Technion – Israel Institute of Technology *Summer 2015*

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 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 including electronic devices as readers, with downloaded materials.
5. You are not allowed to communicate or browse your device during exam hours.
6. PLEASE TURN OFF ALL COMMUNICATION.
7. Write in a clear and organized manner.
8. The exam sheet includes 7 pages, including this page.

Good Luck

Question 1 – SVM (30 points)

Unless specified otherwise, the following sections are independent.

1. A separable SVM problem is solved on a labeled training set $\{x_i\}_{i=1}^n$. The resulting solution is w,b. A new training set is generated by applying the transformation $f(\cdot)$ on each data point. We denote the modified training set by $\tilde{x}_i = f(x_i)$ for $i = 1,\ldots,n$.

   The modified data is used to solve an SVM problem, and the solution is denoted by $\tilde{w},\tilde{b}$. For each of the following transformations, write $\tilde{w},\tilde{b}$ in terms of w,b.

   a. $f_1(x) = ax$, where a is a scalar constant.

   b. $f_2(x) = Ux$, where U is a square, orthogonal matrix

   c. $f_3(x) = x + x_0$' where x_0 is a constant vector

2. An **inseparable** SVM problem is solved using the training set $\{x_i\}_{i=1}^n$ and a constant C. We denote the resulting solution by w,b,ξ_i ($i = 1,\ldots,n$). We apply the transformation f_1 on the training set, and solve an SVM problem with a different constant $\tilde{C}$. We denote the resulting solution as $\tilde{w},\tilde{b},\tilde{\xi}_i$. Find $\tilde{C}$ such that the errors would not increase, namely $\tilde{\xi}_i = \xi_i$ for $i = 1,\ldots,n$.

3. Consider the following 1D labeled training set $\{-1,+\};\{0,-\};\{1,+\}$. An SVM problem is solved using a kernel function $K(x_i,x_j) = \left(x_i^T x_j + 1\right)^2$ and assuming no bias ($b = 0$). The solution to the dual problem is $\alpha_1 = \alpha_3 = 1, \alpha_2 = 3$. Find the classifier's decision boundary.

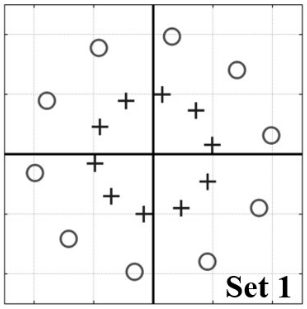
4. Repeat section 3 when the bias might be different than zero, and the solution to the dual problem is $\alpha_1 = \alpha_3 = 1, \alpha_2 = 2$.

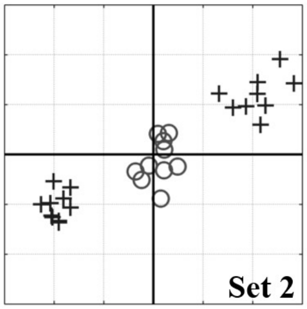
5. A separable SVM problem is solved using a kernel function. For each of the following training set, choose the kernel function that achieves zero error. If there is more than one function, pick the one with the lowest index. Explain your answer.

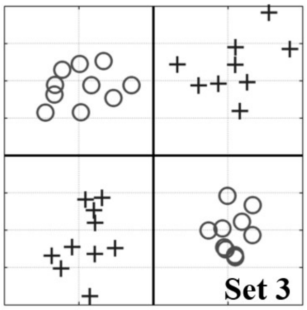
   i. $K_1(x_i, x_j) = x_i^T x_j$

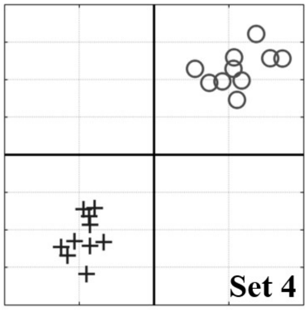
   ii. $K_2(x_i, x_j) = \left(x_i^T x_j + 1\right)^2$

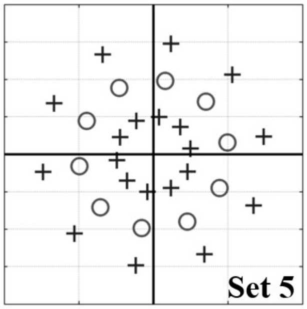
   iii. $K_3(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$ for some. γ


Set 1


Set 2


Set 3


Set 4


Set 5

Question 2 – PCA and Regression (30 points)

Let $R \in \Re^{p\times k}$ be a real matrix. Let z be a random, normally distributed state vector of dimension k, $\mathrm{P}(\mathrm{z})=\mathcal{N}(0,I)$. The observation x is a noisy version of z, the conditional probability of x given z is $P(x|\mathrm{z})=\mathcal{N}(Rz,10I)$.

a. Given z, find the MAP estimator for x, namely $\mathrm{x}=\arg\max\ \mathrm{P}(\mathrm{x}|\mathrm{z})$.

b. Given x, find the MAP estimator for z, namely $\mathrm{z}=\arg\max\ \mathrm{P}(\mathrm{z}|\mathrm{x})$.

In the following, assume that $k=1, p=2, R=\begin{pmatrix}1\\1\end{pmatrix}$.

c. Let a be real constant. Set $y=az$ (a scalar). Consider the regression problem for predicting y given x according to a linear model $\hat{y}=w\cdot x$.
   1. Write the squared error of some pair (x,y) as a function of w,z,x,R,a .
   2. Write the expected value of the squared error.
   3. Find w which minimized the expectation of the squared error.

d. Find the expectation and covariance matrix of x.

e. PCA is applied on samples drawn according to $P(\mathrm{x})$. Denote the first principal vector as u_1 . Calculate u_1 and find the reconstruction error $E\left(\| u_1u_1^\top x-x\|^2\right)$.

