**Exam: 23 January 2018**
**Duration: 2h45min**

**Name:** ____________________________________________

1. Indicate whether each of the following assertions is true or false. Give a **very short explanation**.

a) The optimal value of the objective function for the estimation of a probability density through the EM algorithm, using a mixture of p+1 Gaussians, cannot be higher than the objective function for the same estimation, using a mixture of p Gaussians.

b) The naive Bayes classifier is a special case of the Bayes classifier.

2. Consider the dataset D described in the Table. The set D is to be used as training data for a binary classifier to identify whether a point ($x_1$; $x_2$) falls inside some given target shape or not. Positive class labels (+) correspond to data-points falling inside the target shape, while negative class labels (-) correspond to data-points not falling inside the target shape.
Table: Training dataset containing 20 data-points pertaining to two different classes.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $-2.4$ | $-2.1$ | $-1.7$ | $-1.6$ | $-1.5$ | $-1.2$ | $-1.1$ | $-0.5$ | $0.0$ | $0.0$ |
| $X_2$ | $0.4$ | $-0.3$ | $-1.6$ | $-1.3$ | $1.5$ | $1.9$ | $-2.0$ | $0.1$ | $0.4$ | $2.0$ |
| Class | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ | $+$ | $+$ | $-$ |
| $X_1$ | $0.1$ | $0.1$ | $0.1$ | $0.2$ | $0.3$ | $0.4$ | $0.8$ | $1.0$ | $1.7$ | $2.0$ |
| $X_2$ | $-0.7$ | $-0.6$ | $0.0$ | $-0.5$ | $-0.5$ | $0.9$ | $0.2$ | $0.1$ | $-1.0$ | $0.4$ |
| Class | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | $-$ | $-$ |

(A) Given a data-point p = ($x_1$; $x_2$), consider the two binary attributes, $A_1$ and $A_2$, where attribute $A_i(p)$ is 1 if $|x_i| > 1$ and 0 otherwise, i = 1; 2. Using these attributes to represent each of the data-points in D, compute the parameters of a Naïve Bayes classifier for the dataset D. Using this classifier, compute the class label for the point (0.9; 0.9).

(B) Suppose that you want to use a perceptron to classify the points in D. Indicate any pre-processing steps necessary for the perceptron to perfectly classify all points in the data-set D. Note: You don't actually need to perform these steps, just explain what they would consist of.

3. Consider the following dataset.

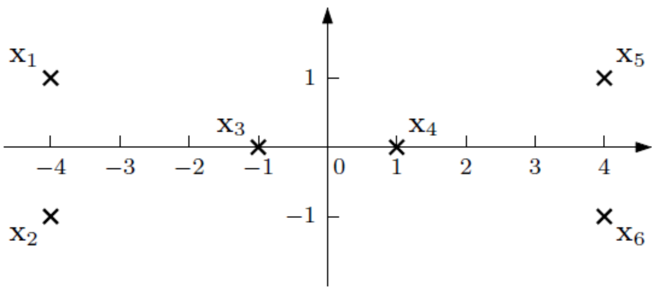
| $x_1$ | $x_2$ | $y_1$ | $y_2$ |
|---|---|---|---|
| -1.5 | -2 | -5.1 | 2.7 |
| -1 | -0.5 | -2.5 | -1.05 |
| 0 | 0 | 0.05 | 0.01 |
| +1 | +1 | 3.2 | 3.9 |
| +1.5 | +2 | 4.9 | 8.7 |

a) Consider that $y_1 = w_{11}x_1 + w_{21}x_2$. Estimate $w_{11}$ and $w_{21}$ by linear regression.
b) assume now that $y_2 = w_{12}x_1x_2 + w_{22}x_2$. Estimate $w_{11}$ and $w_{21}$ by linear regression.
c) Field knowledge tells us that $w_{11} = w_{12}$. Simultaneously estimate the four parameters imposing this constraint.

4. Consider an ensemble learning algorithm that uses simple majority voting among K learned hypotheses. Suppose that each hypothesis has an error $\epsilon$ and that the errors made by each hypothesis are independent of the others'. Calculate a formula for the error of the ensemble algorithm in terms of K and $\epsilon$, and evaluate it for the cases where K=5 and 20, and $\epsilon$= 0.1 and 0.4. If the independence assumption is removed, is it possible for the ensemble error to be worse than $\epsilon$?

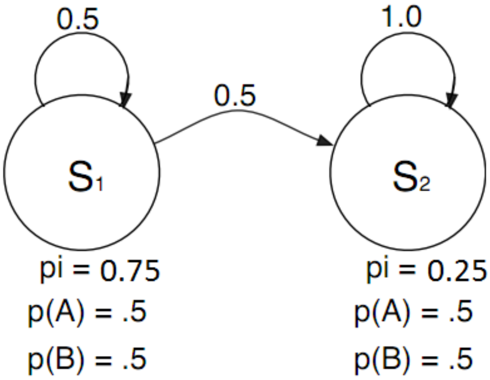
5. Consider a classification support vector machine (SVM) that uses a second-degree polynomial kernel, i.e., a kernel of the form $K(\bar{x},\bar{y})=(\bar{x}\cdot\bar{y}+a)^2$, where $a$ is a scalar. Show that the classification boundary of the SVM in the input space is described by an equation of the form $\bar{x}^T A\bar{x}+\bar{v}^T\bar{x}+c=0$, in which $A$ is a matrix, $\bar{v}$ is a vector, and $c$ is a scalar.

6. Consider the 6-point data-set D = {$x_1$,…, $x_6$}, depicted in Figure, where the data-points are marked with an "x".




Suppose now that we wish to partition the dataset D into three clusters. Manually run K-means on this data-set, indicating the cluster associations and prototype vectors after each iteration.
Initialize your vector prototypes to $\mu_1 = [-2.5, 0]^T$, $\mu_2 = [0, 0]^T$ and $\mu_3 = [2.5, 0]^T$.

7. Use the HMM depicted in the Figure to work out the following questions (A and B).




(A) In the given HMM, what is the probability that an observation sequence {AAB} was generated?
(B) Given the observed sequence {AAB}, what's the most likely (to be observed) value for t=5?

Solve one of the following:
(C) In the HMM described in the classes, observations are generated for all time steps. If we only observe the outputs $x_{t1}, \ldots, x_{tk}$ at the time steps $t_1, \ldots, t_k$, how could you modify the forward algorithm to calculate P(x)?
(C) Show that a stochastic matrix P always has the value $\lambda = 1$ as its eigenvalue.