COMP90051 Statistical Machine Learning

Semester 2, 2015

# Workshop Week 10: Gaussian Mixture Models

**Introduction**

In the first part of this workshop you will be working with Gaussian Mixture Models (GMM). You will practice fitting GMM to data and estimating the number of clusters using Akaike Information Criterion corrected for finite sample sizes (AICc). The second part of the workshop is dedicated to Project 2.

The overall plan for the first part is to generate data with known number of clusters, and then fit different GMMs to the same data. Each GMM will correspond to a different number of clusters. After the fit, AICc can be used to choose the best model, and it would be interesting to see if the chosen model corresponds to the correct number of clusters. A more interesting variation of the above scheme would be to swap data samples with another students, so that you don't know the actual number of clusters, but the other student knows. If you don't have anyone to swap the samples with, you can use an obfuscated sample generation as explained below.

Step 1

Choose the number of clusters $K$ between 3 and 10, and generate $K$ random samples. For an obfuscated number, choose $K$ using the random number generator, make sure the program saves the number in some file, but don't look up the value of this variable yourself. Each of the $K$ samples should comprise $round\left(\frac{1000}{K}\right)$ random points drawn from a bivariate normal distribution. You can vary the locations and spreads of these distributions making the samples more or less overlapping. In the first instance, start with a well separated samples. You will end up with a dataset with about $1000$ points and $K$ clusters.

Step 2

Figure out how to use the expectation-maximisation (EM) algorithm in your programming language of choice. If this algorithm is not implemented, you will have to use another programming language. Set $K$ to a particular value and use the EM algorithm to fit a GMM to the data. Fitting in this context means estimating the maximum likelihood parameters. Repeat the procedure several times, each time using a different $K$. Refer to lecture notes if you need a refresher on fitting a GMM using EM.

Step 3

Use the maximum likelihood estimates for different $K$ to compute AICc scores (see lecture notes). The suggested true number of clusters is then $K$ that corresponds to a model with the smallest score. Figure out what the actual $K$ was, and compare this with the estimated value.

Step 4

Try steps 1 to 3 using clusters that are more overlapping. Also try these steps generating a dataset where the number of data points varies between the clusters.