CS 189 Introduction to Machine Learning
Fall 2018

Midterm

After the exam starts, please write your student ID (or name) on **EVERY PAGE**.

There are **4** questions for a total of **13** parts. You may consult your sheet of notes. Calculators, phones, computers, and other electronic devices are not permitted. There are **17** pages on the exam. **Notify a proctor immediately if a page is missing.** You may, without proof, use theorems and lemmas that were proven in the notes and/or in lecture, unless we explicitly ask for a derivation. However, you must clearly state what theorem or lemma you are using and where/how you are using it.

Please write legibly if you want full credit on all problems.

You have 75 minutes.

PRINT and SIGN Your Name: ________________, ________________, ________________
(last) (first) (signature)

PRINT Your Student ID: ____________________________

Person before you: ____________________, ____________________
(name) (SID)

Person behind you: ____________________, ____________________
(name) (SID)

Person to your left: ____________________, ____________________
(name) (SID)

Person to your right: ____________________, ____________________
(name) (SID)

Seat Number: __________

Do not turn this page until your instructor tells you to do so.

1 Finding the Centroid (3 parts, 20 points)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$. We consider computing the centroid of this dataset. Consider the loss function

$$\mathcal{L}(\mathbf{w}) := \frac{1}{2n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{w}\|_2^2.$$

(a) (5 points) First, we compute the gradient of the loss function. **Show that**

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \mathbf{w} - \bar{\mathbf{x}},$$

where $\bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$.

Solution: We have that $\nabla_{\mathbf{w}} \|\mathbf{x}_i - \mathbf{w}\|_2^2 = \nabla(\mathbf{x}_i^\top \mathbf{x}_i - 2\mathbf{x}_i^\top \mathbf{w} + \|\mathbf{w}\|_2^2) = 2\mathbf{w} - 2\mathbf{x}_i$. Hence,

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{2n} \sum_{i=1}^{n} (2\mathbf{w} - 2\mathbf{x}_i) = \mathbf{w} - \bar{\mathbf{x}}.$$

(b) (5 points) **Show that the minimizer of the loss function is given by $\bar{\mathbf{x}}$, i.e.** $\arg\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) = \bar{\mathbf{x}}$. Make sure to justify your answer.

Solution: Since $\mathcal{L}$ is convex, at the minimum the gradient is equal to zero. Thus, by the previous problem, this is when $\mathbf{x} = \bar{\mathbf{x}}$.

(c) (10 points) Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are identically and independently distributed according to a normal distribution with mean $\mathbf{x}_*$ and diagonal covariance, i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_*, \sigma^2 \mathbf{I}_d)$ for $i = 1, \ldots, n$.

Calculate $\mathbb{E}[\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2^2]$**.**

Solution:

$$
\begin{aligned}
\mathbb{E}[\|\bar{\mathbf{x}} - \mathbf{x}_*\|_2^2] &= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i - \mathbf{x}_*\|^2] \\
&= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}_*)\|^2] \\
&= \frac{1}{n^2}\mathbb{E}[\|\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}_*)\|^2] \\
&= \frac{1}{n^2}\mathbb{E}[\sum_{i=1}^{n} \|(\mathbf{x}_i - \mathbf{x}_*)\|^2 + 2\sum_{i \neq j} \langle \mathbf{x}_i - \mathbf{x}_*, \mathbf{x}_j - \mathbf{x}_* \rangle] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{E}[\|(\mathbf{x}_i - \mathbf{x}_*)\|^2 + 2\sum_{i<j} \mathbb{E}[\langle \mathbf{x}_i - \mathbf{x}_*, \mathbf{x}_j - \mathbf{x}_* \rangle]\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{E}[\|(\mathbf{x}_i - \mathbf{x}_*)\|^2]\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma^2 d\right) \\
&= \sigma^2 d/n.
\end{aligned}
$$

2 A Spectral View of Linear Regression (5 parts, 25 points)

Assume we are given training data in the form of the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ where the rows are the d-dimensional feature vectors $\mathbf{x}_i$ and $\mathbf{y} \in \mathbb{R}^n$ which is the vector of the corresponding target values. We do not assume that $\mathbf{X}$ is full rank, and take its rank to be r. Note that $d \leq n$.

