Mixture Models and EM Algorithm

Slide 1:

In statistics, a mixture model is a probabilistic model used for representing and confirming the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which it belongs.

**Slide 2:**

How do we know how to fit the observed data? A hint that allows us to suspect that data might follow a mixture model is that it looks multimodal. Where there is more than one "peak" in the distribution of data. The bottom left image is trying to fit a multimodal distribution with a unimodal model. This is usually a very poor fit. Another more intelligent way to model a multimodal distribution would be to assume that it is generated by multiple unimodal distributions and fit it accordingly. Generally, the most common distribution in modelling unimodal data is the normal distribution. Thus, the normal mixture model will be the one used to represent normally distributed subpopulations within an overall population.

**Slide 3:**

The general case of the finite mixture models is presented here. Formally, let Y be a random sample where each Yi is a continuous, p- dimensional random vector belonging to a heterogenous population. It can be useful to denote the finite mixture model as the integral of the kernel function.

Where k is the number of components. Phi is the vector of unknowns. fj (y; θj ) are the component densities corresponding to a parametric family of distributions with unknown parameters θj. We see the kernel function again. And psi of theta is the mixing distributions Alpha j are the mixing proportions with the property that they must all sum to 1.

**Slide 4:**

Here we have the notation specific to the Gaussian mixture model. We see that the component densities are now those of the parametric family of normal distributions instead of just an arbitrary notation. We also take the likelihood and the log likelihood for we will need these later to determine and estimate our unknown parameters. Recall that In statistics, a likelihood function (often simply a likelihood) is function of parameters theta within the parameter space big theta that describes the probability of obtaining the observed data y

**Slide 5:**

Contrary to most intuition, the Gaussian mixture models are the ones with the most problematic and undesired mathematical properties. The likelihood functions are unbounded based on a set of random samples unless we assume the component variances. The models may also fail to fulfil the regularity conditions including the finiteness of the Fisher information.

Observing the log likelihood, it is clear that and setting μ1 = Y1 while letting σ1 → 0. Keeping the other parameters fixed, ln(φ) → ∞ and is thus unbounded. The unbounded likelihood prevents a straightforward implementation of maximum likelihood estimation. Which is a method of estimating the unknown parameters of our statistical model.

We also observe the score function with respect to alpha. If the variance of the score, ie the fisher information is not finite at alpha =0 the asymptotic results that rely on the finiteness of the matrix must be entirely re-examined.

Lets assume a two component GMM: Let Y = (Y1,...,Yn) be a random sample from a two-component normal mixture model such that each Yi is iid to this NMM for i = 1, ..., n. We then see that this entry of the information matrix tends to infinity when the first component’s variance is > 2.

**Slide6**:

Lets Assume that Y1 to Yn are an iid to the parametric family f with theta in the parametric space of real numbers with a certain dimension. This is how the loglikelihood and the maximum likelihood estimator theta hat are defined.

The two key properties of the MLE are its consistency and its normally asymptotic

An estimator is considered consistent if, as the number of observations n increases towards infinity, the sequence of estimates converges toward the true value of the parameter. If this is satisfied then it follows that the MLE is asymptotically normal.

**Slide 7:**

Before we even dive into the EM algorithm we want to introduce some essential components which we will use. Let Zi be a categorical random variable. Therefore, Zi is a k-dimensional vector that takes on the value of one at position j when Yi belongs to sub-population j for j = 1-k.it follows that zi follows a multinomial dis w 1 trial and k categories.

We again see the incomplete data likelihood and its corresponding log likelikihood. Followed by the complete data likelihoods.

Recapping what we know: Define the maximum likelihood estimate θ hat n of the vector of unknown parameters as the global maximizer of φ. Then, if the parameter space is compact and the regularity conditions hold then the likelihood equation is almost surely maximized in a neighbourhood of the specified parameter space for φ.

However, the likelihood equation can be unbounded as we have seen, which contradicts the regularity conditions. Thus, θˆn does not exist as a global maximizer of the likelihood in these cases.

SO WHAT DO WE DO??

**Slide8:**

We now turn to the subject of ultimate interest, which is a particular iterative procedure for numerically approximating maximum-likelihood estimates for mixture density problems. This procedure is known as the EM algorithm.

The EM algorithm consists of two major steps: the expectation step followed by the maximization step.

**Slide 9:**

The expectation is with respect to the underlying variables. We compute the expected value of the x data using the estimate of the parameters and conditioned upon the observed data. This step also assigns sampling weights corresponding to the probability that an observation belongs to a certain subpopulation.

**Slide 10:**

maximization step: Given the values you computed in the last step and treating them as observed data, estimate new values for parameters that maximize our previously defined variant of the likelihood function. As it iterates, it is clear that the log likelihood of incomplete data is bounded, and the sequence of iterations converges to its local maximum.

The updated parameter estimates (αˆ(t+1),θˆ(t+1)) are then computed by maximizing the two sums individually. This is what allows the algorithm to determine the maximum likelihood estimates in a more straightforward manner than the rudimentary procedures.

The updated mixing proportions and corresponding desired parameters can now be obtained by direct calculation.