Technion – Israel Institute of Technology *Summer 2016*

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 laptops).
   Disable any connectivity.
5. Write in a clear and organized manner.
6. The exam sheet includes 7 pages, including this page.

Good Luck

Question 1 – Bayesian Classification (30 points)

Note: Part A and part B are independent and unrelated.

Part A

Consider a set of *k+1* Gaussian distributions, given by:

$$f\left(x\,|\,\theta\right)=\frac{2\left(|\theta|+1\right)}{\sqrt{2\pi}}\exp\left(-2\left(x-\theta\right)^{2}\left(|\theta|+1\right)^{2}\right)$$

Each distribution is characterized by an integer parameter for $\theta=0,1..,k-1,k$.

1. Write an expression for the mean and variance of each distribution. Draw the four distributions for *k=3*. For each distribution, sketch the mean and variance on the drawing.
2. Calculate the MAP estimator of θ for k=3, assuming a uniform prior.

In the following assume *k=10*.

3. We measure a single value *x*, and it was found out that $x\in\left[1000,1001\right]$. In each of the following cases, find the MAP and the MLE estimator for θ. It is sufficient to use quantitative arguments rather than an exact calculation.
   a. The prior over θ is uniform.
   b. The prior is $P(\theta)=z\left|\sin\left(\pi\theta/2\right)\right|$ where z is a normalization constant.
4. Assume a uniform prior over θ. We define a variable,

   $$A(x)=x-\lfloor x\rfloor$$

   For example, $A(2.42)=0.42$. Write the MAP estimator of θ given that $A(x)\in[0,0.001]$.
   Tip: Check qualitatively which of the following events

   $$x\in[i-0.001,i+0.001]\qquad i=-k,..,0,1..k$$

   has the highest probability.

Question 2 – K means & PCA

We denote the center of the α cluster at iteration t as $C_\alpha^{(t)}$ and by $N_\alpha^{(t)}$ the number of points that are associated with it at time t. **Note that the some of the question's sections are independent.**

We define a new clustering algorithm:

Modified Clustering Algorithm

Input: Initial center list $\left\{C_\alpha^0\right\}_{\alpha=1}^n$ and a set of points $\left\{x_i\right\}$.

1. Associate each point x_i with the nearest center. We denote the set of indices of the points that are associated with cluster α as N_α^0. Namely, $i \in N_{\alpha'}^0 \Leftrightarrow \alpha' = \arg\min_{\alpha=1..n}\left(\left\|x_i - C_\alpha^0\right\|\right)$.
2. Initialize $t \leftarrow 1$.
3. While (there exists α for which $N_\alpha^{(t)} \neq N_\alpha^{(t-1)}$) or (t=1)
   a. For every α

$$C_\alpha^t \leftarrow \frac{1}{\left|N_\alpha^{t-1}\right|}\sum_{i \in N_\alpha^{t-1}} x_i$$

   b. Take a point x_i and associate it with the center α' which satisfies

$$\alpha' = \arg\min_{\alpha=1..n}\left(\left\|x_i - C_\alpha^t\right\| + 2\left|N_\alpha^{t-1}\right|\right)$$

   and fix $i \in N_{\alpha'}^t$. Repeat for all points.
   c. Set $t \leftarrow t+1$.

