Technion – Israel Institute of Technology *Summer 2013*

Electrical Engineering

Introduction to Machine Learning – International School Final Exam

1. Exam duration is three hours.
2. It is highly recommended to read the entire exam before you start.
3. Include brief explanations. You should answer all questions. The value of each question is given in the body of the question. The total number of points is 100.
4. You may use any material during the exam.
5. Write in a clear and organized manner.
6. The exam sheet includes 5 pages, including this page.

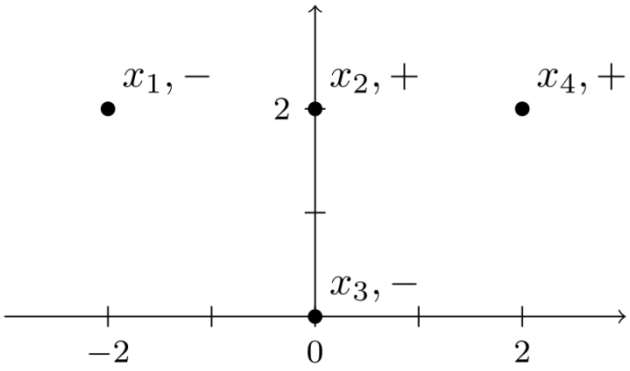
Good Luck

Distributions Table

Distribution	Notation	Support	PDF	Mean	Variance
Uniform	$x \sim U[a,b]$	$x \in [a,b]$	$f(x) = \frac{1}{b-a}$	$\frac{b+a}{2}$	$\frac{1}{12}(b-a)^2$
Normal	$x \sim N(\mu,\sigma^2)$	$\Re$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential	$x \sim \text{Exp}(\lambda)$	$0 \le x \in \Re$	$f(x) = \frac{1}{\lambda} e^{-x/\lambda}$	λ	λ^2

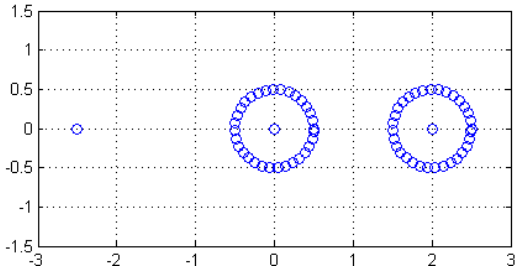
Question 1 – Assorted topics (35 points)

1. (9 Points) AdaBoost
   You are given a binary classification problem with the following set of 4 labeled examples, where $x \in \mathbb{R}^2, y \in \{-1,1\}$:



   We denote the coordinates of a sample x by $x = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}$. The AdaBoost algorithm is used in order to build a classifier for this data set, and it is given that in the first iteration $t = 1$, the algorithm uses the weak classifier $h_1 = \text{sign}(x^1 + 1)$ (a linear classifier parallel to the x^2 axis). The empirical distribution in step t is denoted by D_t.
   a. What is the <u>initial</u> distribution D_1?
   b. Compute the distribution after a single iteration D_2?

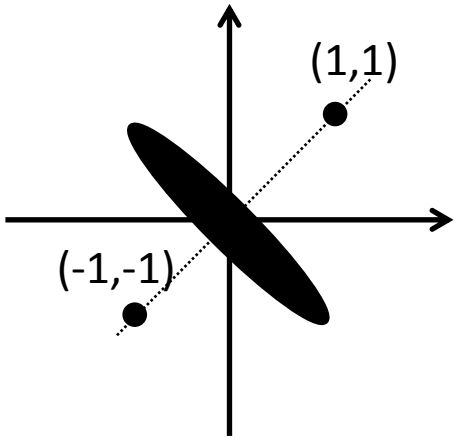
2. (9 points) K-Means
   For the data set sketched to the right, the K-means algorithm is run with $K = 2$. A solution $\{\mu_1, \mu_2\}$ is the centroids to which the algorithm converges (i.e., centroids which the algorithm does not update).



   a) Is there a single solution?
   (2 permutations of the same centroids are considered as a single solution).
   b) If there is only one solution, explain what it is and why there are no more solutions. If there is more than one solution, give at least two possible solutions, and give an example to initial conditions which converge to each of the two solutions.

3. (8 points) Decision Trees
   1. Determine if the following claim is correct, explain briefly: Two different decision trees that label the (same) training set identically with zero training error, will label every new input identically.
   2. Given a training set $\{(x_i, y_i)\}_{i=1}^n, x_i \in \Re^d, y_i \in \{-1,+1\}$ where $x_i \neq x_j$ for $i \neq j$. Is there a decision tree with zero training error, where the nodes conditions are of the form $1_{x^k > \theta}$ or $1_{x^k < \theta}$ for some feature $k = 1 \ldots d$ and $\theta \in \Re$?

4. (9 points) A data set is generated by sampling from a Gaussian with mean $\mu = (0,0)$ and covariance $\Sigma = \begin{pmatrix} 11 & -9 \\ -9 & 11 \end{pmatrix}$ with probability (1-p), or the point $(1,1)$ with probability p/2, or the point $(-1,-1)$ with probability p/2



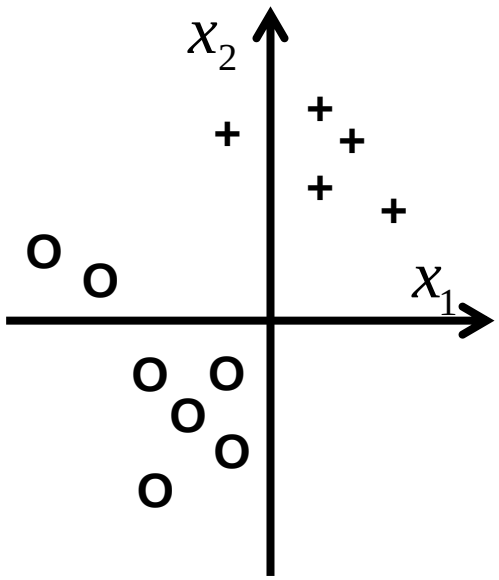
    1. What is the mean of this distribution? What is the covariance matrix?
    2. We like to reduce the dimension of this dataset from d=2 to d=1 using PCA. What is the 1st principal direction for p=0?
    3. What is the 1st principal direction for p=1?
    4. Is there a value for p for which there are two principal directions (ie both eigenvalues are equal)? There is no need to compute this value of p.

