Finite Mixture Models

Data $D = \{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$, $x^{(i)} \in \mathbb{R}^d$.

Assumption: points are drawn in an IID fashion from an underlying density function $p(x)$. Furthermore we assume that $p(x)$ is defined as a finite mixture model with K components.

$$p(x|\theta) = \sum_{k=1}^{K} \alpha_k \, p_k(x | z_k, \theta_k)$$

- $p_k(x|z_k, \theta_k)$ are mixture components. Each is density or distribution defined over $p(x)$ with parameters θ_k.
- $\alpha_k = p(z_k)$ are the mixture weights. These represent the probability that randomly generated x was generated by component k. $\sum_{k=1}^{K} \alpha_k = 1$
- $z = (z_1 \cdots z_k)$ is a vector of k binary indicator variables. Only one of these can be non-zero i.e., these are mutually exclusive and exhaustive.

Parameters for a finite mixture model with K components $\theta = \{\alpha_1, \cdots, \alpha_K, \theta_1, \cdots, \theta_K\}$

* Membership Weights

We can compute the membership weight of data point $x^{(i)}$ in component k given parameter θ as

$$w_{ik} = p(z_{ik}=1 | x^{(i)}, \theta) = \frac{p_k(x^{(i)} | z_k, \theta_k) \cdot \alpha_k}{\sum_{m=1}^{K} p_m(x^{(i)} | z_m, \theta_m) \, \alpha_m}$$

↓ normalization constant

$$p(A|B)\,p(B) = p(B|A)\,p(A)$$
$$p(A|B) \propto p(B|A)\,p(A)$$

The membership weights reflect our uncertainty about which of K components generated vector $x^{(i)}$. We continue to assume that $x^{(i)}$ is generated from one of these components. So these weights doesn't assume any mixing of components.

Gaussian Mixture Model

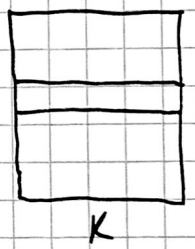
For $x \in R^d$ we can define a Gaussian mixture model by making each of the K components a Gaussian density with parameters μ_k and Σ_k. Each component is a multivariate Gaussian density.

$$P_k(x|\theta_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$$

With its own parameters $\theta_k = \{\mu_k, \Sigma_k\}$.

The EM algorithm for Gaussian mixture model

* E-Step: Given the current parameter values θ, compute w_{ik} for all data points and all mixture components. Note that $\sum_{k=1}^{K} w_{ik} = 1$. So we get $N \times K$ matrix of membership weights.



$\rightarrow 1$ Each row sums to 1.

* M-Step: Use membership weights and data to calculate new parameter values. The effective number of data points assigned to component k are

$$\alpha_k^{new} = \frac{N_k}{N}, \text{ where } N_k = \sum_{i=1}^{N} w_{ik}$$

We can similarly compute the updated mean.

$$\mu_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot x_i^{(i)} \quad 1 \leq k \leq K$$

and the updated covariance

$$\Sigma_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \left(x^{(i)} - \mu_k^{new}\right)\left(x^{(i)} - \mu_k^{new}\right)^t$$

Both mean and covariance are computed similar to how we would compute these quantities empirically. Except ~~the weights are computed~~ each data point is weighted.

Likelihood of Mixture Model

Under the iid assumption

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \alpha_k P_k(x | z_k, \theta_k)$$

Log likelihood

$$\log L = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \alpha_k P_k(x | z_k, \theta_k)$$

Probabilistic LSA (pLSA)

Documents $D = \{d_1, \ldots, d_N\}$

Vocabulary $W = \{w_1, \ldots, w_m\}$

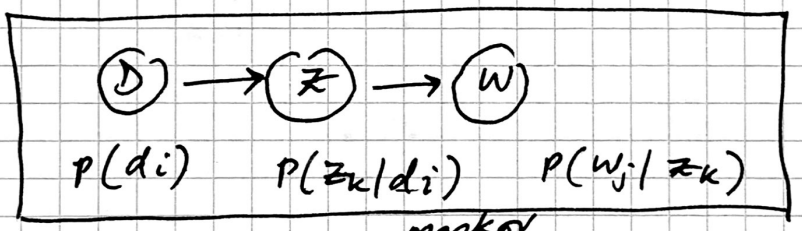
Co-occurence table of counts $N = (n(d_i, w_j))_{ij}$

Associate with each observation $\langle d_i, w_j \rangle$ an un-observed class variable $z_k \in \{z_1, \ldots, z_k\}$.