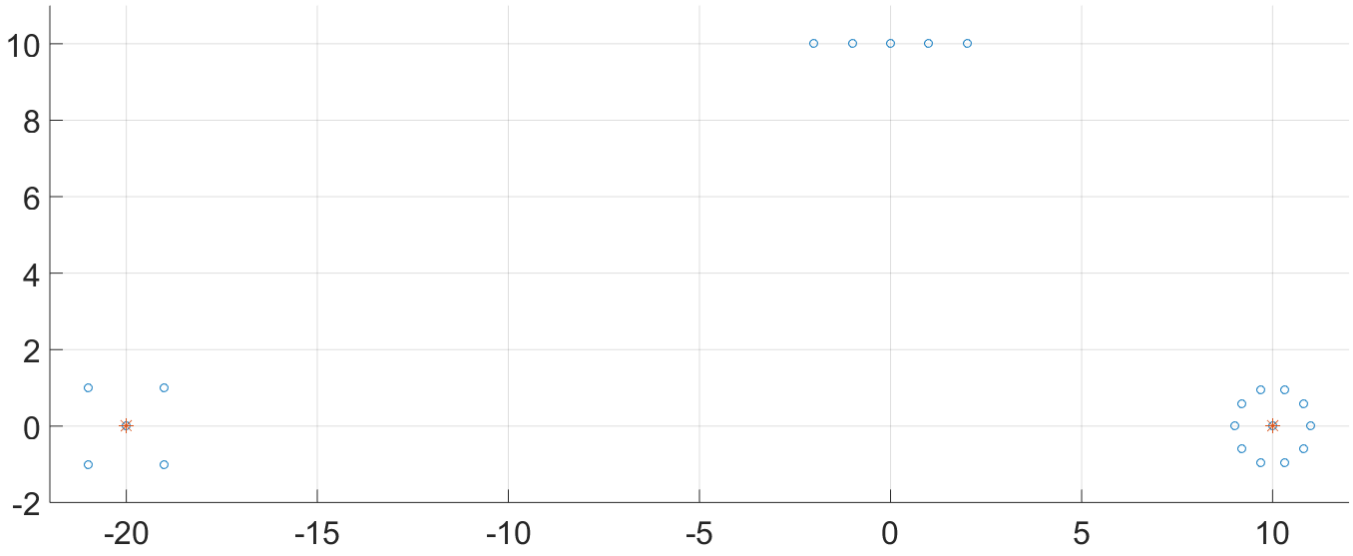
1. Apply the algorithm on the data points depicted in the following figure. The two initial centers are $C_1^0 = (-20,0), C_2^0 = (10,0)$.
   The data points are

$$x_1 = (-20,0)$$
$$x_i = (-20,0) + (\pm 1, \pm 1) \quad i = 2..4$$
$$x_i = (i-8,10) \qquad i = 6..10$$
$$x_{11} = (10,0)$$
$$x_i = (10,0) + \left(\cos\left(\frac{2\pi(i-12)}{10}\right), \sin\left(\frac{2\pi(i-12)}{10}\right)\right) \qquad i = 12..21$$

   Calculate the final centers and to which center each point is associated with.
   Tip: In order to avoid unnecessary calculation, consider which are the "limiting" / "boundary" points and do your calculation only for those points instead of the whole set of points.
2. We apply the PCA algorithm only on the points which are associated with the final first cluster, in other words, only on points that were associated with C_1^{final}. Draw on the last page of the exam

the projection of these points on the first principal component. There no need for an exact calculation, a rough sketch is sufficient.

3. We add two points $x_{23} = (250, 250)$ $x_{22} = (0, 0)$ and we apply the PCA algorithm on all points (with no preprocessing by the modified clustering algorithm). Draw on the last page of the exam the projection of these points on the first principal component. There no need for an exact calculation, a rough sketch is sufficient.
4. Give an example of an input for which the algorithm does not stop (is in an infinite loop).



Misc. Topics

- Question 3: Decision Trees

Consider a classification problem with a sample of size *n*, and k classes. We denote the input as the matrix $\mathcal{X} \in \Re^{d \times n}$ and the labels as $\mathcal{Y} \in \{1,2,..,K\}^n$. Namely, the i-th input is the i-th column of the matrix $\mathcal{X}[:,i]$. A decision tree learning algorithm was deployed on the training set, and it resulted in a decision tree T with a zero training error. For each of the following cases, decide if the resulting tree will be the same as T. Two trees are the same if, for every input, the output labels are the same.

1. Assume there are two features $s \neq t$ for which all samples have the same values, i.e. $\mathcal{X}[s,:] = \mathcal{X}[t,:]$. We now remove the s row from the matrix (we erase the feature s) and reapply the learning algorithm.
2. Assume there are two samples $i \neq j$ for which all features have the same values, i.e. $\mathcal{X}[:,\mathrm{i}] = \mathcal{X}[:,\mathrm{j}]$. We now remove the i-th column and reapply the learning algorithm.
3. Assume we duplicate each of the sample point, so that we have a sample of size 2n, and then the training algorithm is applied again.

