Mixture models

- Recall types of clustering methods
  - hard clustering: clusters do not overlap
    - element either belongs to cluster or it does not
  - soft clustering: clusters may overlap
    - strength of association between clusters and instances

- Mixture models
  - probabilistically-grounded way of doing soft clustering
    - each cluster: a generative model (Gaussian or multinomial)
    - parameters (e.g. mean/covariance are unknown)

- Expectation Maximization (EM) algorithm
  - automatically discover all parameters for the K “sources”

Gaussian Mixture Model

- Data with D attributes, from Gaussian sources $c_1 \ldots c_L$
  - how typical is $x_i$ under source $c_i$?
  - how likely that $x_i$ came from $c_i$?
  - how important is $x_i$ for source $c_i$: $w_{i,c} = \frac{P(x_i | c_i) P(c_i)}{\sum_{c} P(x_i | c) P(c)}$
  - mean of attribute $a$ in items assigned to $c$: $\mu_a = \sum w_{i,c} x_{i,a}$
  - covariance of $a$ and $b$ in items from $c$: $\Sigma_{ab} = \sum w_{i,c} (x_{i,a} - \mu_{a,c})(x_{i,b} - \mu_{b,c})$
  - prior: how many items assigned to $c$: $P(c) = \sum_c P(x_i | c_i) P(c)$

How to pick $K$?

- Probabilistic model: $L = \log P(x | \theta) = \sum_{i=1}^{N} \log P(x_i | \theta) P(\theta)$
  - tries to “fit” the data (maximize likelihood)
- Pick $K$ that makes $L$ as large as possible?
  - $K = n$: each data point has its own “source”
  - may not work well for new data points
- Split points into training set $T$ and validation set $V$
  - for each $K$: fit parameters of $T$, measure likelihood of $V$
  - sometimes still best when $K = n$
- Occam’s razor: pick “simplest” of all models that fit
  - Bayes Inf. Criterion (BIC): $\max_L L - \frac{1}{2} p \log n$
  - Akaike Inf. Criterion (AIC): $\min_L \{2 p - L\}$

Summary

- Walked through 1-d version
  - works for higher dimensions
  - $d$-dimensional Gaussian, can be non-spherical
  - works for discrete data (lex)
  - $d$-dimensional multinomial distributions prior
- Maximum likelihood of the data: $P(x_{1:n} | \theta) = \prod_{i=1}^{N} P(x_i | \theta)$
- Similar to K-means
  - sensitive to starting point, converges to a local maximum
  - convergence: when change in $P(x_{1:n} | \theta)$ is sufficiently small
  - cannot discover K (likelihood keeps growing with $K$)
- Different from K-means
  - soft clustering: instance can come from multiple “clusters”
  - co-variance: notion of “distance” changes over time
- How can you make GMM = K-means?