is from cluster k, what's the likelihood of seeing $\mathbf{x}_i$? (e.g., just look at distribution for "clouds")

$$p(x_i \mid z_i = k, \mu_k, \Sigma_k) = N(x_i \mid \mu_k, \Sigma_k)$$



10

Formally: An application of Bayes' rule

Responsibility cluster k takes for observation i

$$r_{ik} = p(z_i = k \mid \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^K, \text{params})$$

$$= \frac{\pi_k \; N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

$p(z_i = k \mid \text{params})$

$p(\text{params} \mid z_i = k, \text{params})$

11

Formally: An application of Bayes' rule

Responsibility cluster k takes for observation i

$$r_{ik} = p(z_i = k \mid \{\pi_j, \mu_j, \Sigma_j\}_{j=1}^K, \text{params})$$

$$= \frac{\pi_k \; N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)}$$

$p(z_i \in j \mid \text{params})$

$p(\text{params} \mid z_i \in j, \text{params})$

12

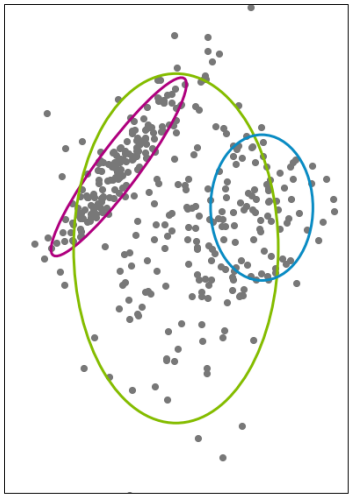
Formally: An application of Bayes' rule

$$r_{ik} = p(A|B,\text{params})$$

$$= \frac{p(A|\text{params})p(B|A,\text{params})}{\sum_C p(C|\text{params})p(B|C,\text{params})}$$

$$= \frac{p(A|\text{params})p(B|A,\text{params})}{p(B|\text{params})}$$

Part 1 summary



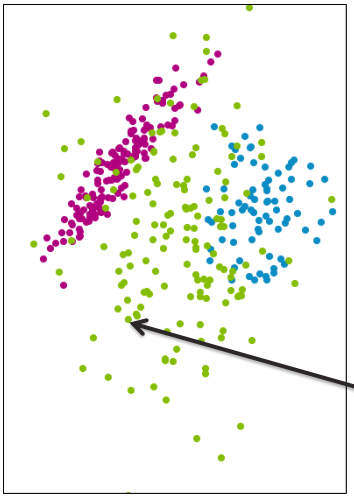
Desired soft assignments (responsibilities) are **easy** to compute when cluster parameters $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$ are known

But, we don't know these!

Part 2a:
Imagine we knew the cluster (hard) assignments z_i

15

Estimating cluster parameters



Imagine we know the cluster assignments

Estimation problem decouples across clusters

Is green point informative of fuchsia cluster parameters? NO!

16

Data table decoupling over clusters

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3
$x_4[1]$	$x_4[2]$	$x_4[3]$	1
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

17

Maximum likelihood estimation

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

Estimate $\{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$ given data assigned to cluster k

maximum likelihood estimation (MLE)

Find parameters that maximize the score, or *likelihood*, of data

18

Mean/covariance MLE

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i \text{ in } k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Scalar case:

