Mixture Models

Base form	Hierarchical representation
$X \sim f$ $f(x) = \sum_{k=1}^{K} \omega_k g_k(x\|\theta_k)$ with $\sum_{k=1}^{K} \omega_k = 1$	$X\|C \sim g_c$ $C \sim Pr(C=k) = \omega_k$ $p(x) = \sum_{k=1}^{K} p(x\|C=k)p(C=k)$

Observed data likelihood	Complete data likelihood
$L(X\|\omega,\theta) = \prod_{i=1}^{n}\sum_{k=1}^{K} \omega_k g_k(x_i\|\theta_k)$	$L(X\|\omega,\theta,c) = \prod_{i=1}^{n}\prod_{k=1}^{K} [\omega_k g_k(x_i\|\theta_k)]^{\mathbb{1}_{c_i=k}}$

Probability of a data point x_i being generated by a component g_k

$$Pr(c_i = k|\omega,\theta) = \frac{\omega_k g_k(x_i|\theta_k)}{\sum_l \omega_l g_l(x_i|\theta_l)} =: v_{i,k}(\omega,\theta)$$

Maximum likelihood framework	Bayesian framework

Model fitting

Expectation-maximization algorithm

output: maximum likelihood estimates of ω and θ

procedure:

$\hat{\omega}, \hat{\theta} \leftarrow$ initial values

repeat

$\hat{\omega}, \hat{\theta} \leftarrow \operatorname{argmax}_{\omega,\theta} Q(\omega,\theta|\hat{\omega},\hat{\theta})$

until convergence

where

$$Q(\omega,\theta|\hat{\omega},\hat{\theta}) = E_{c|X,\hat{\omega},\hat{\theta}}[\log L(X|\omega,\theta,c)]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} v_{i,k}(\hat{\omega},\hat{\theta})\,[\log\omega_k + \log g_k(x_i|\theta_k)]$$

MCMC algorithm/Gibbs sampling

output: a sample from the posterior distribution $p(\omega,\theta,c|X)$

procedure:

set priors $p(\omega)$ and $p(\theta)$ on ω and θ

convenient: $(\omega_1,\ldots,\omega_K) \sim \text{Dirichlet}(a_1,\ldots,a_K)$

repeat

$\omega \hookleftarrow \text{Dirichlet}(a_1^*,\ldots,a_K^*)$,

with $a_k^* = a_k - 1 + \sum_{i=1}^{n} \mathbb{1}_{c_i=k}$

$c_i \hookleftarrow p(c_i = k|\ldots) = v_{i,k}(\omega,\theta)$

$\theta_k \hookleftarrow p(\theta_k|\ldots) \propto \left[\prod_{\{i:\,c_i=k\}} g_k(x_i|\theta_k)\right] p_k(\theta_k)$

Determining the number of components

For each candidate model $\mathcal{M}$ compute its Bayesian information criterion value

$$\text{BIC}(\mathcal{M}) = -2\log L(\hat{\eta}_{\mathcal{M}}) + r_{\mathcal{M}}\log n,$$

where

$L(\hat{\eta}_{\mathcal{M}}) = L(X|\hat{\omega}_{\mathcal{M}},\hat{\theta}_{\mathcal{M}})$,

$r_{\mathcal{M}}$ – # independent parameters in the model.

Select the model with the lowest BIC value.

With

K – upper bound on # components,

K^* – best guess for actual # components $\overline{K}$,

solve

$K^* = \alpha\log\frac{n+\alpha-1}{\alpha}$

for α.

Set the prior $p(\omega) \sim \text{Dirichlet}(\frac{\alpha}{K},\ldots,\frac{\alpha}{K})$.

Fit a model with K components.

Posterior distribution for $\overline{K}$ obtained by counting the number of unique values assumed by indicators c_i per iteration of Gibbs sampling.

Classification with mixture models

Unsupervised: fit a model, classify as $c_i = \operatorname{argmax}_k v_{i,k}(\hat{\omega},\hat{\theta})$

Semi-supervised: fit a model to the whole dataset, keeping $v_{i,k}$ constant on the training data; for the test data, classify as $c_i = \operatorname{argmax}_k v_{i,k}(\hat{\omega},\hat{\theta})$

Supervised: estimate $\hat{\theta}_k$ by fitting g_k to training data from class k; set $\hat{\omega}_k = (\#\text{obs. in class } k)/n$; for the test data, classify as $c_i = \operatorname{argmax}_k v_{i,k}(\hat{\omega},\hat{\theta})$