Recall that the compact singular value decomposition is $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{n\times d}$, $\mathbf{V} \in \mathbb{R}^{d\times d}$, and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$. We denote the n-dimensional column vectors of $\mathbf{U}$ by $\mathbf{u}_i$ and the d-dimensional column vectors of $\mathbf{V}$ by $\mathbf{v}_i$. Furthermore, let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$.

In this problem, we consider the result of two different linear regression techniques: ridge regression and applying ordinary least squares after using PCA to reduce the feature dimension from d to k (PCA-OLS). In particular, we compare the predicted value $\widehat{y}$ of a new datapoint $\mathbf{x}$ by writing an expression of the form:

$$\widehat{y}(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} = \mathbf{x}^\top \sum_{i=1}^{d} \rho(\sigma_i)\mathbf{v}_i\mathbf{u}_i^\top \mathbf{y}. \tag{1}$$

In the following questions you will find the form of the spectral function $\rho(\sigma)$ for ridge regression and PCA-OLS.

(a) (5 points) Recall that the ridge regression optimizer is defined (for $\lambda > 0$) as

$$\mathbf{w}_{\text{ridge}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 .$$

Show that the closed-form solution for $\mathbf{w}_{\text{ridge}}$ has the form

$$\mathbf{w}_{\text{ridge}} = \mathbf{V}\operatorname{diag}(\rho_\lambda(\sigma_1), \ldots, \rho_\lambda(\sigma_d))\mathbf{U}^\top \mathbf{y},$$

and find the ridge-regression spectral function ρ_λ.

Solution: First, recall that

$$\mathbf{w}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} .$$

Then plugging in the SVD of $\mathbf{X}$,

$$\begin{aligned}\mathbf{w}_{\text{ridge}} &= (\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y} \\ &= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{y}\end{aligned}$$

Thus we see that

$$\rho_\lambda(\sigma_i) = \frac{\sigma_i}{\lambda + \sigma_i^2} .$$

(b) (5 points) Using the expression for $\mathbf{w}_{\text{ridge}}$ from the previous part, **write down the ridge regression predictor function in the form of** (1).

Solution: The resulting prediction for ridge reads

$$\begin{aligned}\hat{\mathbf{y}}_{\text{ridge}} &= \mathbf{x}^\top \mathbf{V}\,\text{diag}\,\Big(\frac{\sigma_i}{\lambda+\sigma_i^2}\Big)\mathbf{U}^\top \mathbf{y} \\ &= \mathbf{x}^\top \sum_{i=1}^{d} \frac{\sigma_i}{\lambda+\sigma_i^2}\mathbf{v}_i \mathbf{u}_i^\top \mathbf{y}\end{aligned}$$

(c) (5 points) The ordinary least squares problem on the reduced k-dimensional PCA feature space (PCA-OLS) can be written as

$$\tilde{\mathbf{w}}_{\text{PCA}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{V}_k\mathbf{w}-\mathbf{y}\|^2$$

where the columns of $\mathbf{V}_k$ consist of the first k right singular vectors of $\mathbf{X}$. This expression embeds the raw feature vectors onto the top k principle components by the transformation $\mathbf{V}_k^\top \mathbf{x}_i$. Assume the PCA dimension is less than the rank of the data matrix, $k \leq r$.

Write down the expression for the optimizer $\tilde{\mathbf{w}}_{\text{PCA}} \in \mathbb{R}^k$ in terms of $\mathbf{U}$, $\mathbf{y}$ and the singular values of $\mathbf{X}$.

Hint: $k \leq r$ implies that the matrix of PCA embedded data matrix $\mathbf{X}\mathbf{V}_k$ is full rank.

