CAP 6617 ADVANCED MACHINE LEARNING

NAME: SANKET KUMAR UF ID: 4513-3352

THE INFINITE GAUSSIAN MIXTURE MODEL

AIM: In this paper, a **Markov Chain Monte Carlo (MCMC**) implementation of a hierarchical infinite Gaussian mixture model is presented.

- Monte Carlo methods are so important because they are often relatively easy to understand and implement, yet are powerful enough to enable us to compute relevant statistical summaries even when fitting highly structured models. The finite Gaussian mixture model with k components can be written as $\sum_{j=1}^{k} \pi_j N(\mu_j, s_j^{-1})$…(1) where μ_j are the means, s_j are precisions(inverse variances), π_j are the mixing proportions and N is (normalized) Gaussian with specified mean and variance. This paper also introduces stochastic indicator variables c_i that tells which class has generated the observation. A method for Markov Chain Monte Carlo(MCMC) is **Gibbs sampling**. For Gibbs sampling, the priors and conditional distributions on component parameters and hyper parameters will be required.
- The component means and precisions are given Gaussian…(2) and Gamma…(3) priors respectively, hyper parameters for mean and precision distributions are given Normal, Gamma…(4) and inverse Gamma, Gamma…(5) priors respectively. The shape parameter β for mean is set to unity so that the distribution is very broad and vague. The conditional posterior distributions for the means are obtained by multiplying (1) and (2) to get:

$$p(\mu_j|\mathbf{c},\mathbf{y},s_j,\lambda,r) \sim \mathcal{N}\Big(\frac{\bar{y}_j n_j s_j + \lambda r}{n_j s_j + r}, \frac{1}{n_j s_j + r}\Big), \qquad \bar{y}_j = \frac{1}{n_j}\sum_{i:c_i=j} y_i$$

  where n_j is the number of observations belonging to class j. The conditional posterior distributions for the precisions are obtained by multiplying (1) and (3) to get:

$$p(s_j|\mathbf{c},\mathbf{y},\mu_j,\beta,w) \sim \mathcal{G}\Big(\beta + n_j, \big[\frac{1}{\beta + n_j}\big(w\beta + \sum_{i:c_i=j}(y_i - \mu_j)^2\big)\big]^{-1}\Big)$$

- Likewise, the conditional posterior distributions for hyper parameters can be equated. The mixing proportions are given a symmetric Dirichlet(multivariate beta) prior with concentration parameter α/k. Mixing proportions are positive and sum to 1. Using Dirichlet integral in predictive distributions of the form $p(y|\mathbf{y}) = \int p(y|\Phi)p(\Phi|\mathbf{y})d\Phi$, we may integrate out the mixing proportions and write the prior directly in terms of the indicators as:

$$p(c_1,\ldots,c_k|\alpha) = \int p(c_1,\ldots,c_k|\pi_1,\ldots,\pi_k)p(\pi_1,\ldots,\pi_k)d\pi_1\cdots d\pi_k$$

- The above integral allows us to work directly with finite number of indicator variables, rather than infinite number of mixing proportions. For using Gibbs sampling for indicators c_i, the conditional prior for a single indicator given all the others can be obtained from the above equations, such that –i indicated all indexes except i:

$$p(c_i = j|\mathbf{c}_{-i},\alpha) = \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha}$$

- The concentration parameter is given an inverse Gamma prior and likewise the conditional posterior can be obtained by multiplying the prior and the integral above. The conditional posterior for α only depends on n and k and not on "how" the observations are distributed among components.
- With $k \to \infty$, for all model variables except indicators, conditional posteriors are obtained by substituting k in the equations previously derived for finite model with the number of classes that have data associated with them. For indicators, the conditional prior is given:

$$\begin{aligned} &\text{components where } n_{-i,j} > 0: && p(c_i = j|\mathbf{c}_{-i},\alpha) = \frac{n_{-i,j}}{n-1+\alpha} \\ &\text{all other components combined:} && p(c_i \neq c_{i'} \text{ for all } i' \neq i|\mathbf{c}_{-i},\alpha) = \frac{\alpha}{n-1+\alpha} \end{aligned}$$

- The above equations show that greater the n, greater are the chances of joining the same component(rich gets richer). The above equation can be combined with (1) to get conditional posteriors for the indicators. Now that all classes have parameters associated with them, we can easily evaluate their likelihoods and priors. The likelihood pertaining to the currently unrepresented class(having no parameters) is obtained through integration over the prior distribution for these. We need not differentiate between infinitely many unrepresented classes, since their parameter distributions are identical. Since this Monte Carlo estimate is unbiased, the resulting chain will sample from exactly the desired distribution. For multivariate observations, the means and precisions become vectors and covariance matrices respectively. The advantages of infinite mixture model are: 1) delimiting the number of classes 2) number of represented classes is automatically determined 3) avoids local minima 4) simpler to handle.