- Question 4: Perceptron

We apply the perceptron algorithm on the sample of $n = 400$ two dimensional points $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{400}$. All the input points are confined (within) the unit ball. It is given that this set is linearly seperable with a unit vector $u \mid \|u\| = 1$ and margin $\gamma = 0.001$.

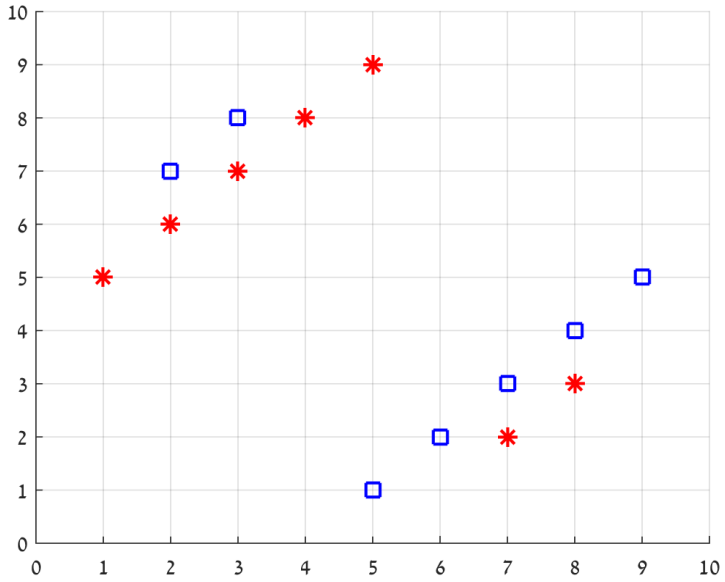
1. Find an upper and lower bound that the algorithm will incur during a single pass on the whole set.
2. Find an upper and lower bound that the algorithm will incur during two passes on the whole set.
3. Find an upper and lower bound that the algorithm will incur during three passes on the whole set.