Solution: Apply OLS on the new matrix $\mathbf{X}\mathbf{V}_k$ to obtain

$$\begin{aligned}\tilde{\mathbf{w}}_{\text{PCA}} &= [(\mathbf{X}\mathbf{V}_k)^\top(\mathbf{X}\mathbf{V}_k)]^{-1}(\mathbf{X}\mathbf{V}_k)^\top\mathbf{y} \\ &= [\mathbf{V}_k^\top\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top\mathbf{V}_k]^{-1}\mathbf{V}_k^\top\mathbf{X}^\top\mathbf{y} \\ &= \mathbf{\Sigma}_k^{-1}\mathbf{U}_k^\top\mathbf{y}\end{aligned}$$

(d) (5 points) Now, use the expression for $\tilde{\mathbf{w}}_{\text{PCA}}$ from the previous part to **write down the predictor function in the form of** (1). In doing so, you should **define the form of the PCA-OLS spectral function** ρ_k.

Solution: The resulting prediction for PCA reads (note that you need to project it first!)

$$\begin{aligned}\widehat{\mathbf{y}}_{\text{PCA}} &= \mathbf{x}^\top \mathbf{V}_k \tilde{\mathbf{w}}_{\text{PCA}} \\ &= \mathbf{x}^\top \mathbf{V}_k \boldsymbol{\Sigma}_k^{-1} \mathbf{U}_k^\top \mathbf{y} \\ &= \mathbf{x}^\top \sum_{i=1}^{k} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^\top \mathbf{y}\end{aligned}$$

$$\rho_k(\sigma_i) = \begin{cases} \frac{1}{\sigma_i} & i \leq k \\ 0 & i > k \end{cases}$$

(e) (5 points) The ridge regression regularization parameter λ and the PCA dimension k are both hyperparameters that affect the resulting model and predictions. In practice, we would tune these parameters based on the dataset we were given. **Briefly describe a principled method for choosing λ.**

Solution: Cross validation or holdout

3 Classification (3 parts, 25 points)

(a) (5 points) The plots below show labeled data $\{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^2$. For each plot, points corresponding to $y_i = -1$ are denoted by an O, and points corresponding to $y_i = +1$ are denoted by an X. The origin is labeled as the point `(0,0)`. Now, consider classifiers of the form

