Fundamentals of Probabilistic Data Mining

Graded lab and homeworks

`http://chamilo.grenoble-inp.fr/courses/ENSIMAGWMM9AMO17/`

---

1 Mixture models

The Unistroke alphabet, closely related to Graffiti[1], is an essentially single-stroke shorthand hand-writing recognition system used in PDAs. The data set is composed of 50×6 time-trajectories representing the drawing of letters A, E, H, L, O and Q in a plane.

Here you will focus on modelling letter A (actually drawn as a Λ). After some pre-processing, we obtain the data set "Amerge.txt" (you can find it in the zip file), which is composed of every stroke of every trial for the gestures associated with that letter (the temporal aspect of sequences and the separations between sequences were lost here).

1.1 Lab work

1.1.1 Preparatory work and modelling

Do this before the class. Questions about this part will be answered only at the beginning of the practical session.

1. Prove the reestimation formula for Gaussian Mixture Model (GMM) (exercise 2 in the slides).

2. Simulate a sample of size 500 of the following bivariate GMM:

$$0.3\mathcal{N}(\mu_1; \Sigma_1) + 0.7\mathcal{N}(\mu_2; \Sigma_2)$$

   with

$$\mu_1 = \begin{pmatrix} -3 \\ 0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \text{ and } \Sigma_1 = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

   Hint: `numpy.random.multivariate_normal`.

   Plot the synthetic data set and check if it corresponds to the figures in the slides of the class (Page 6).

3. Download (from `chamilo`), load and plot the Unistroke data set (letter A) and provide the figure.

4. Do you think a 2-components GMM could be appropriate for letter A? Why?

---

[1] `http://en.wikipedia.org/wiki/Graffiti_(Palm_OS)`

1.1.2   Data analysis: Gaussian model

1. Estimate a bivariate GMM on the letter A data set and provide the estimated parameters.

2. Label the data using the estimated model and show the pdf of the estimated GMM
   (Provide one figure with the data labeled in color overlapping on the contours of the log(pdf),
   please add inline labels for the contours)
   Hint: `mixture.GaussianMixture.predict`, `numpy.meshgrid`.

3. To validate the assumption of bivariate Gaussian mixture:

   (a) Plot each marginal histogram (in x and y) and add the estimated mixture of univariate
       Gaussian pdfs to the figure.

   (b) For each marginal, provide separate histograms of each cluster and add the estimated
       univariate Gaussian pdf to the figure.

   Hint: `scipy.stats.norm`.

4. Comment the results of questions 3 (a) and (b). What to think about the bivariate Gaussian
   mixture assumption? Why?

5. Plot each data point x_i with some colourmap corresponding to $P(Z_i = 1|X_i)$ (you may plot
   $\log P(Z_i = 1|X_i)$ instead). How to interpret that plot?

1.2   Mandatory additional questions

The aim of this part is to compare mixture of von Mises distributions with Gaussian mixtures.

1. Transform the Unistroke data to angular data. Plot the histogram of angles and comment.

2. Define von Mises and mixtures of von Mises distributions.

3. A priori, would a mixture of von Mises distributions be more or less adequate than Gaussian
   mixtures on the real data set of part 1.1? Why?

4. Provide equations for the E-step and M-step of the EM algorithm for mixtures of von Mises
   distributions. Justify these results with formal computations.

5. Fit a 2-components mixture of von Mises distributions on the Unistroke data set of part 1.1.
   List the estimated parameters and color the data (in original form) by the estimated labels.
   Hint: `You may find an existing python library for mixtures of von Mises`.

1.3 Optional additional questions

1. Consistent estimators of the number of components.

   (a) Give a formal definition of consistent estimators of the number of components in a mixture model. Write some state-of-the-art on that topic, choose one of the references therein, justifying your choice. Provide a one-page description of the approach developed in that reference.

   (b) Imagine, describe and implement a protocol to evaluate the consistency of any arbitrary estimator of the number of components. Test this protocol on Gaussian mixtures to check the consistency of that estimator.

2. Implementation of the mixtures of von Mises distributions

   (a) Write your own sampling function and pdf function of Mixtures of von Mises distributions.

   (b) Use your functions to simulate a 3-components mixture, with sample size of 1,000. Provide the figure showing the data colored by the true labels and the contour plot of the log(pdf) of the simulated model (you may visualize them on 2D euclidean space).

   (c) Estimate the parameters on the simulated data using your implementation. Comment the results using parameters, histograms and bivariate plots with clusters (the same plot as for (b) but using the estimated parameters).