Technion – Israel Institute of Technology *Summer 2014*

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 – Inference (29 points)

A decision problem is given, in which the input space is the real numbers, $X = \Re$. Input examples belong to one of three classes: $\Omega = \{e, g, u\}$. Class conditional distributions for the three classes are (see also 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) Given are n independent samples $x_1, \ldots x_n$, drawn from class e. Write the maximum likelihood estimator (MLE) for the parameter λ_e of the exponential distribution. Is the estimator a biased estimator?

Assume until the remaining of the question 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.

Assume until the remaining of the question the prior distribution over states is $p(e) = 0.6$, $p(g) = p(u) = 0.2$.

4. (7 points) We are given n i.i.d observations $x_1 \ldots x_n$ from a **single** class, for a very large n. 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 $[x_{\min}, x_{\max}]$. The rule should be of the form:

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

   Find a values for 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.

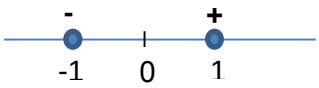
Question 2 – SVM (30 points)

This question deals with a SVM problem in its non-separable formulation, with a constant C. The algorithm minimizes the following objective:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t. \quad \xi_i \geq 0 \text{ for } i = 1\ldots n$$
$$\xi_i \geq 1 - y_i w \cdot x_i \text{ for, } i = 1\ldots n$$

Denote the optimal solution of the SVM problem by w^*, ξ^* (where ξ stands for the set $\{\xi_1, \ldots, \xi_n\}$). As a reminder, given arbitrary w (not necessarily optimal), the optimal value of ξ_i is $\xi_i^* = \max\{0, 1 - y_i \cdot w^T x_i\}$.

a. (6 points) Given is a training set of 2 one-dimensional samples, as shown in the plot. Consider two possible values, $w = 1$ or $w = \frac{1}{2}$ (**not necessarily optimal**), and in addition it is given that $C >> 1$. For which value, $w = 1$ or $w = \frac{1}{2}$, you expect that the objective function will get a higher value? Calculate the value of the objective function for each of these values (the answer can depend on C).



b. (6 points) For an arbitrary value of C (not necessarily large), find an optimal solution for w^*. Draw a graph of w^* as a function of C, for $0 < C < \infty$.

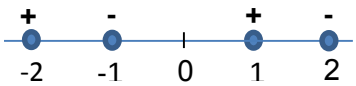
c. (6 points) For each one of the following kernel functions, determine whether the given training set is linearly separable, if the samples are transformed to a new space in which the inner product is given by the kernel function:

   i. $K_1(x_i, x_j) = (x_i x_j)^2$

   ii. $K_2(x_i, x_j) = (1 + x_i x_j)^2$

   iii. $K_3(x_i, x_j) = e^{-\frac{(x_i - x_j)^2}{\sigma^2}}$ for $\sigma \ll 1$.

d. (6 points) Now two more samples are added to the training set, as shown in the following plot:



   Prove that an optimal solution satisfies $w^* < 0$.

e. (6 points) For the training set from section (d), calculate the value of the objective function for $w = -1$ and for $w = -\frac{1}{2}$.

Question 3 – Assorted topics (41 points)

1. (8 Points) AdaBoost

We saw in class the AdaBoost algorithm (see the appendix at the end of the exam sheet). Given are three functions $h_i : X \to \{+1,-1\}$ for $i = 1,2,3$.

Denote the majority function $H(x) = \text{maj}\left[\text{h}_1(x), h_2(x), h_3(x)\right]$. Given $x \in X$ the function $H(x)$ outputs a label which is the majority of the set $\{h_1(x), h_2(x), h_3(x)\}$. Assume that all classifiers have the same error denoted by $error\left(h_i\right) = \epsilon$.

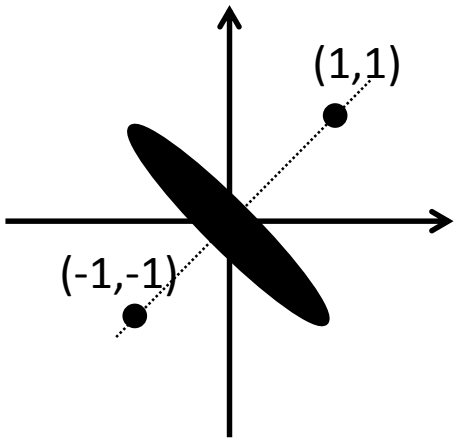
   a. (2 points) Find coefficients $\alpha_1, \alpha_2, \alpha_3$ such that $H(x) = sign\left(\sum_{i=1}^{3} \alpha_i h_i(x)\right)$.

   b. (3 points) Calculate a tight upper and lower bounds on the training error of $H(x)$ for $\epsilon = 0.3$.

   c. (3 points) Repeat the former section (b) for $\epsilon = 0.4$.

2. (8 points) Decision Trees
   a. 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.
   b. Given a training set $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, $x_i \in \Re^d$, $y_i \in \{-1,+1\}$ 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 $\theta \in \Re$?

3. (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.

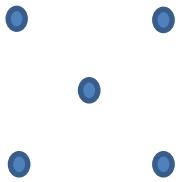


   a. What is the mean of this distribution? What is the covariance matrix?
   b. We like to reduce the dimension of this dataset from d=2 to d=1 using PCA. What is the 1st principal direction for p=0?
   c. What is the 1st principal direction for p=1?
   d. Is there a value for p for which there are two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

4. (8 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 solution.

5. (8 points) Consider the k-means algorithm. The algorithm changes the association of a sample to a centroid only if its distance to another centroid is strictly smaller than the distance to the currently associated centroid.

   Given is a set of 5 samples, 4 in the vertices of a square, and a 5^{th} point in the intersection of the diagonals (see plot).



   a. For $k = 2$ and Euclidean distance metric, find all possible partitions which the algorithm can converge to.
   b. For each of the partitions you found, determine whether it is stable or not. A partition is defined stable if when a small change is made to one of the centroids (and the algorithm continues to run), the algorithm converges to the same partition.

Appendix: AdaBoost Algorithm

1. Initialization: A uniform distribution $D_1 = 1/m$.
2. Find a weak classifier $h_t : X \to \{1,-1\}$ with small average error (necessarily smaller than half) with respect to D_t.
3. Denote the error by: $\varepsilon_t = \Pr_{i\sim D_t}\left[h_t(x_i) \neq y_i\right]$.
4. Set: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.
5. Update:

$$D_{t+1}(i) = D_t(i)\frac{\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t},$$

   Where Z_t is a normalization constant such that $\sum_{i=1}^{m} D_{t+1}(i) = 1$.
6. Return to step 2 until some stopping criterion is satisfied.
7. The final hypothesis is: $H(x) = \operatorname{sgn}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right)$