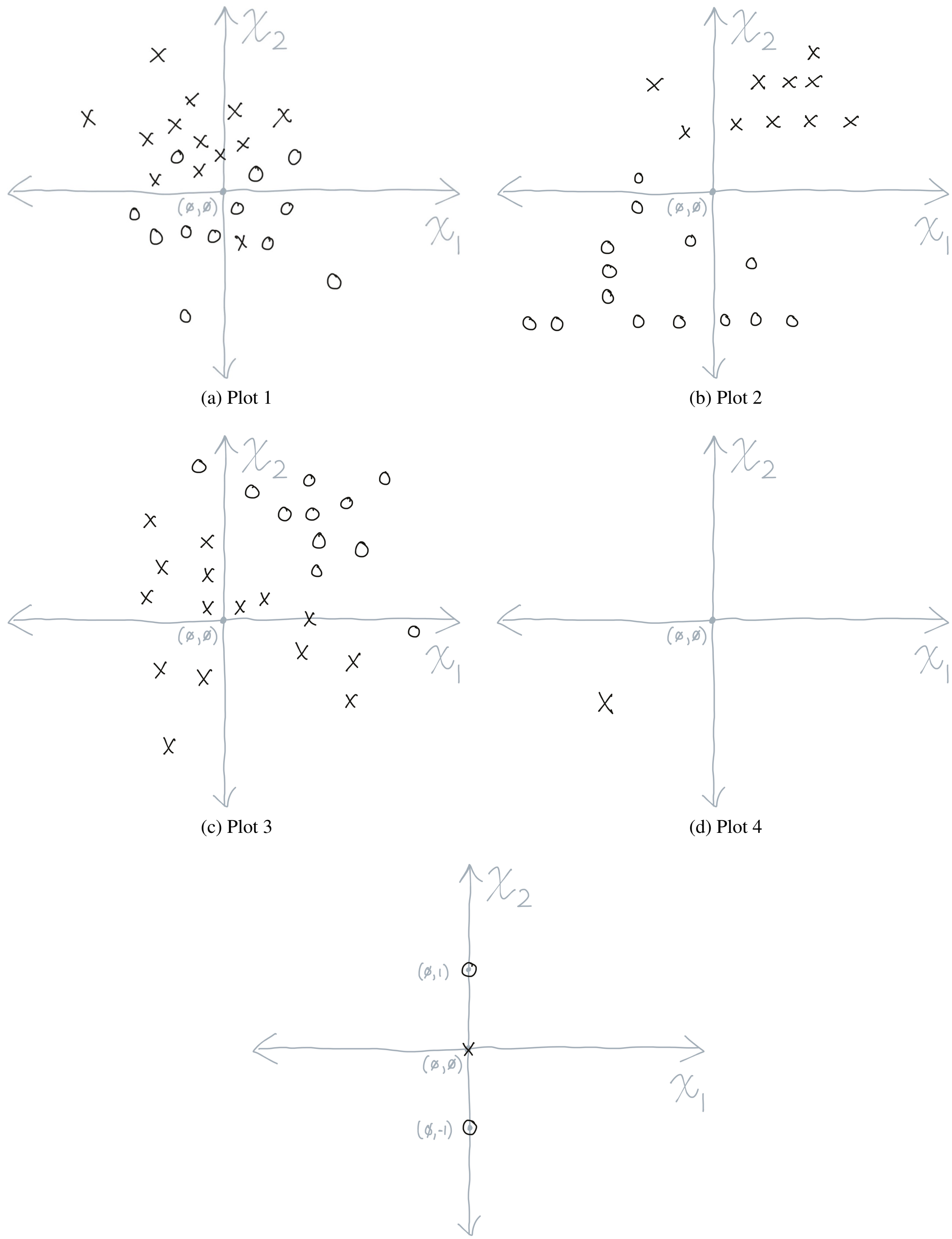
$$\phi_{\mathbf{w}}(\mathbf{x}) = \begin{cases} +1, & \mathbf{w}^\top \mathbf{x} \geq 0 \\ -1, & \mathbf{w}^\top \mathbf{x} < 0 \end{cases}$$

where $\mathbf{w} \in \mathbb{R}^2$.

For each of the five plots, determine if the data can be perfectly classified by a classifier of this form.

- **If so, draw the decision boundary of the classifier on the plot.**
- **If not, write "not separable" in the appropriate cell in the following table.**

Plot	Separable?
1	
2	
3	
4	
5	



(a) Plot 1

(b) Plot 2

(c) Plot 3

(d) Plot 4

(e) Plot 5

Figure 1: Problem 3(a)

Solution: Only the second and the fourth are separable by a linear (not affine!) classifier.

(b) (10 points) Consider the data shown in Figure 2.

