Gaussian mixture model (GMM)

Motivation

Where approaches such as linear regression and sub-space models are based on reducing the dimensionality of a signal to capture the essential information in a signal, in many cases we want to model the full range of possible signals. For that purpose we can design models the *statistical distribution* of the signal. For example, it is possible to model a signal as a Gaussian process, where every observation has a (multivariate) Gaussian (normal) distribution.

Speech signals however feature much more structure than simple Gaussian processes. For example voiced signals are very different from unvoiced signals, and within both voiced and unvoiced signals we have a multitude of distinct groups of utterances whose statistical characteristics are clearly different. Modelling them all with a Gaussian process would ignore such structures and the model would therefore be inefficient.

Mixture models is a type of models, where we assume that the signal under study consists of several distinct classes, where each class has its own unique statistical model. That is, for example the statistics of voiced sounds is clearly different from those of unvoiced sounds. We model each class with its own distribution and their joint distribution is the weighted sum of the class distributions. The weights of each distribution correspond to the frequency with which they appear in the signal. So if unvoiced signals would in some hypothetical language constitute 30% of all speech sounds, then the weight of the unvoiced class would be 0.3.

The most typical mixture model structure uses Gaussian (normal) distributions for each of the classes, so that the whole model is known as a *Gaussian mixture model* (GMM). Depending on application the class-distributions can obviously take other forms than Gaussian, for example a Beta mixture model could be used if the individual classes follow the Beta distribution. In this document we however focus on Gaussian mixture models because it is most common among mixture models and demonstrates application in an accessible way.

Model definition

The multivariate normal distribution for a variable x is defined as

$$ f\left(x;\Sigma,\mu\right) = \frac{1}{\sqrt{\left(2\pi\right)^N |\Sigma|}} \exp\left[-\frac12 (x-\mu)^T \Sigma^{-1}(x-\mu)\right], $$

where Σ and μ are the covariance and mean of the process, respectively, with N dimensions. In other words, this is the familiar Gaussian process for vectors x.

Suppose then that we have K classes in the signal, where each class has its own covariance and mean Σ_k and $\mu_k.$ The *Gaussian mixture model* is then defined as

$$ \boxed{f\left(x\right) = \sum_{k=1}^K \alpha_k f\left(x; \Sigma_k, \mu_k\right),} $$

where the weights α_k add up to unity $\sum_{k=1}^K \alpha_k=1.$

Applications

- In recognition/classification applications, we can, for example, model a system which has two distinct states (like speech and noise) and train a GMM with mixture components matching those states. When receiving a microphone signal, we can then determine the likelihood of each mixture component and thus obtain the likelihood that the signal is speech or noise.
- In transmission applications, our objective is to model the signal such that we can transmit likely signals with a small amount of bits and unlikely signals with a large number of bits. If we train a GMM on a speech database, we can determine which signals are speech-like, such that those can be transmitted with a low number of bits.