Cluster proportion MLE

R	G	B	Cluster
$x_4[1]$	$x_4[2]$	$x_4[3]$	1

R	G	B	Cluster
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

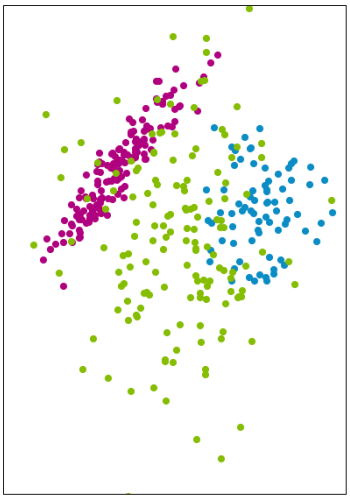
obs in cluster k

$$\hat{\pi}_k = \frac{N_k}{N}$$

total # of obs

True for general mixtures of i.i.d. data, not just Gaussian clusters

Part 2a summary



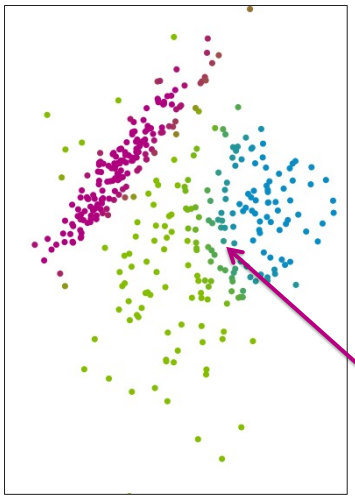
needed to compute soft assignments

Cluster parameters are simple to compute if we know the cluster assignments

But, we don't know these!

Part 2b:
What can we do with just soft assignments r_{ij}?

Estimating cluster parameters from soft assignments



Instead of having a full observation $\mathbf{x}_i$ in cluster k, just allocate a portion r_{ik}

$\mathbf{x}_i$ divided across all clusters, as determined by r_{ik}

Maximum likelihood estimation from soft assignments

Just like in boosting with weighted observations…

R	G	B	r_{i1}	r_{i2}	r_{i3}
$\mathbf{x}_1[1]$	$\mathbf{x}_1[2]$	$\mathbf{x}_1[3]$	0.30	0.18	0.52
$\mathbf{x}_2[1]$	$\mathbf{x}_2[2]$	$\mathbf{x}_2[3]$	0.01	0.26	0.73
$\mathbf{x}_3[1]$	$\mathbf{x}_3[2]$	$\mathbf{x}_3[3]$	0.002	0.008	0.99
$\mathbf{x}_4[1]$	$\mathbf{x}_4[2]$	$\mathbf{x}_4[3]$	0.75	0.10	0.15
$\mathbf{x}_5[1]$	$\mathbf{x}_5[2]$	$\mathbf{x}_5[3]$	0.05	0.93	0.02
$\mathbf{x}_6[1]$	$\mathbf{x}_6[2]$	$\mathbf{x}_6[3]$	0.13	0.86	0.01
Total weight in cluster: (effective # of obs)			**1.242**	**2.8**	**2.42**

52% chance this obs is in cluster 3

Maximum likelihood estimation from soft assignments

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30	0.18	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01	0.26	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002	0.008	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75	0.10	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05	0.93	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13	0.86	0.01

25

Maximum likelihood estimation from soft assignments

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30

R	G	B	Cluster 2 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.18

R	G	B	Cluster 3 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.01

26

Cluster-specific location/shape MLE

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

Total weight in cluster k = effective # obs

Compute cluster parameter estimates with weights on each row operation

MLE of cluster proportions $\hat{\pi}_k$

r_{i1}	r_{i2}	r_{i3}
0.30	0.18	0.52
0.01	0.26	0.73
0.002	0.008	0.99
0.75	0.10	0.15
0.05	0.93	0.02
0.13	0.86	0.01

Total weight in cluster:

1.242	2.8	2.42

Total weight in dataset: 6 ← # datapoints N

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

Total weight in cluster k = effective # obs

Estimate cluster proportions from relative weights

Defaults to hard assignment case when r_{ij} in {0,1}

Hard assignments have:

$$r_{ik} = \begin{cases} 1 & i \text{ in } k \\ 0 & \text{otherwise} \end{cases}$$

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0	0	1
$x_2[1]$	$x_2[2]$	$x_2[3]$	0	0	1
$x_3[1]$	$x_3[2]$	$x_3[3]$	0	0	1
$x_4[1]$	$x_4[2]$	$x_4[3]$	1	0	0
$x_5[1]$	$x_5[2]$	$x_5[3]$	0	1	0
$x_6[1]$	$x_6[2]$	$x_6[3]$	0	1	0
Total weight in cluster:			1	2	3

