P1. k-Nearest Neighbor Classifier (25 pts: 3+3+5+3+3+5+3)

In Figure 1(a) and Figure 1(b), we are given 9 data points each in the 2-D space. A point labeled with a circle is in Class 1 (with label 1), and a point labeled with a square is in Class 2 (with label 2). For each figure, assume that the lower-left corner is the origin (0, 0), and the grid is of unit length, hence the upper-right corner has coordinate (12, 12). **Note that the x-coordinate comes before the y-coordinate**, and we use Euclidean (L_2) distance to measure the distance between points.

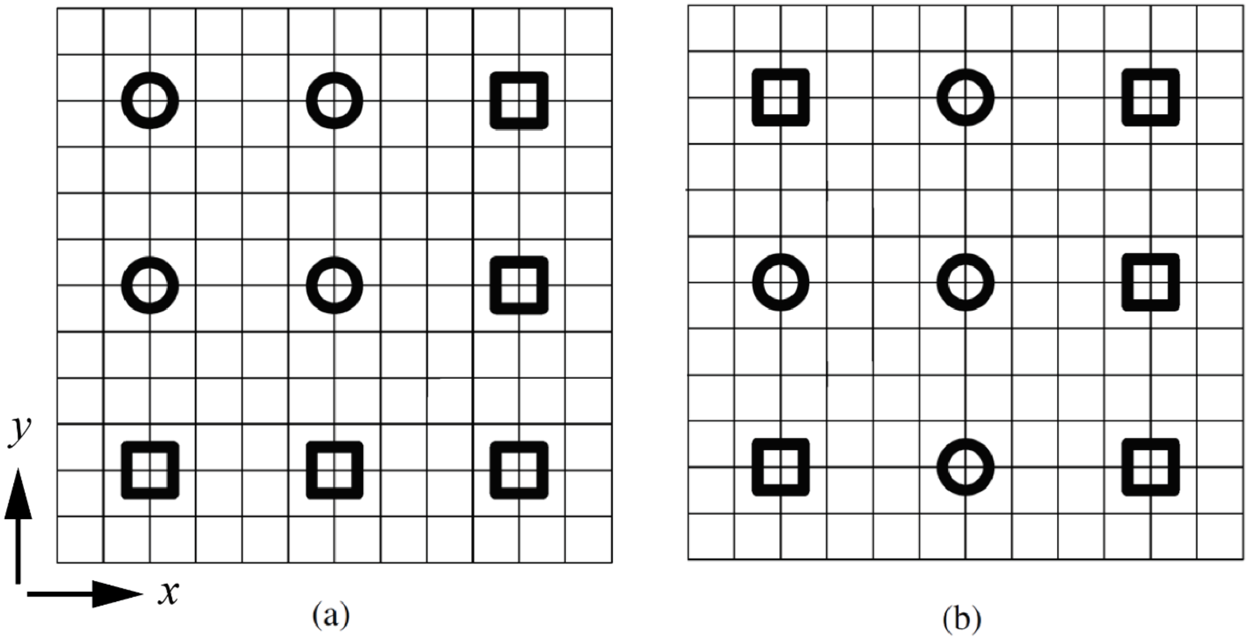


Figure 1: Training Data Set for k-NN Classifiers

(A) Using training set in Figure 1(a), for point **X** = (3, 6), what is the 1-NN label?

Solution: The 1-NN label is Class 1. [3pts]

(B) Using training set in Figure 1(a), for point **X** = (3, 11), what is the 1-NN label?

Solution: The 1-NN label is Class 1. [3pts]

(C) Using training set in Figure 1(a), for point **X** = (3, 11), what is the 7-NN label?

Solution: The 1-NN label is Class 1. [5pts]

(D) Using training set in Figure 1(b), for point **X** = (3, 6), what is the 1-NN label?

Solution: The 1-NN label is Class 1. [3pts]

(E) Using training set in Figure 1(b), for point **X** = (3, 11), what is the 1-NN label?

Solution: The 1-NN label is Class 2. [3pts]

(F) Using training set in Figure 1(b), for point **X** = (3, 11), what is the 7-NN label?