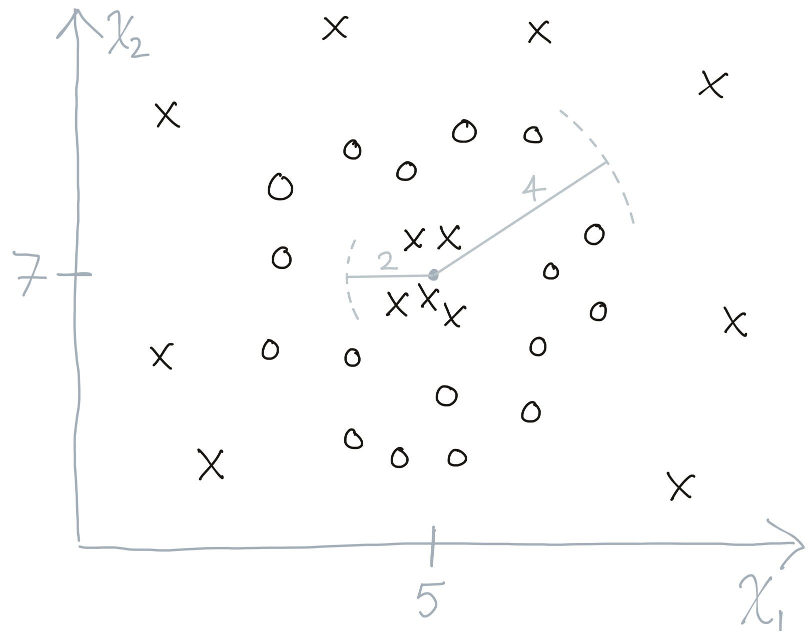


Figure 2: Problem 3(b)

Again, points corresponding to $y_i = -1$ are denoted by an O, and points corresponding to $y_i = +1$ are denoted by an X. Note that the O points (and only the O points) are contained between two circles of radii 2 and 4, both centered at the point $(5,7)$. This data can not be perfectly classified by a classifier described in the previous problem. However, we can make a nonlinear transformation of the data to make it easier to classify. Specifically, we seek a transformation $\varphi(\mathbf{x}) : \mathbb{R}^2 \to \mathbb{R}$ such that each transformed point $z_i = \varphi(\mathbf{x}_i)$ can be perfectly classified by a classifier $h_b : \mathbb{R} \to \{-1,+1\}$ of the form

$$h_b(z) = \begin{cases} +1, & z \geq b \\ -1, & z < b \end{cases}.$$

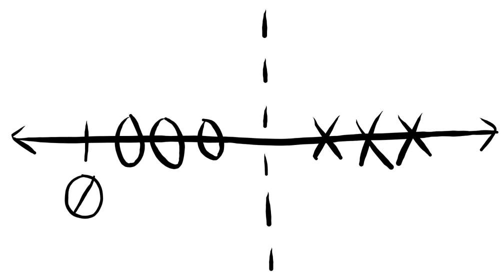
(i) **Give such a transformation** $\varphi(\mathbf{x})$. *(You should not need to estimate exact locations of points.)*

(ii) **Plot the (nonlinear) decision boundary on the original plot** (Figure 2).

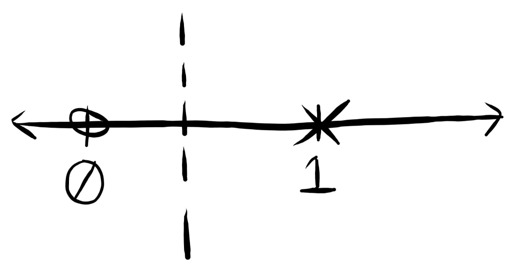
(iii) **Plot the transformed data and the decision boundary in the transformed space $\mathbb{R}$, i.e. on a number line (you should have a tick mark for** 0**).** This plot should be qualitative to illustrate the situation; you do not need to find an explicit b for the decision boundary, nor do you need to exactly plot every transformed point.

Solution: Some example correct solutions for (i) and the corresponding plots (iii):

$$\varphi(\mathbf{x}) = |(x_1 - 5)^2 + (x_2 - 7)^2 - 10|$$
$$\varphi(\mathbf{x}) = |\sqrt{(x_1 - 5)^2 + (x_2 - 7)^2} - 3|$$



$$\varphi(\mathbf{x}) = 1 - \mathbf{1}\{2 \leq \sqrt{(x_1 - 5)^2 + (x_2 - 7)^2} \leq 4\}\,.$$



The (not connected) decision boundary for part (ii) should be the two circles of radii 2 and 4 centered at $(5, 7)$.

(c) (10 points) Now consider classifying two data points $x_1 = (a, b)$, $x_2 = (-a, -b)$, with labels $y_1 = +1$ and $y_2 = -1$, respectively, shown in Figure 3.