* Generative model for an observation pair $\langle d_i, w_j \rangle$
  1. Select a document d_i with probability $p(d_i)$
  2. Pick a latent class z_k with probability $p(z_k | d_i)$
  3. Generate a word w_j with probability $p(w_j | z_k)$

We can describe the above process using the following expression

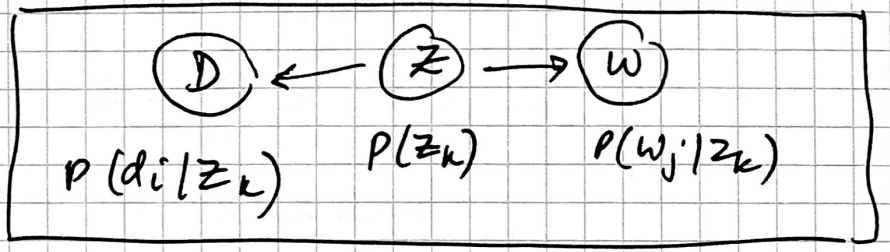
$$p(d_i, w_j) = \sum_{k=1}^{K} p(d_i)\, p(z_k | d_i)\, p(w_j | z_k) \quad \rightarrow ①$$



BN - Bayesian Network Model

The expression in ① is (Markov) equivalent to the following

$$p(d_i, w_j) = \sum_{k=1}^{K} p(z_k)\, p(d_i | z_k)\, p(w_j | z_k)$$



Implicit conditional independence assumption: d_i and w_j are independent conditioned on the state of the associated latent variable z_k (sometimes called aspect z_k).

Aside: $P(d_i, w_j) = \sum_{k=1}^{K} P(d_i)\, P(z_k|d_i)\, P(w_j|z_k)$ (?)

$= P(d_i)\, P(w_j|d_i)$ (Bayes Rule)

$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j, z_k|d_i)$ (marginalization)

$= \sum_{k=1}^{K} P(w_j|z_k)\, P(z_k|d_i)$ (Bayes Rule)

$P(d_i, w_j) = P(d_i|z_k)\, P(w_j|z_k)\, P(z_k)$

? $\sum_{k=1}^{K} P(d_i, w_j|z_k)\, P(z_k) = P(d_i, w_j)$

$P(d_i, w_j) = \sum_{k=1}^{K} P(d_i, w_j|z_k)\, P(z_k)$

$= \sum_{k=1}^{K} P(d_i|z_k)\, P(w_j|z_k)\, P(z_k)$ (conditional independence of d_i and w_j given z_k.)

---

pLSA learns the unobservable probabilities in a maximum likelihood fashion, learning the hidden aspect in the process.

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(d_i, w_j)^{n(d_i, w_j)}$$

$n(d_i, w_j)$ ← # word w_j appears in document d_j.

And now log likelihood

$$\log L = \sum_{i=1}^{N} \sum_{j=1}^{N} n(d_i, w_j) \log P(d_i, w_j)$$

Aside: Some log identities. $\log x^n = n \log x$

$\log(xy) = \log x + \log y$

$\log\left(\frac{x}{y}\right) = \log x - \log y$

$$\log L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log\left[\sum_{k=1}^{K} P(d_i)\, P(z_k|d_i)\, P(w_j|z_k)\right]$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log\left[P(d_i) \sum_{k=1}^{K} P(z_k|d_i)\, P(w_j|z_k)\right]$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\left[\log P(d_i) + \log\left\{\sum_{k=1}^{K} P(z_k|d_i)\, P(w_j|z_k)\right\}\right]$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log P(d_i) + n(d_i, w_j) \log\left\{\sum_{k=1}^{K} P(z_k|d_i)\, P(w_j|z_k)\right\}$$

$$= \sum_{i=1}^{N} n(d_i) \log P(d_i) + \frac{n(d_i)}{n(d_i)} \sum_{j=1}^{M} n(d_i, w_j) \log\left\{\sum_{k=1}^{K} P(z_k|d_i)\, P(w_j|z_k)\right\}$$

$$= \sum_{i=1}^{N} n(d_i)\left[\log P(d_i) + \sum_{j=1}^{M} \frac{n(d_i, w_j)}{n(d_i)} \log\left\{\sum_{k=1}^{K} P(z_k|d_i)\, P(w_j|z_k)\right\}\right]$$

Expectation - Maximization (EM)

EM consists of two steps: ① E-step calculates the posterior probabilities for latent variables given the observations by using the current estimates of the parameters. ② M-step update the parameters such that data log-likelihood ($\log L$) increases using the posterior probabilities.

E-step:

$$P(z_k|d_i, w_j) = \frac{P(w_j, z_k|d_i)}{P(w_j|d_i)}$$

$$= \frac{P(w_j|z_k, d_i)\, P(z_k|d_i)}{P(w_j|d_i)}$$

$$= \frac{P(w_j|z_k)\, P(z_k|d_i)}{\sum_{\ell=1}^{K} P(w_j|z_\ell)\, P(z_\ell|d_i)}$$

M-step: Maximize $\log L$. $P(d_i)$, $n(d_i)$ and $n(d_i, w_j)$ can be estimated from the data.

In order to maximize log L.

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)\, P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)\, P(z_k|d_i,w_m)}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)\, P(z_k|d_i,w_j)}{n(d_i)}$$

E-step and M-step are applied until convergence condition is met.

Applications:

- Topic detection and tracking corpus
- Image classification (images are represented as a bag-of-visual words)
- Action classification.