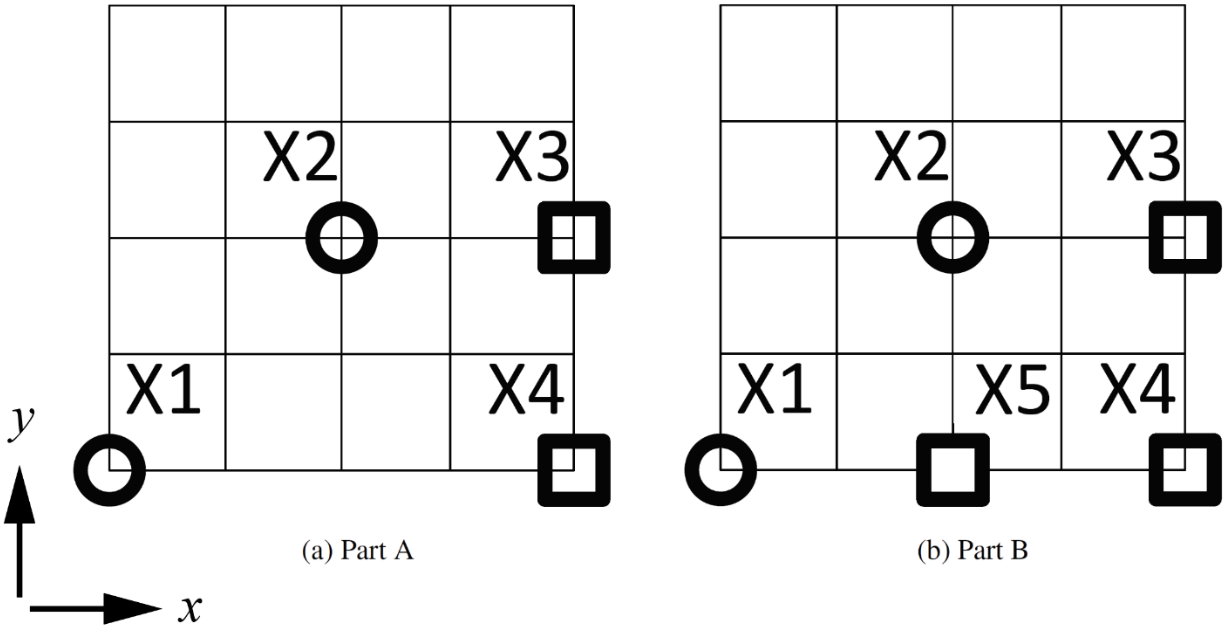
Solution: The 1-NN label is Class 2. [5pts]

(G) Given a training set with N labeled points, what is the time complexity to compute the label of a new point **Y** using 1-NN? Write your answer using asymptotic notations as a function of N.

Solution: The time complexity is $O(n)$ or $\Theta(n)$. [3pts]

Grading: 2pts for $O(n\log n)$ or $\Theta(n\log n)$

P2. Support Vector Machine (25 pts: 1+7+12+5)



(a) Part A

(b) Part B

Figure 2: Problem 2: SVM

In Figure 2(a) and Figure 2(b), we are given some data points in 2-D space, and we aim to train a hard margin linear SVM classifier. A point labeled with a circle is in Class 1 (with label 1), and a point labeled with a square is in Class 2 (with label -1). For each figure, assume that the lower-left corner of the grid is at the origin (0, 0), and the length of each cell in the grid is 1, hence the upper-right corner of the grid has coordinate (4, 4). **Note that the x-coordinate comes before the y-coordinate**.

(A) For the case in Figure 2(a), what is the size of the margin of your SVM classifier?

Solution: The size of the margin is 1. [1pt]

Grading: 0.5pt for a margin of 2

(B) Suppose we add a new data sample (with label -1) at position (2, 0) (as shown in Figure 2(b)). Will this addition change the decision boundary of the SVM classifier in problem (A)? If so, what is the size of the margin of the new SVM classifier?

Solution: Yes. [3pts] The size of the margin of the new SVM classifier is $\frac{\sqrt{2}}{2}$. [4pts]

Grading: 2pts for a margin of $\sqrt{2}$

(C) Suppose we consider the data points shown in Figure 2(a). Write down the (convex) primal optimization formulation for this problem. Write down the dual optimization formulation for this problem. Suppose the solution to the Lagrange dual problem is $[0,0.5.0.5,0]^{\mathrm{T}}$, where the i-th element of the solution corresponds to point Xi for $\mathrm{i} = 1{,}2{,}3{,}4$. What are the support vectors of the SVM classifier?

Solution: Let $\mathbf{w} = [w_1, w_2]^{\mathrm{T}}$, plugging in the coordinates of the training data to the general formulation of primal problem, namely

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \ \text{ s.t. } t_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b) \geq 1,\ n = 1, 2, \ldots, N,$$

with $\mathrm{X1} = [0,0]^{\mathrm{T}}$, $\mathrm{X2} = [2,2]^{\mathrm{T}}$, $\mathrm{X3} = [4,2]^{\mathrm{T}}$, and $\mathrm{X4} = [4,0]^{\mathrm{T}}$, we have the primal problem

$$\begin{aligned} &\min_{w_1,w_2,b} \frac{1}{2}\left(w_1^2 + w_2^2\right) \\ &\text{s.t.} \qquad b \geq 1 \\ &\qquad\quad 2w_1 + 2w_2 + b \geq 1 \qquad\qquad \text{[4pts].} \\ &\qquad\quad -4w_1\text{-}2w_2 - b \geq 1 \\ &\qquad\quad -4w_1 - b \geq 1 \end{aligned}$$