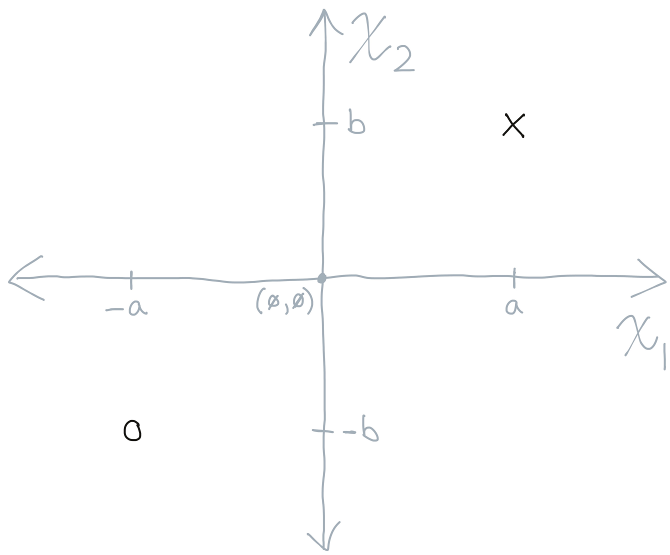


Figure 3: Problem 3(c)

For this data, calculate the form of the maximum margin separating hyperplane which goes through the origin. Make sure you justify your answer mathematically. Recall that for linear classifiers, the maximum margin is defined as:

$$\max_{\mathbf{w}\in\mathbb{R}^d} \min_{1\leq i\leq n} \left(\frac{\mathbf{w}^\top \mathbf{x}_i}{\|\mathbf{w}\|_2} y_i \right)$$

Solution: There are only two data points, so the margin is

$$\begin{aligned}
&\max_{w} \min_{i=1,2} \left(\frac{w^\top x_i}{\|w\|_2} y_i \right) \\
&= \max_{w:\|w\|_2=1} \min \left\{ w^\top \begin{pmatrix} a \\ b \end{pmatrix} (1), w^\top \begin{pmatrix} -a \\ -b \end{pmatrix} (-1) \right\} \\
&= \max_{w:\|w\|_2=1} \min \left\{ w^\top \begin{pmatrix} a \\ b \end{pmatrix}, w^\top \begin{pmatrix} a \\ b \end{pmatrix} \right\} \\
&= \max_{w:\|w\|_2=1} w^\top \begin{pmatrix} a \\ b \end{pmatrix}
\end{aligned}$$

The maximizing w is the unit vector in the direction $(a, b)^\top$, so we have that the maximizing hyperplane is defined by

$$\{x : x^\top w = 0\}$$

where $w = \begin{pmatrix} a \\ b \end{pmatrix}$.

4 Checking Kernels (2 parts, 10 points)

Recall that for a function k to be a valid kernel, it must be symmetric in its arguments and its Gram matrices must be positive semi-definite. More precisely, for every sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the Gram matrix

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & k(\mathbf{x}_i, \mathbf{x}_j) & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

must be positive semi-definite. Also, recall that a matrix is positive semi-definite if it is symmetric and all its eigenvalues are non-negative.

(a) (5 points) **Give an example of two positive semi-definite matrices A_1 and A_2 in $\mathbb{R}^{2\times 2}$ such that $A_1 - A_2$ is not positive semi-definite.**

As a consequence, **show that the function k defined by $k(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_i, \mathbf{x}_j) - k_2(\mathbf{x}_i, \mathbf{x}_j)$ is not necessarily a kernel even when k_1 and k_2 are valid kernels.**

Solution: Take $A_1 = 0_2$ and $A_2 = I_2$. We can define k_1 and k_2 to have 2×2 Gram matrices equal to A_1 and A_2 respectively.

(b) (5 points) **Show that the function k defined by $k(\mathbf{x}_i, \mathbf{x}_j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2$ is not a valid kernel.**

Solution: Consider the dataset $\{x_1, x_2\} = \{0, 1\}$. The gram matrix induced by k on this dataset is $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The eigenvalues of this matrix are $-1, 1$, which means this matrix is not positive semidefinite. Hence k is not a valid kernel.

