Model Selection

Peter Johansen

April 8, 2001

By means of the *EM* algorithm we can find a Gaussian mixture model of a finite data set from R^n. To apply the algorithm we are required to supply the number of gaussian components in advance. This note explains in principle how to determine the optimal number of components. In principle means that I know how to do it, but did not do it. The math scared me, maybe next time around.

We invoke the principle of parsimony[1], often called Occam's razor.

Choose the simplest among several possibilities.

We must make precise what we mean by simple. One data representation is simpler than another if it can be transmitted using fewer bits. The sender and the reciever agrees on a set of conventions which are:

1. The finite data set $X = \{x_i\}, i = 1, \ldots, N$ are integers in the range $\{1, \cdots, N\}$.

2. Each real x_i has error 1. Formally x_i is a stochastic variable $N(x_i, 1)$ with mean x_i and variance 1.

3. The data is represented by a gaussian mixture model with k components $p(x|j), j = 1, \ldots, k$ and k mixture coefficients $P(j), j = 1, \ldots, k$. Sender and reciever have agreed upon a common value σ^2 for the variance. Bishop [1] shows how to also include the variance (the covariance matrix) into the optimization.

$$p(x) = \sum_{j=1}^{j=k} p(x|j)P(j) \quad (1)$$

$$= \sum_{j=1}^{j=k} N(\mu_j, \sigma^2)P(j) \quad (2)$$

4. First, the parameters of the model is transmitted. Second, the data are transmitted after encoding by an entropi coding method, for instance a prefix code or arithmetic coding. The receiver first decodes the model parameters, then uses the model to decode the data.

[1] dansk: sparsommelighed

We are now in a position to make the above principle operational. The encoder tries a range of k-values and chooses the value that gives the shortest transmission length.

Message length

The message length is easy to predict. We know from Shannon's theorem that if we know the probability distribution p, the expected code length is the entropy $H(p)$ per data item.

$$\text{code length} = H(p)$$

Model length

The interesting part is to encode the model parameters. We assume that the model parameters are uniformly distributed in their range. To first approximation this implies that we should use the same number of bits for each parameter. But in addition we must determine their precision, we need to determine the number of significant bits in μ_j. The reason is that the sender knows the precision and can compute the distribution $p(x)$ based upon this high precision. The reciever, on the other hand, is not able to compute the centers from the data, hence it is necessary to transmit the extra bits for center positions. Let us choose a fixed number of bits to transmit N, s, d, k and $P(j|x)$, $j = 1, \ldots, k$.

As an initial observation let us determine the number of significant bits in $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_i$. The sensitivity of μ_j is the size of the gradient $|\nabla\mu_j|$. It expresses the change in the size of μ_j due to the uncertainty in the value of x_i.

$$\begin{aligned}
\mu_j &= \frac{1}{N}\sum_{i=1}^{i=N} x_i \\
\frac{\partial \mu_j}{\partial x_i} &= \frac{1}{N} \\
\nabla(\mu_j) &= \frac{1}{N}(1, 1, \ldots, 1) \\
|\nabla\mu_j| &= \frac{1}{\sqrt{N}}
\end{aligned}$$

As a consequence we can measure the result with a yardstick with measuring unit $\frac{1}{\sqrt{N}}$. This means that μ_j can be expressed as an integer of size $\frac{\mu_j}{|\nabla\mu_j|}| = \mu_j\sqrt{N}$. It requires $\lceil \log \mu_j \sqrt{N} \rceil = \lceil \log \mu_j + \frac{1}{2}\log N \rceil$ bits. We need - disregarding $\lceil . \rceil$ - $d + \frac{1}{2}\log N$ bits to transmit each coordinate of the parameter μ_j.

Thus the code for any of the coordinates of μ_j is the d bits precision of the data items plus an additional cost $\frac{1}{2}\log_2 N$. If there are k centers, each having n coordinates, the total extra cost is $n\frac{k}{2}\log_2 N$.

The incremental cost incurred by adding one more center is the n coordinates of the additional center, the probability $P(j)$ of the new center and in addition $\frac{n}{2}\log_2 N$.

This initial observation has shown us how to compute the bits necessary to transmit a model parameter under a simplifying assumption. The computation in the general case for a gaussian mixture was not shown. In principle one can find the gradient of the center coordinates and of the mixing coefficients from the formulas in the lecture note describing the expectation-maximation algorithm. One should by implicit differentiation of the formulas find the size of the gradient.

The bottom line is that we can compute the incremental cost for an additional parameter and weigh it against N times the entropy of the gaussian mixture. This way we can choose the minimum total message length and the optimum number of gaussians in a mixture.

References

[1] Christopher M. Bishop. *Neural Nets for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[2] Mark Hansen and Bin Yu. Wavelet thresholding via mdl for natural images. *IEEE Trans. Inform. Theory (Special Issue on Information Theoretic Imaging*, 2000.