One-hot encoding of cluster assignment

29

Equating the estimates…

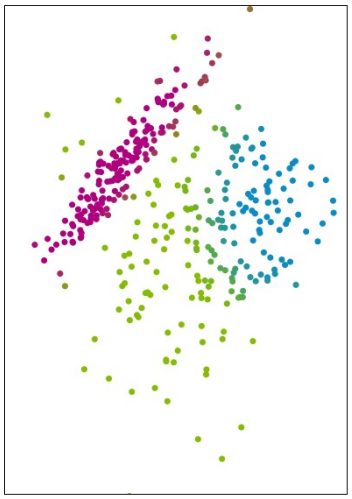
$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N} \leftarrow N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^{N} r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

30

Part 2b summary



Still straightforward to compute cluster parameter estimates from soft assignments

31

Expectation maximization (EM) for MoG model

32

Expectation maximization (EM) for MoG: An iterative algorithm

Motivates an iterative algorithm:

1. **E-step:** estimate cluster responsibilities given current parameter estimates

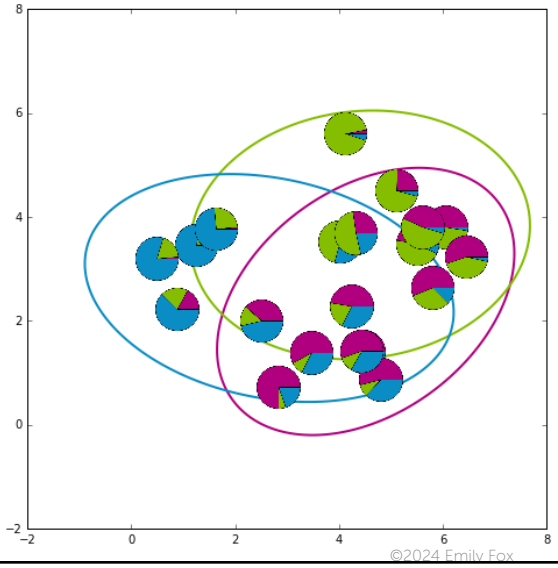
$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}$$

2. **M-step:** maximize likelihood over parameters given current responsibilities

$$\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k \mid \{\hat{r}_{ik}, x_i\}$$

