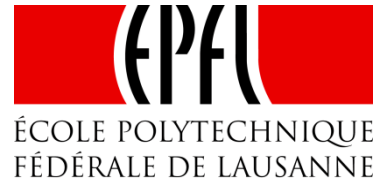


LEARNING ALGORITHMS AND SYSTEMS LABORATORY  
FACULTE DES SCIENCES ET TECHNIQUES DE L'INGENIEUR  
SCHOOL OF ENGINEERING / SWISS INSTITUTE OF  
TECHNOLOGY



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](mailto:aude.billard@epfl.ch)

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

Assistant:

**Klas Kronander**

*E-mail:* [klas.kronander@epfl.ch](mailto:klas.kronander@epfl.ch)

**Basilio Noris**

*E-mail:* [basilio.noris@epfl.ch](mailto:basilio.noris@epfl.ch)

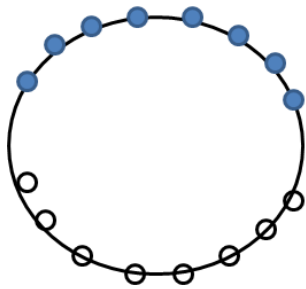
## Applied Machine Learning

### Exercises - II

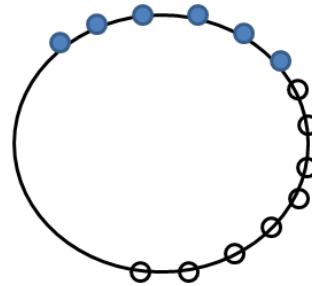
WINTER 2011-2012

**Exercise 1: LDA**

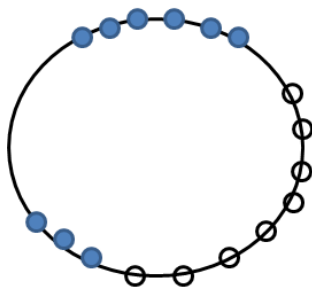
- i) Determine the line that LDA would compute in the following four examples. The filled circles correspond to points belonging to class +1 while the empty circles correspond to points belonging to the class -1. All points are located on a circle. Draw the resulting projections. Use your geometrical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



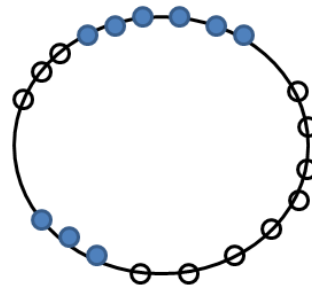
(a)



(b)



(c)

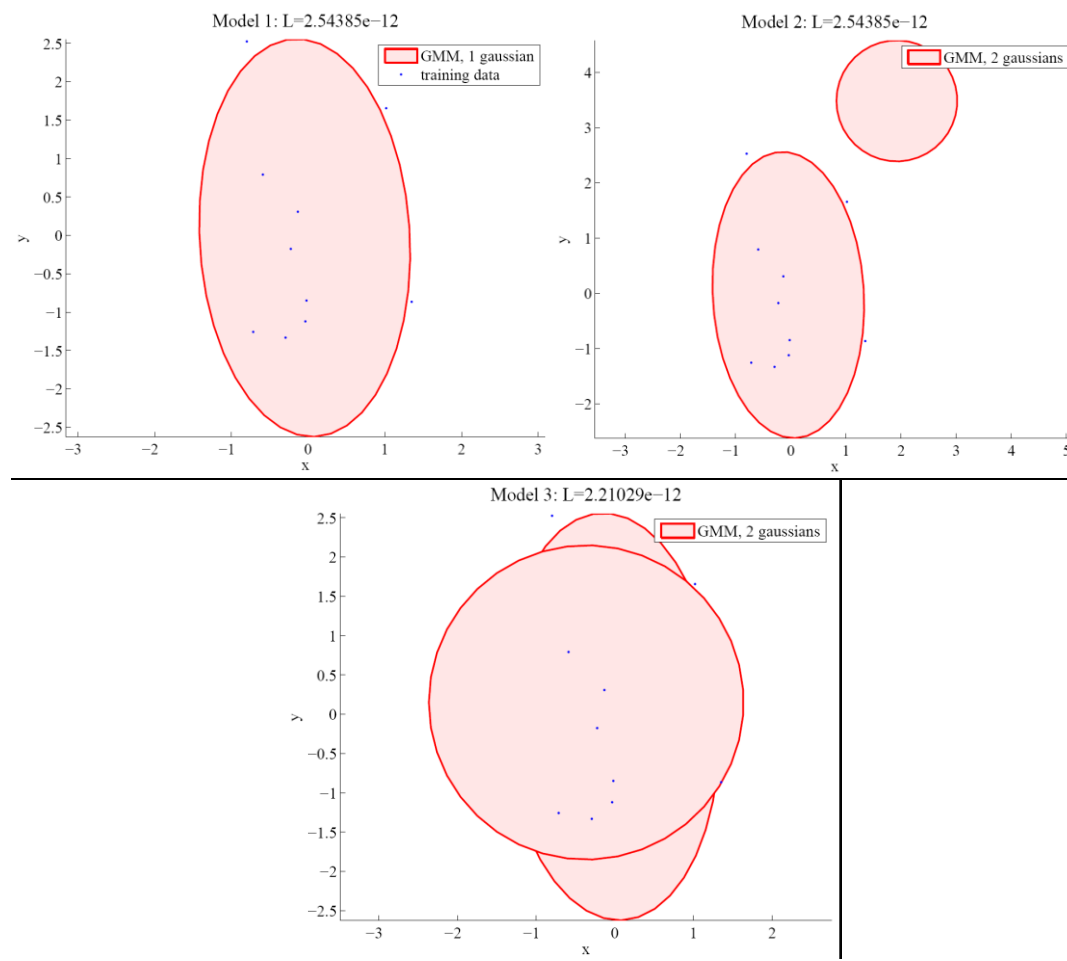


(d)

- 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 (almost) 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

**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).

## 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  $u(x)=x^T Ax$ , with  $A$  a square matrix.

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

### 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

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