

LEARNING ALGORITHMS AND SYSTEMS LABORATORY
FACULTE DES SCIENCES ET TECHNIQUES DE L'INGENIEUR

SCHOOL OF ENGINEERING / SWISS INSTITUTE OF
TECHNOLOGY



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Lecturer:

Prof. Aude Billard

Office: ME A3 464

direct: +41-21-693 54 64

secretary: +41-21-693 09 39

fax: +41-21-693 78 50

E-mail: aude.billard@epfl.ch

Web: <http://lasa.epfl.ch/>

Assistant:

Klas Kronander

E-mail: klas.kronander@epfl.ch

Basilio Noris

E-mail: basilio.noris@epfl.ch

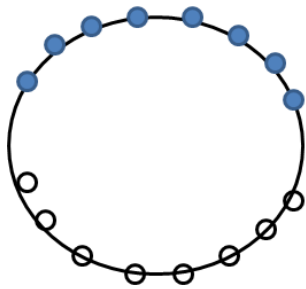
Applied Machine Learning

Exercises - II

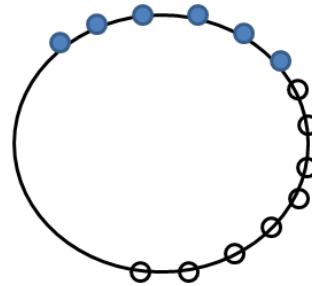
WINTER 2011-2012

Exercise 1: LDA

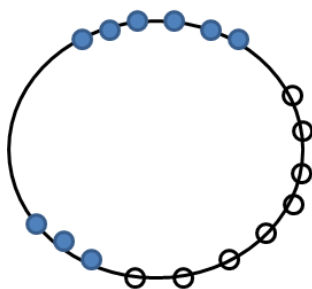
- i) Determine the line that LDA would compute in the following four examples. The filled circles correspond to points belonging to class +1 while the empty circles correspond to points belonging to the class -1. All points are located on a circle. Draw the resulting projections. Use your geometrical intuition only!



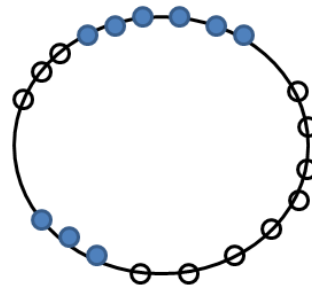
(a)



(b)



(c)