The Lagrangian function is

$$L(\mathbf{w}, b, \boldsymbol{a}) = \frac{1}{2}\left(w_1^2 + w_2^2\right) - a_1(b-1) - a_2(\ 2w_1 + 2w_2 + b - 1) - a_3(-4w_1\text{-}2w_2 - b - 1) - a_4(-4w_1 - b - 1)$$

Again, plugging in the coordinates of the training data to the general formulation of dual problem, namely

$$\begin{aligned}&\max_{\boldsymbol{a}} L(\boldsymbol{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n{}^{\mathrm{T}}\mathbf{x}_m \\ &\text{s.t. } a_n \geq 0,\, n = 1,2,\ldots,N \qquad\qquad , \\ &\quad \sum_{n=1}^{N} a_n t_n = 0\end{aligned}$$

we have the dual optimization formulation for this problem

$$\begin{aligned}&\max_{a_1,a_2,a_3} -\frac{1}{2}\left(8a_2^2 + 20a_3^2 + 16a_4^2 - 24a_2a_3 - 16a_2a_4 + 32a_3a_4\right) + a_1 + a_2 + a_3 + a_4 \\ &\text{s.t.} \qquad a_1 \geq 0, a_2 \geq 0, a_3 \geq 0, a_4 \geq 0 \\ &\qquad\quad a_1 + a_2 - a_3 - a_4 = 0\end{aligned}$$

[6pts]

Alternatively, we can have set the derivatives of the Lagrangian function to zero:

$$\begin{aligned}\frac{\partial L(\mathbf{w}, b, \boldsymbol{a})}{\partial w_1} &= w_1 - 2a_2 + 4a_3 + 4a_4 = 0 \\ \frac{\partial L(\mathbf{w}, b, \boldsymbol{a})}{\partial w_2} &= w_2 - 2a_2 + 2a_3 = 0 \qquad . \\ \frac{\partial L(\mathbf{w}, b, \boldsymbol{a})}{\partial b} &= -a_1 - a_2 + a_3 + a_4 = 0\end{aligned}$$

Substituting the above equations into the Lagrangian function and using the definition of the dual formulation, we obtain the same dual formulation.

The support vectors are X2 and X3. [2pts]

Grading: 6pts for writing the correct dual problem; take 1 mark off for missing a constraint in primal or dual problems; 4pts for writing the dual problem in abstract form; 2pts for writing the primal problem in abstract form; 1pt for writing the correct Lagrangian function, but getting the incorrect dual problem; 3pts for getting the correct derivatives of the Lagrangian function.

(D) We now consider training SVM on a general training set. Suppose that the training set is linearly separable. The data points in the training set are in a high-dimensional space. What is the smallest possible number of support vectors?

Solution: The smallest possible number of support vectors is 2. [5pts]

P3. K-Means Clustering (25 pts: 9+12+4)

Given N data points x_i $(i = 1,2,\ldots N)$, K-means will group them into K clusters by minimizing the loss/cost/distortion function $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2$, where μ_k is the cluster center of the kth cluster, and $r_{nk} = 1$ if $x_n$belongs to the kth cluster and $r_{nk} = 0$ otherwise. In this question, we will use the following iterative procedure.

- Initialize the cluster centers $\mu_k, k = 1,2,\ldots K$
- Iterate until convergence
  - Step 1: Update the cluster assignments r_{nk} for each data point x_n
  - Step 2: Update the cluster center μ_k for each cluster k

Suppose we are given 6 data points in 1-D space: $x_1 = -3, x_2 = -1, x_3 = 0, x_4 = 1, x_5 = 3, x_6 = 4$.

(A) [9 points] Suppose the initial cluster centers are $\mu_1 = -4$ and $\mu_2 = 5$. If we only run K-means for one iteration on the above dataset (i.e., $K = 2$), what is the cluster assignment for each data point after Step 1? What are the updated cluster centers after Step 2?

Solution: After Step 1, cluster 1 contains the points x_1, x_2 and x_3. Cluster 2 contains points x_4, x_5, and x_6 [5pts, taking 2pts off for each wrong cluster]. After Step 2, $\mu_1 = -\frac{4}{3}$ and $\mu_2 = \frac{8}{3}$. [2pts for each cluster center]

(B) [12 points] Let the K-means algorithm with the above initialization run to convergence (i.e. the cluster assignment of all data points remains unchanged). What are the values of the two cluster centers? What is the cluster assignment? What is the corresponding value of the loss function J?