Question 2 – Empirical Risk Minimization (36 Points)

The items are independent.

1. (9 Points) You build an Optical character recognition (OCR) for English. Given an image your classifier should output vs the letter in the image is U or not-U (binary classification problem). Assume a uniform distribution over all 26 letters. You are offered to buy for $10 a classifier with a guaranteed accuracy of 95% over the test set. Will you buy this classifier? Please explain your answer.

2. (9 Points) Consider the classification problem shown in the right figure. Note that the data is linearly separable.
We train a classifier using SVM for the separable case, but with the addition of a regularization term on one of the parameters, $\{b, w_1, w_2\}$, for very large regularization parameter D:



$$\min_{w,b} \tfrac{1}{2}\|w\|^2 + Dw_j^2$$

$$\text{s.t.}: \; y_k(w^T x_k + b) \geq 1, \quad k = 1,2,\ldots,n$$

where $j \in \{0,1,2\}$ and for simplicity we define $w_0 \triangleq b$. In other words, only one variable is added to the objective in each case. Given the training data shown above, how does the training error change with regularization of each parameter w_j? State what is the training error for each w_j, for very large D. Provide a brief justification.

a) What are the number of training mistakes for j=0 (ie b)
b) What are the number of training mistakes for j=1
c) What are the number of training mistakes for j=2

3. (9 Points) Denote by (w^1, b^1) the solution of the following separable SVM problem.

$$(w^1, b^1) = \arg\min_{w,b} \tfrac{1}{2}\|w\|^2 \qquad \text{s.t.: } y_k(w^T x_k + b) \geq 1, \quad k = 1, 2, \ldots, n$$

Consider the following variant of SVM

$$(w^\alpha, b^\alpha) = \arg\min_{w,b} \tfrac{1}{2}\|w\|^2 \qquad \text{s.t.: } y_k(w^T x_k + b) \geq \alpha, \quad k = 1, 2, \ldots, n$$

Write (w^α, b^α) in terms of (w^1, b^1) for all real values of α

4. (9 Points) You learn a logistic regression over n=50 examples by minimizing

$Q(w) = \sum_{i=1}^{n} q(w, (x_i, y_i))$ for $q(w, (x_i, y_i)) = \log_2\left(1 + e^{-y_i (x_i \cdot w)}\right)$ where $x, w \in \Re^d, y \in \{-1, +1\}$.

Note that the logistic function bounds the zero-one loss $q\left(w, (x, y)\right) \geq 1_{y(x \cdot w) \leq 0}$. You trained a classifier and found that the empirical loss is $Q(n) = 9.2$. Provide the tightest bounds required below, these should be a natural number (or zero).

a. What is the tightest lower bound on the number of training mistakes? What is the tightest upper bound on the number of training mistakes?
b. Repeat part a for $Q(n) = 0.92$

Question 3 – Inference (29 points)

A decision problem is given, in which the input space is the real numbers, $X = \Re$. Each input example belongs to one of 3 classes: $\Omega = \{e, g, u\}$. The class conditional distributions for the 3 classes are (see distributions table in page 1 of the exam):

- Class u: $x \sim U[\lambda_u - 1, \lambda_u + 1]$
- Class g: $x \sim N(\lambda_g, 1)$
- Class e: $x \sim \text{Exp}(\lambda_e)$

1. (7 points) What is the maximum likelihood estimator (MLE) for the parameter of the exponential distribution λ_e, given n independent samples $x_1, \ldots x_n$, drawn from class e? Is it a biased estimator?

From this point onward assume that the parameters are known: $\lambda_u = \lambda_g = \lambda_e = 1$.

2. (7 points) Plot the conditional probability densities of the input given each one of the classes.
3. (8 points) Assuming the prior distribution over states is uniform, what is the optimal Bayes classifier of the state, given a single input x? Give a function from the real numbers to Ω .

From this point onward assume that the prior distribution over states is

$p(e) = 0.6$, $p(g) = p(u) = 0.2$.

4. (7 points) Assume now that the observations are n iid samples from a **single** class, $x_1 \ldots x_n$, for a very large n. We define two random variables, the maximal value $x_{\max} = \max_i x_i$ and the minimal value $x_{\min} = \min_i x_i$. We look for a decision rule for predicting the state using the pair $[x_{\min}, x_{\max}]$, of the form

$$\begin{aligned} &\text{if } (x_{\min} < \theta_A) \text{ then state is } \omega = \text{A} \\ &\text{else if } (x_{\max} > \theta_B) \text{ then state is } \omega = B \\ &\text{else} \qquad\qquad\qquad\quad \text{state is } \omega = C \end{aligned}$$

Find the thresholds $\theta_A, \theta_B \in \Re$ and the states $A, B, C \in \Omega$ such that the decision error given the state u is 0, and the decision error given the other 2 states is minimal.