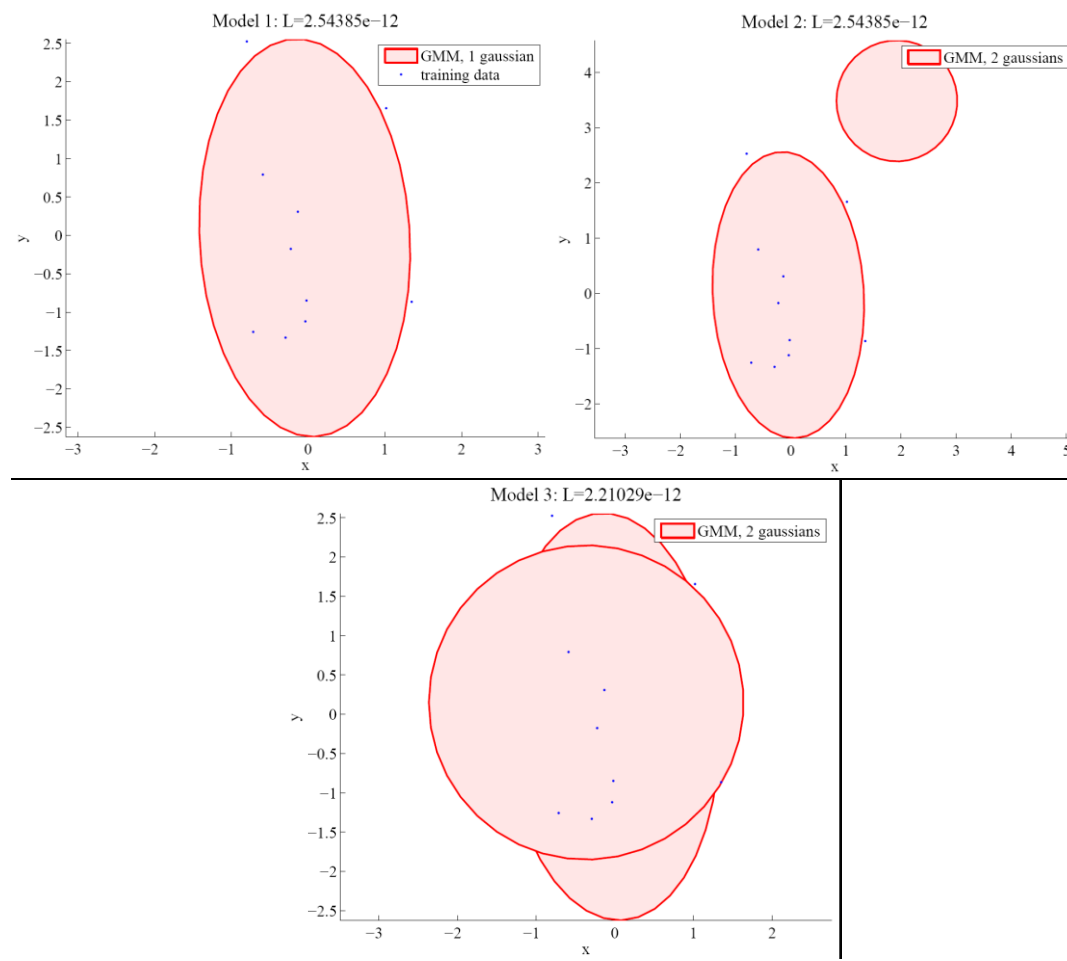


(d)

- ii) Create examples where points are all located on the boundary of a square and where the line would pass through the means of the two classes, (almost exactly) on one mean of the two classes and through neither of the two means.

Exercise 2: Mixture of Gaussians - Likelihood

Consider that we have three mixtures of Gaussians. The first one has a single Gaussian. The second and the third have each two Gaussians, see figure below.



- Write the formal definition of the pdf and the likelihood of each of the two mixtures
- Explain how come the likelihood is the same for the three mixtures even though they yield very different fits;
- Discuss how this can affect maximum likelihood estimation and classification using groups of GMM with Bayes

Solutions:

a) The pdf of a GMM at datapoint x^i is given by:

$$p(x^i) = \sum_{k=1}^K \alpha_k \cdot p(x^i | \mu^k, \Sigma^k), \text{ where } \alpha_k = \frac{\prod_{i=1}^M p(x^i | \mu^k, \Sigma^k)}{\sum_{l=1}^K \prod_{i=1}^M p(x^i | \mu^l, \Sigma^l)}$$

The likelihood is: $p(X) = \prod_{i=1}^M \sum_{k=1}^K \alpha_k \cdot p(x^i | \mu^k, \Sigma^k)$

For one Gaussian:

$$p(x^i) = \alpha_1 p(x^i | \mu^1, \Sigma^1) \sim N(\mu^1, \Sigma^1) \text{ and } \alpha_1 = 1$$

$$p(X) = p(X | \mu^1, \Sigma^1) = \prod_{i=1}^M p(x^i | \mu^1, \Sigma^1)$$

For two Gaussians:

$$p(x^i) = \tilde{\alpha}_1 \cdot p(x^i | \mu^1, \Sigma^1) + \tilde{\alpha}_2 \cdot p(x^i | \mu^2, \Sigma^2)$$

$$p(x^i) = \prod_{i=1}^M (\tilde{\alpha}_1 \cdot p(x^i | \mu^1, \Sigma^1) + \tilde{\alpha}_2 \cdot p(x^i | \mu^2, \Sigma^2))$$

b) When the second Gaussian is very far, its associated weight becomes very small ($\alpha_2 \ll 1$) and hence hardly affect the computation of the likelihood. When the two Gaussians are completely superimposed, they result in the same estimation. Hence their relative weights are almost the same $\alpha_1 \cong \alpha_2$ (half that of the single Gaussian). As a result, the addition of each of their effect does not bring any increase in the likelihood.

c) When using E-M, one penalizes for an increase in number of parameters that yield the same value of likelihood. This prevents overfitting data with more than the required number of Gaussians. In the example above, the BIC criterion would show that the fit with one Gaussian is better than any of the two other fits with two Gaussians, as it yields the same likelihood value but with the added cost of computing two Gaussians instead of one.

When comparing the likelihood of two GMM models (as in classification using GMM + Bayes), one should make sure that they contain the same number of Gaussians; otherwise, the absolute value could be biased. If the distribution of one of the two classes requires much more components to be well represented (e.g. if it entails much more non-linearities), then the two likelihood could be “normalized” by their number of Gaussians. i.e:

X has label +1 if:

$$\frac{1}{K_+} p_+(X) \geq \frac{1}{K_-} p_-(X)$$

where $p_+(X), p_-(X)$ are the pdf for the class +1 and -1 resp.

and K_+, K_- denote the number of Gaussians in each of these two GMM-s.