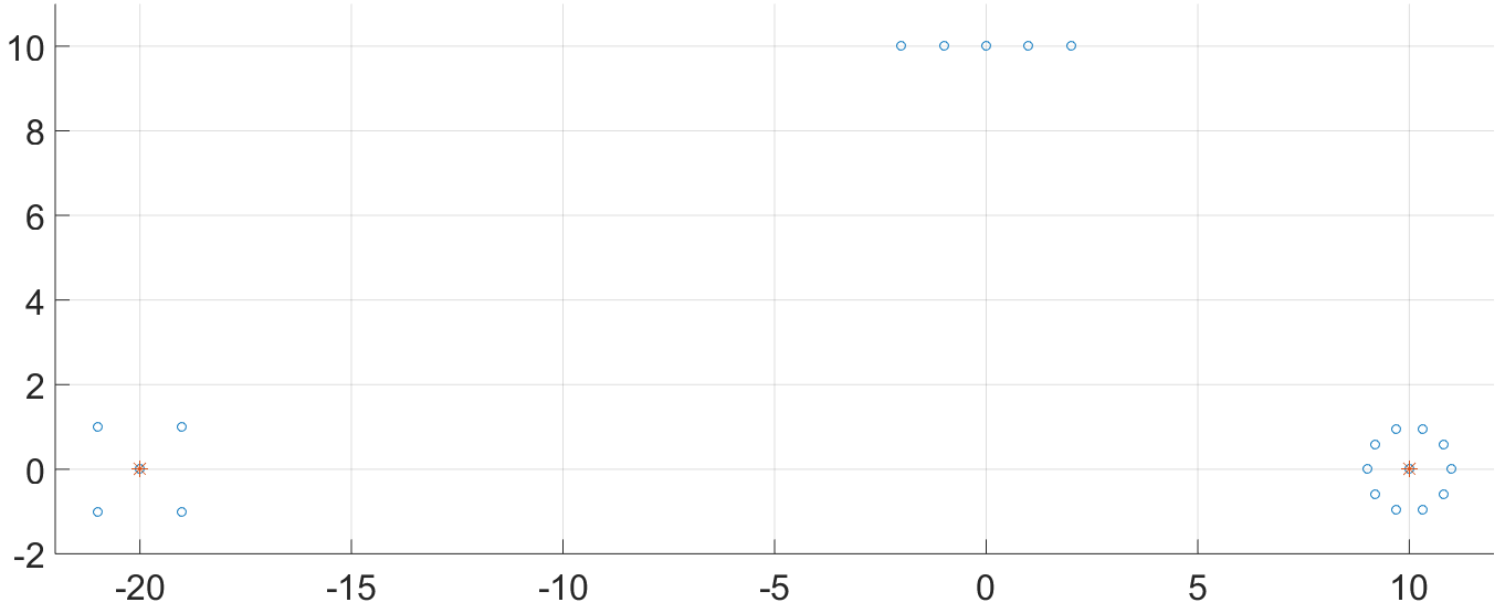
- Question 5: k Nearest Neighbors

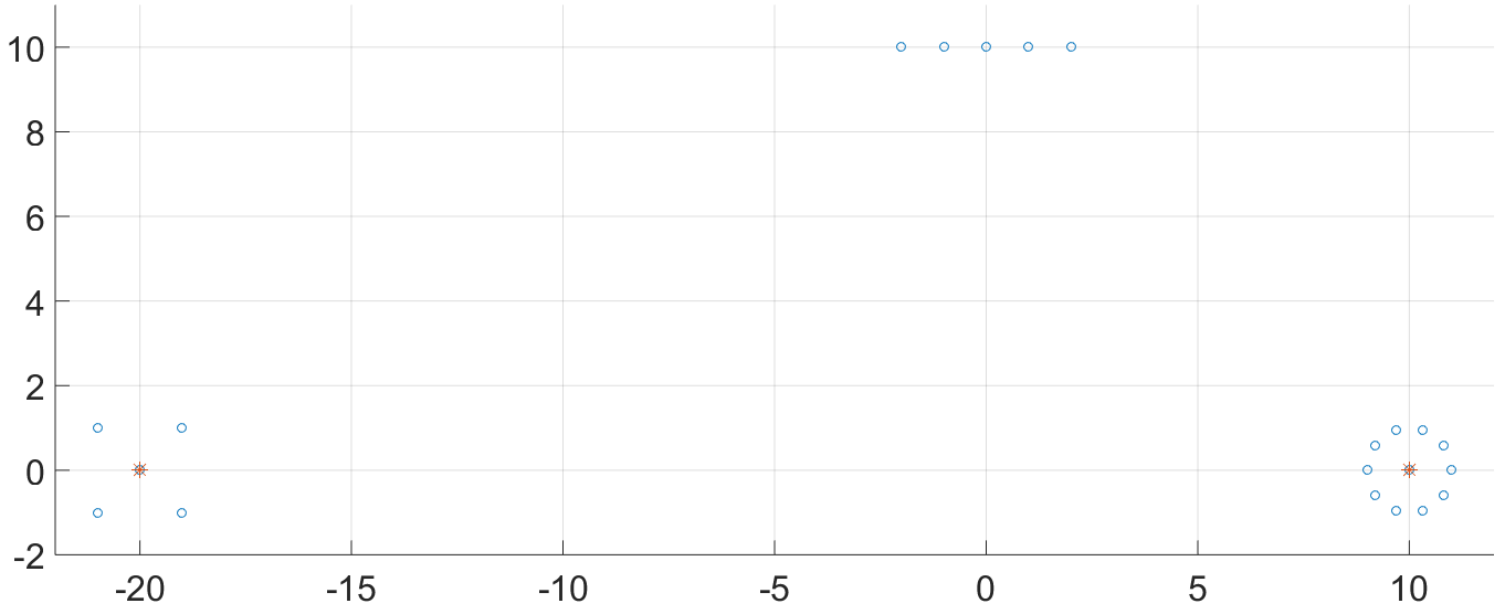
Consider the following figure, which depicts a labeled set of 2D points.

1. **All** points are provided as input for a 1-nn classifier. I.e., each point is classified based on the values of all points, including itself). What is the number of errors?
2. Now, each point is classified with only the other points (we do not use the point to classify itself, leave one out). Calculate the number of errors of a k-NN classifier for $k = 1,3,5,7,11$.

3. Which value of k would you pick?




10
8
6
4
2
0
-2
-20
-15
-10
-5
0
5
10


10
8
6
4
2
0
-2
-20
-15
-10
-5
0
5
10