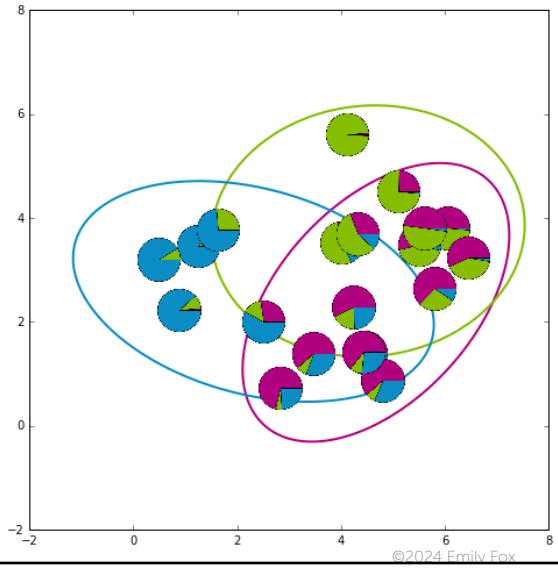
33

EM for mixtures of Gaussians in pictures – initialization



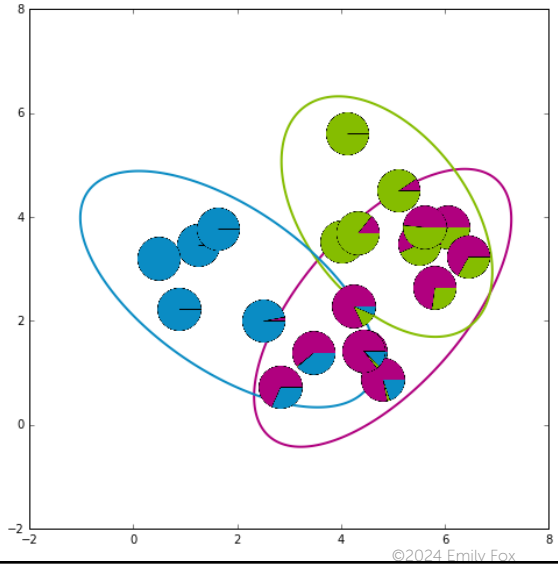
34

EM for mixtures of Gaussians in pictures – after 1st iteration



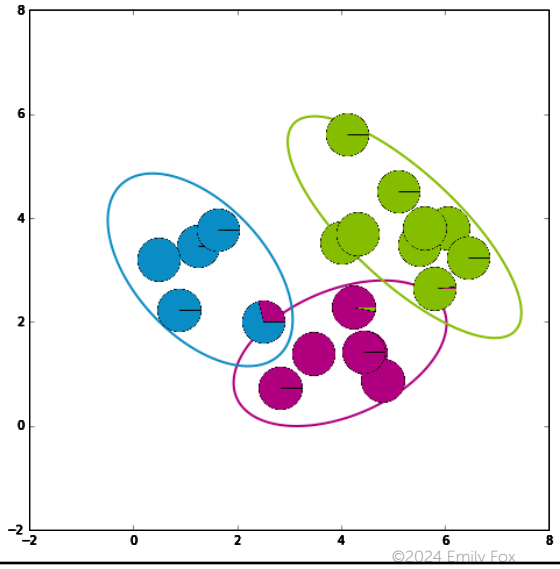
35

EM for mixtures of Gaussians in pictures – after 2nd iteration



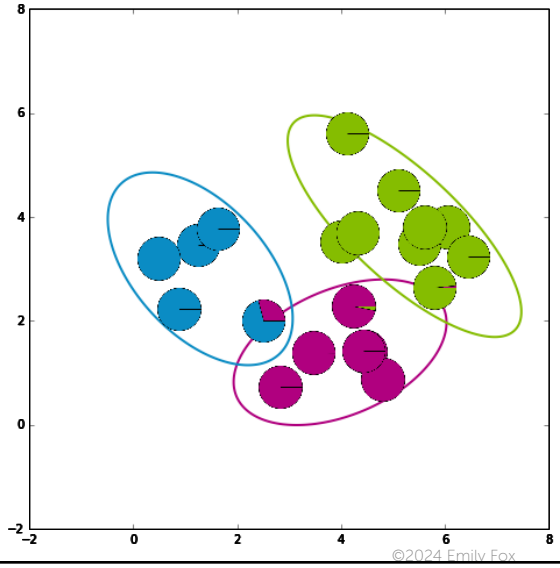
36

EM for mixtures of Gaussians in pictures – converged solution



37

EM for mixtures of Gaussians in pictures – replay



38

The nitty gritty of EM

39

EM more formally / generally

40

MLE of mixture model parameters

- Log likelihood

$$L_x(\theta) = \log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \log \sum_{z_i} p(x_i, z_i \mid \theta)$$

- Want MLE

$$\hat{\theta}^{MLE} = \arg\max_{\theta} L_x(\theta)$$

- Assume exponential family for $p(x_i, z_i \mid \theta)$
- Neither convex nor concave and local optima

41

Formalizing EM for mixtures of Gaussians

Complete data likelihood

$$p(z_{1:n}, x_{1:n} \mid \Theta = \{\pi_k, \mu_k, \Sigma_k\}) = \prod_{i=1}^{n} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

E-Step: Expected log-likelihood given previous parameters

$$U(\Theta \mid \Theta^{(j-1)}) = \mathbb{E}\left[\log p\left(z_{1:n}, x_{1:n} \mid \Theta\right) \mid x_{1:n}, \Theta^{(j-1)}\right]$$

M-Step: Maximize expected log-likelihood

$$\Theta^{(j)} = \arg\max_{\Theta} U\left(\Theta | \Theta^{(j-1)}\right)$$

EM Algorithm: Initialize $\Theta^{(0)}$; **iterate** over E and M steps **until convergence**

42

Example: Mixture models (again)

E-step: Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$

M-step: Compute $\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$

Consider $y_i = \{z_i, x_i\}$ i.i.d.