Exercise 3: Computational Cost of K-means, Soft K-Means, GMM clustering

The performance of a machine learning technique must often be evaluated in terms of its computational costs. The more computational steps are required the more unlikely it is that the algorithm could be ported for real-time computation on small portable hardware (robots, cell phones, PDA-s, etc). Computational costs are also tightly linked to the “curse of dimensionality”. The larger the dimension of the dataset is, the heavier the computational costs. Knowing whether computational costs grow linearly or exponentially with the number of datapoints, M , and the dimension of the dataset, N , is hence crucial. One will prefer a method that grows only linearly with M and N .

- i) Compute the computational cost *per iteration for the update step* of K-means, soft K-means and GMM clustering
- ii) Discuss the pros and cons of these three clustering techniques given your answer to (i).

Solutions:

- i) The update step of K-means, soft-K-mean and GMM with full covariance matrices grows with $O(K*M)$, $O(K*(M+1))$ and $O(K*M*(N*(N+1)/2))$, respectively (where K is the number of clusters/Gaussians).
- ii) K-means is the cheapest method. It however can fit solely clusters with isotropic distributions; it is also very sensitive to the choice of K ; soft-K-means and GMM relax this constraint.
GMM with full covariance matrices will hence be preferred only when computational costs are no real concern, as it offers maximal flexibility; Variants on GMM with either diagonal or isotropic covariance matrices may be considered to reduce computational costs while relaxing the constraint of choosing the number of clusters (as the optimal number K can then be estimated using BIC and AIC).

Supplementary Exercises (To be done at home)

Exercise 1: ICA, negentropy

ICA uses a fundamental property of Gaussian distributions to estimate the independent component. This property is that the entropy of a Gaussian distribution is larger than any other distribution with same mean and variance.

a) For a variable x and two associated distributions $g(x)$ (*Gaussian distribution*) and $f(x)$ with same mean and variance, show that the above property is true.

Hint:

Use the fact that the relative entropy $D(f \parallel g)$ of two distribution f and g is positive,

$$\text{i.e. } D(f \parallel g) = \int f(x) \ln \left(\frac{f(x)}{g(x)} \right) \geq 0$$

and that

$\int u(x)f(x) = \int u(x)g(x)$ if $u(x)$ is a quadratic form, i.e. such that $u(x) = x^T A x$, with A a square matrix.

b) Show that the negentropy is thus always positive and discuss what this means for ICA.

Solution:

a) Recall that $h(f(x)) = -\int f(x) \ln(f(x)) dx$ is the *differential entropy* of $f(x)$.

$$\begin{aligned} 0 &\leq D(f \parallel g) \\ &= \int f(x) \ln(f(x)) - f(x) \ln(g(x)) \\ &= -h(f(x)) - \int f(x) \ln(g(x)) \\ &= -h(f(x)) + h(g(x)) \quad (\ln(g(x)) \simeq x^T x \text{ is a quadratic form}) \\ &\Leftrightarrow h(f(x)) \leq h(g(x)) \end{aligned}$$

Since $f(x)$ and $g(x)$ have same mean and covariance

b) The negentropy is given by $g(x) = h(x)$. In ICA, we find the optimum of $J(x)$, i.e. its minimum. There is only 1 minimum $J(x)=0$, when the distribution of the data follows a Gaussian distribution. Thus, the larger the negentropy, the further away the distribution is from that of a Gaussian.

Exercise 2: ICA, Whitening

Recall that in ICA, whitening is done by projecting a zero mean distribution x through the matrix $V = D^{-\frac{1}{2}}E^T$, where E is the matrix of the eigenvector of the covariance matrix of x and D is a diagonal matrix composed of the eigenvalues of the corresponding eigenvector in E .

- Explains how such a projection whitens the data, i.e. ensures that the data once projected is uncorrelated and has variance equal to 1.
- While this projection is done in one time step, show that $z = Vx$ is a stationary point of the iterative learning rule

$$\Delta V = \gamma(I - zz^T)V$$

Solution:

$$E\{\Delta V\} = 0$$

$$\Leftrightarrow I - E\{zz^T\} = 0$$

$$\Leftrightarrow I - E\{Vxx^TV^T\} = 0$$

$$\Leftrightarrow I - VE\{xx^T\}V^T = 0$$

$$\Leftrightarrow I - VE\{xx^T\}V^T = 0$$

$$\Leftrightarrow I - D^{-\frac{1}{2}}E^TE\{xx^T\}ED^{-\frac{1}{2}} = 0$$

$$\Leftrightarrow I - D^{-\frac{1}{2}}E^TEDE^TED^{-\frac{1}{2}} = 0$$

$$\Leftrightarrow I - I = 0$$