Question 3 – Assorted topics (41 points)

1. (8 Points) Perceptron

Given a labeled sample (x, y) and a linear classifier w, we define the margin as $yw^T x$. Recall that in the t-th iteration of the Perceptron algorithm, the input is a labeled sample (x, y). If the sample is mislabeled, namely $y_t \neq \text{sign}(w_t^T x_t)$, the algorithm updates $w_{t+1} = w_t + y_t x_t$. We denote this update rule as P. Consider the following alternative update rule: if the point is misclassified, set $w_{t+1} = -w_t + y_t x_t$. We denote this update rule as $\tilde{P}$.

The Perceptron algorithm was applied on t-1 samples, and the resulting classifier in the t-th iteration is w_t . Assume that (x_t, y_t) is mislabeled and $x_t \neq 0$. We denote the margin (of (x_t, y_t)) after applying P as γ , and, correspondingly, the margin after applying $\tilde{P}$ as $\tilde{\gamma}$. Which of the following statements is correct? Explain your answer

a. $\gamma \geq \tilde{\gamma}$ and after applying $\tilde{P}$ the point (x_t, y_t) is correctly classified.

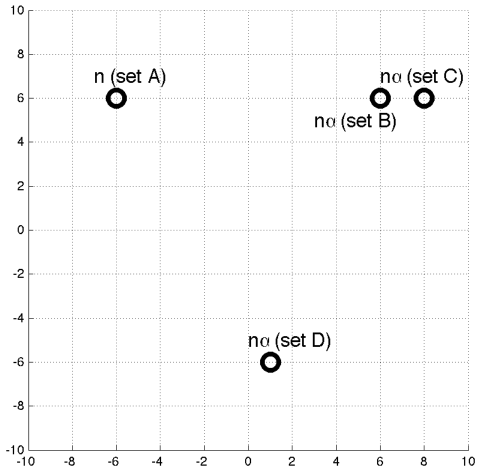
b. $\gamma \geq \tilde{\gamma}$ and after applying $\tilde{P}$ the point (x_t, y_t) may be misclassified.

c. $\gamma \leq \tilde{\gamma}$ and after applying $\tilde{P}$ the point (x_t, y_t) is correctly classified.

d. $\gamma \leq \tilde{\gamma}$ and after applying $\tilde{P}$ the point (x_t, y_t) may be misclassified.

2. (8 points) K-Means

Consider the following figure, which presents the distribution of $n(1+3\alpha)$ points in the plane. We would like the cluster the points into three clusters.



a. The algorithm is initialized by setting of the initial center of one of the three (out of four) points presented in the figure. In each of the four possible initializations, what is the final partition of the points into clusters?

b. Which partition minimizes the cost function? You may assume that all points at the same coordinate are associated to the same cluster. Write the solution as a function of α .

3. (8 points) AdaBoost

Consider a training set with n=5 samples. In each of the following section, AdaBoost is applied using a set of weak classifier (the sets may be different in different sections). In each iteration, a weak classifier with a weighted error of less than 0.5 is used. In the first three steps, the coefficients $\alpha_1, \alpha_2, \alpha_3$ are obtained. For each of the following series, find as tight as possible upper bound on the final training error of the classifier, or state that the presented series is infeasible.

a. $\alpha_1 = \frac{1}{2}\log 5, \alpha_2 = \frac{1}{2}\log 7, \alpha_3 = \frac{1}{2}\log 9$

b. $\alpha_1 = \frac{1}{2}\log 4, \alpha_2 = \frac{1}{2}\log 7, \alpha_3 = \frac{1}{2}\log 13$

c. $\alpha_1 = \frac{1}{2}\log 4, \alpha_2 = \frac{1}{2}\log 7, \alpha_3 = \infty$

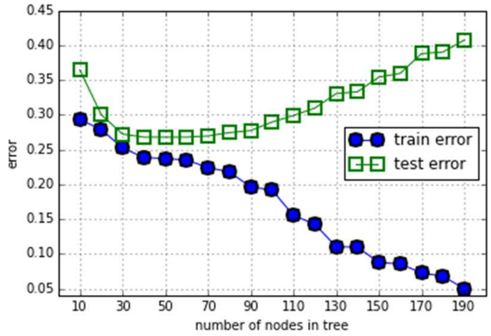
4. Concepts (8 points)

In a given classification problem we sampled a training set and a test set. Inputs came from some fixed unknown distribution P(x), and labels are a deterministic function of inputs. We learned few decision trees using the training set. Each tree uses different number of nodes. In the figure below we plot the training error and test error as a function of the number of nodes (of a learned tree). Reminder: training error is the fraction of mistakes on the training set, test error is the fraction of mistakes on examples on test set, generalization error is the fraction of mistakes according to P(x).

a. If we were drawing another training set, would the curve of training error would change? Would the graph describing the test error change?

b. If we were drawing another test set, would the curve of training error would change? Would the graph describing the test error change?

c. Use the plots and choose the number of nodes in the trees you would use to classifiy future inputs. Your goal is to minimize generalization error.



5. (8 points) Decision Trees

Let $S=(i,j)\quad i,j=1...10$ be a set of 100 different vectors. We denote by T_n the set of decision trees with exactly n leaves.

a. Is there a binary labelling of the vectors in S such that there exists a tree in T_{20} that correctly classifies S ?

b. Is there a binary labelling of the vectors in S such that there exists a tree in T_{100} that correctly classifies S ?

c. Is there a binary labelling of the vectors in S such that there does not exists a tree in T_{20} that correctly classifies S ?

d. Is there a binary labelling of the vectors in S such that there does not exists a tree in T_{100} that correctly classifies S ?