Solution: $\mu_1 = -\frac{4}{3}$ and $\mu_2 = \frac{8}{3}$. [4pts] Cluster 1 contains the points x_1, x_2 and x_3. Cluster 2 contains points x_4 , x_5 , and x_6 [4pts, taking 2pts off for each wrong cluster]. $J = \left(-3+\frac{4}{3}\right)^2 + \left(-1+\frac{4}{3}\right)^2 + \left(0+\frac{4}{3}\right)^2 + \left(1-\frac{8}{3}\right)^2 + \left(3-\frac{8}{3}\right)^2 + \left(4-\frac{8}{3}\right)^2 = \frac{84}{9}$. [4pts]

(C) [4 points] Assume that there are 4 data points in 2-D space: X_1 = (0, 0), X_2 = (0, 1), X_3 = (1, 0), X_4 = (1, 1). We want to run K-means algorithm to cluster them into two nonempty clusters C_1 and C_2. How many possible clustering results will we have? Assume that the initialization satisfies the condition $\mu_1 \neq \mu_2$. Note that $C_1 = \{X_1, X_2\}, C_2 = \{X_3, X_4\}$ and $C_1 = \{X_3, X_4\}, C_2 = \{X_1, X_2\}$ are two different clustering results. Among all the possible clustering results, what is the maximum size of cluster C_1?

Solution: We will have 12 possible clustering results. [2pts] The maximum size of cluster C_1 is 3. [2pts]

P4. Unconstrained Optimization (25 pts: 5+5+5+5+5)

We want to minimize a function

$$f(x) = (x_1 - 2)^2 + 2(x_2 - 3)^2$$

in a two-dimensional Euclidean space where $x = [x_1, x_2]^{\mathrm{T}}$. Answer the following questions.

(A) [5 points] What is the gradient at point $\bar{x} = [4, 4]^{\mathrm{T}}$?

Solution: Since $\nabla f(x) = \begin{bmatrix} 2x_1 - 4 \\ 4x_2 - 12 \end{bmatrix}$, $\nabla f(\bar{x}) = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$ [5pts].

(B) [5 points] What is the steepest descent direction at point $\bar{x}$?

Solution: The steepest descent direction is $-\nabla f(\bar{x}) = \begin{bmatrix} -4 \\ -4 \end{bmatrix}$ [5pts].

(C) [5 points] Suppose we want to perform the backtrack line search at point $\bar{x}$ along direction $-\nabla f(\bar{x})$, using $\alpha = 0.1$ and $\beta = 0.9$. What is the computed step-size?

Solution: The backtrack line search finds the largest value of t in $\{1, 0.9, 0.9^2, 0.9^3, \ldots\}$ such that $f(\bar{x} - t\nabla f(\bar{x})) \le f(\bar{x}) - 0.1t\left\|\nabla f(\bar{x})\right\|^2$. By numerical calculation, this is equivalent to $t \le 0.6$. The computed step size is $0.9^5 \approx 0.59049$ [5pts].

Grading: 2pts for a correct general formulation ($f(\bar{x} - t\nabla f(\bar{x})) \le f(\bar{x}) - 0.1t\left\|\nabla f(\bar{x})\right\|^2$) but wrong final result.

(D) [5 points] Suppose we want to perform the exact line search at point $\bar{x}$ along direction $-\nabla f(\bar{x})$, what is the computed step-size?

Solution: The exact line search finds the optimal t by solving the problem $\min_{t\in\mathbb{R}} f(\bar{x} - t\nabla f(\bar{x}))$, which is equivalent to $\min_t 48t^2 - 32t + 6$. The computed step size is $\frac{1}{3}$ [5pts].

Grading: 2pts for a correct general formulation ($\min_{t\in\mathbb{R}} f(\bar{x} - t\nabla f(\bar{x}))$). 4pts for $\min_t 48t^2 - 32t + 6$.

(E) Write the Lagrangian function for the following optimization problem:

$$\begin{aligned} &\min_{x_1,x_2} \ (x_1-2)^2+2(x_2-3)^2 \\ &\text{s.t.} \quad x_1+x_2=1 \\ &\qquad\quad x_1\geq 0, x_2\geq 0 \end{aligned}$$

Solution: The Lagrangian function is

$L(x_1,x_2,\lambda,\alpha_1,\alpha_2)=(x_1-2)^2+2(x_2-3)^2+\lambda(x_1+x_2-1)-\alpha_1x_1-\alpha_2x_2$ [5pts]

where $\lambda,\alpha_1,\alpha_2$ are Lagrange multipliers and $\alpha_1\geq 0,\alpha_2\geq 0$.

Grading: Take 1pt off for missing a term associate with a Lagrange multiplier (λ, x_1, x_2) in the Lagrangian function.