Abstract:

In nonparametric Bayesian mixture models we assume that the data is an iid sample from an infinite mixture of component distributions; the mixture weights and the parameters of the components are random. This approach can be applied to cluster analysis; in this context it is convenient to think that we have a prior distribution on the space of all possible partitions of data (e.g. the Chineses Restaurant Process) and the likelihood is obtained by marginalising the parameter out of component distributions. This can be transformed into the posterior on the space of possible clusterings; as is rather common in the Bayesian setting, this posterior is known up to intractable normalising constant. During the talk we are going to present some properties of the MAP partition - the one that maximises the posterior probability. Firstly we will describe the results of (1) that concern the Chinese Restaurant prior on partitions and the component distributions being multivariate Gaussian with unknown mean but known (and the same for all components) covariance. Those findings concern the geometry of the MAP partition (we prove that in this case the MAP clusters are linearly separated from each other) and its asymptotic behaviour. Then we will show generalisations of some of these results to general infinite mixture models, when the component measures and the prior on their parameters belong to conjugate exponential families. Finally we present a 'score function' for clusterings, related to our analysis. This score function can be used to choose among clustering propositions suggested by more computationally efficient algorithms, like K-means.

Hope to see you there.

Behnaz P.