WEEK 7.2

GAUSSIAN MIXTURE MODELS (GMM)

→ In soft clustering, instances have probability that they came from the clusters

→ Mixture models are a probabilistically-grounded way of doing soft clustering

↳ Each cluster corresponds to a probability distribution in the d-dimensional space and the points are samples from that prob. distribution

Can be Gaussian (if data is real valued) or multinomial (if data is discrete) → GENERATIVE

→ For GMM, the parameters are mean μ and covariance Σ which are not known in advance

So, we use...

EXPECTATION-MAXIMIZATION ALGO.

1) Start w/ 2 randomly placed Gaussians (μ_a, σ_a^2) and (μ_b, σ_b^2)

2) $\boxed{\text{E-step}}$: For each data point, calculate $P(b|x_i)$ and $P(a|x_i)$

   ↳ How likely is it that each data pt. was generated by each mixture?

3) $\boxed{\text{M-step}}$: Estimate the parameters of each of the mixtures, given the probabilities of each data pt. having been generated by that mixture.

4) Update the parameters and repeat ② and ③ until convergence

$$P(x_i | b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left\{-\frac{(x_i - \mu_i)^2}{2\sigma_b^2}\right\}$$

Bayes rule

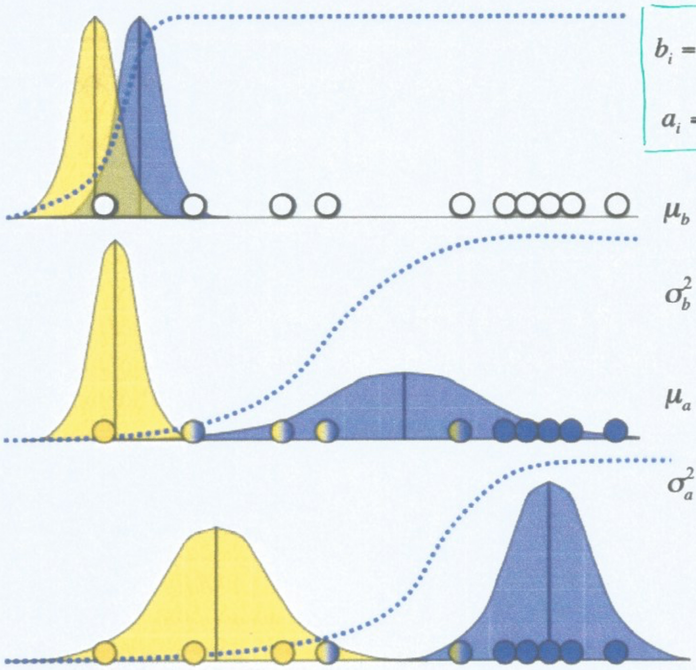
$$b_i = P(b | x_i) = \frac{P(x_i | b)P(b)}{P(x_i | b)P(b) + P(x_i | a)P(a)}$$

$$a_i = P(a | x_i) = 1 - b_i$$

1D - EXAMPLE

1st iteration

2nd iteration



Model Parameters

$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1 - \mu_b)^2 + \ldots + b_n(x_n - \mu_b)^2}{b_1 + b_2 + \ldots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_n}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1 - \mu_a)^2 + \ldots + a_n(x_n - \mu_a)^2}{a_1 + a_2 + \ldots + a_n}$$

could also estimate priors:

$P(b) = (b_1 + b_2 + \ldots b_n) / n$

$P(a) = 1 - P(b)$

FOR ARBITRARY DIM.

- Data with D attributes, from Gaussian sources $c_1 \dots c_k$
  - how typical is $\mathbf{x}_i$ under source **c** $\quad P(\vec{x}_i \mid c) = \frac{1}{\sqrt{2\pi|\Sigma_c|}} \exp\{-\tfrac{1}{2}\underbrace{(\vec{x}_i - \vec{\mu}_c)^T \Sigma_c^{-1}(\vec{x}_i - \vec{\mu}_c)}_{\sum_a \sum_b (x_{ia} - \mu_{ca})[\Sigma_c^{-1}]_{ab}(x_{ib} - \mu_{cb})}\}$
  - how likely that $\mathbf{x}_i$ came from **c** $\quad P(c \mid \vec{x}_i) = \frac{P(\vec{x}_i \mid c)P(c)}{\sum_{c=1}^{k} P(\vec{x}_i \mid c)P(c)}$
  - how important is $\mathbf{x}_i$ for source **c**: $w_{i,c} = P(c \mid \vec{x}_i) / \left(P(c \mid \vec{x}_1) + \dots + P(c \mid \vec{x}_n)\right)$
  - mean of attribute **a** in items assigned to **c**: $\mu_{ca} = w_{c1}x_{1a} + \dots + w_{cn}x_{na}$
  - covariance of **a** and **b** in items from **c**: $\Sigma_{cab} = \sum_{i=1}^{n} w_{ci}(x_{ia} - \mu_{ca})(x_{ib} - \mu_{cb})$
  - prior: how many items assigned to **c**: $P(c) = \frac{1}{n}\left(P(c \mid \vec{x}_1) + \dots + P(c \mid \vec{x}_n)\right)$

HOW MANY GAUSSIANS DO YOU NEED?

$$\boxed{\text{Likelihood } L = \log P(x_1, \dots, x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i \mid k)\, P(k)}$$

→ Pick K that makes L as large as possible? similar problem w/ K-means

(X) ↳ NO! Data pt. will end up having its own 'source' $(K = n)$

→ Split points into training set T and validation set V

(meh) ↳ For each K, fit params. of T and measure likelihood of V

↳ May still end up w/ $K = n$

→ We can pick the 'simplest' of all models that fit. $\quad p$... number of params. (how 'simple' is the model)

(✓)

Bayes Information Criterion (BIC)	Akaike Information Criterion (AIC)
$\max_p \{ L - \frac{1}{2} p \log n \}$	$\min_p \{ 2p - L \}$

how well the model fits the data w/ the complexity of the model