E-step: $p(x_i, z_i \mid \theta) = \pi_{z_i} p(x_i \mid \phi_{z_i}) = \prod_{k=1}^{K} (\pi_k p(x_i \mid \phi_k))^{\mathbb{I}(z_i=k)}$

$E_{q_t}[\log p(y \mid \theta)] = \sum_i E_{q_t}[\log p(x_i, z_i \mid \theta)] = \sum_i \sum_k E_{q_t}[\mathbb{I}(z_i=k) \log \pi_k] + \sum_i \sum_k E_{q_t}[\mathbb{I}(z_i=k) \log p(x_i \mid \phi_k)]$

$= \sum_i \sum_k \hat{r}_{ik} \log \pi_k + \sum_i \sum_k \hat{r}_{ik} \log p(x_i \mid \phi_k)$

M-step: max w.r.t. π_k, ϕ_k w/ $\hat{r}_{ik}$ fixed

E-step: compute resp. $\hat{r}_{ik}$ based on $\theta^{(t)}$

model params $= \theta = \{\pi_k, \phi_k\}$

43

Convergence and initialization of EM

44

Convergence of EM

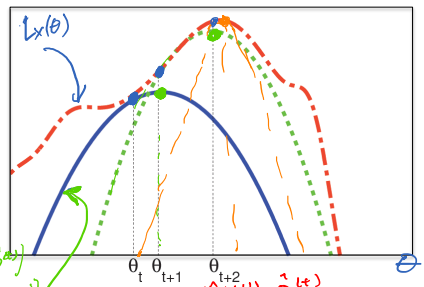
As we just saw…

- EM is a coordinate-ascent algorithm
  - Can equate E-and M-steps with alternating maximizations of an objective function
- Convergences to a local mode
- Assess convergence via (log) likelihood of data under current parameter and responsibility estimates

45

More formally…

Figure from KM textbook



$$L_x(\hat{\theta}^{(t+1)}) \triangleq \log p(x|\hat{\theta}^{(t+1)}) = \log \sum_z p(x,z|\hat{\theta}^{(t+1)})$$

$$= \log \sum_z p(z|x,\hat{\theta}^{(t)}) \frac{p(x,z|\hat{\theta}^{(t+1)})}{p(z|x,\hat{\theta}^{(t)})}$$

Jensen's inequality: $\log E[\,] \geq E[\log(\,)]$

$$\geq \sum_z p(z|x,\hat{\theta}^{(t)}) \log \frac{p(x,z|\hat{\theta}^{(t+1)})}{p(z|x,\hat{\theta}^{(t)})} \quad \text{ELBO}(\hat{\theta}^{(t+1)}, \hat{\theta}^{(t)})$$

$$\geq \sum_z p(z|x,\hat{\theta}^{(t)}) \log \frac{p(x,z|\hat{\theta}^{(t)})}{p(z|x,\hat{\theta}^{(t)})} \quad \text{ELBO}(\hat{\theta}^{(t)}, \hat{\theta}^{(t)})$$

$$= \sum_z p(z|x,\hat{\theta}^{(t)}) \log p(x|\hat{\theta}^{(t)})$$

$$= \log p(x|\hat{\theta}^{(t)}) \triangleq L_x(\hat{\theta}^{(t)})$$

$u(\hat{\theta}^{(t+1)}, \hat{\theta}^{(t)})$

$\sum_z p(z|x,\hat{\theta}^{(t)}) \log p(x,z|\hat{\theta}^{(t+1)})$

$\geq \sum_z p(z|x,\hat{\theta}^{(t)}) \log p(x,z|\hat{\theta}^{(t)})$ by M-step

$u(\hat{\theta}^{(t)}, \hat{\theta}^{(t)})$

$\text{ELBO}(\theta, \hat{\theta}^{(t)}) = \sum_z p(z|x,\hat{\theta}^{(t)}) \log \frac{p(x,z|\theta)}{p(z|x,\hat{\theta}^{(t)})}$

46

Initialization

- Many ways to initialize the EM algorithm
- Important for convergence rates & quality of local mode found
- Examples:
  - Choose K observations at random to define K "centroids". Assign other observations to nearest centriod to form initial parameter estimates.
  - Pick centers sequentially to provide good coverage of data like in k-means++
  - Initialize from k-means solution
  - Grow mixture model by splitting (and sometimes removing) clusters until K clusters are formed

47

Potential of vanilla EM to overfit

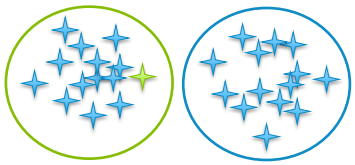
48

Overfitting of MLE

Maximizing likelihood can overfit to data

Imagine a K=2 example with one obs assigned to cluster 1 and others assigned to cluster 2

- What parameter values maximize likelihood?



Set center equal to point and shrink variance to 0

Likelihood goes to ∞ !

Overfitting in high dims

Doc-clustering example:
Imagine no doc assigned to cluster k has word w
(or all docs in cluster agree on count of word w)

Likelihood maximized by setting $\boldsymbol{\mu}_k[w] = \mathbf{x}_i[w]$ and $\sigma^2_{w,k} = 0$

Likelihood of any doc with different count on word w being in cluster k is 0!

Simple regularization of M-step for mixtures of Gaussians

Simple fix: Don't let variances → 0!

Add small amount to diagonal of covariance estimate

Alternatively, take Bayesian approach & place prior on parameters.

Similar idea, but all parameter estimates are "smoothed" via cluster pseudo-observations.

51

Formally relating k-means and EM for mixtures of Gaussians

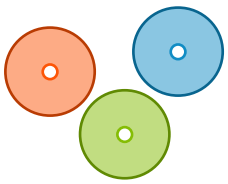
52

Relationship to k-means

Consider Gaussian mixture model with

$$\Sigma = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \ddots \\ & & & & \sigma^2 \end{pmatrix}$$

Spherically symmetric clusters



and let the variance parameter σ → 0

- Spherical clusters w/ equal variances, so relative likelihoods just fcn of dist to cluster center
- As variances→0, likelihood ratio becomes 0 or 1
- Responsibilities weigh in cluster proportions, but dominated by likelihood disparity

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \sigma^2 I)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \sigma^2 I)}$$

Datapoint gets fully assigned to nearest center, just as in k-means

Infinitesimally small variance EM = k-means

1. **E-step:** estimate cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \sigma^2 I)}{\sum_{j=1}^{K} \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \sigma^2 I)} \in \{0, 1\}$$

Infinitesimally small

Decision based on distance to nearest cluster center

2. **M-step:** maximize likelihood over parameters given current responsibilities (hard assignments!)

$$\hat{\pi}_k, \hat{\mu}_k \mid \{\hat{r}_{ik}, x_i\}$$

Summary for the EM algorithm

55

What you can do now…

- Estimate soft assignments (responsibilities) given mixture model parameters
- Solve maximum likelihood parameter estimation using soft assignments (weighted data)
- Implement an EM algorithm for inferring soft assignments and cluster parameters
  - Determine an initialization strategy
  - Implement a variant that helps avoid overfitting issues
- Compare and contrast with k-means
  - Soft vs. hard assignments
  - k-means as a limiting special case of EM for mixtures of Gaussians

56