CSE 575: Homework #2

Due: October 24, 2022

Problem 1

Consider the following data points, also plotted in Figure 1:

$$\text{Labeled } +1 : (-2, 0), (0, 2)$$
$$\text{Labeled } -1 : (2, 2), (2, 0), (3, -1).$$

The positive data points are represented as blue circles, and the negative points as red triangles. Given this data, we wish to train a hard-margin linear SVM classifier.

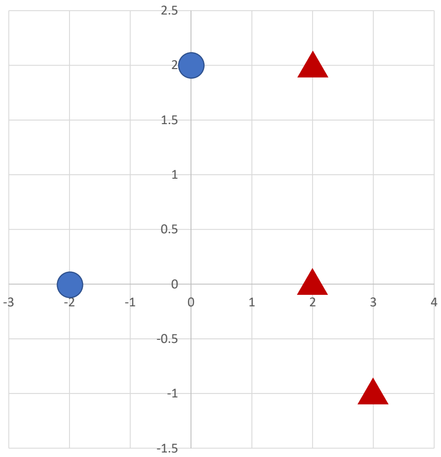


Figure 1: Data points

a) (6pt) What is the optimization problem for the maximum margin classifier (also called the *primal* problem) on this dataset? Denote the optimization variables by w_1, w_2, b.

Solution: The primal problem is:

$$\begin{aligned} \min\ & \frac{1}{2}(w_1^2 + w_2^2) \\ \text{s.t.}\ & -2w_1 + b - 1 \geq 0 \\ & 2w_2 + b - 1 \geq 0 \\ & -2w_1 - 2w_2 - b - 1 \geq 0 \\ & -2w_1 - b - 1 \geq 0 \\ & -3w_1 + w_2 - b - 1 \geq 0 \end{aligned}$$

b) (6pt) What is the dual formulation of the problem in part (a)? Denote the optimization variables by a_1, a_2, a_3, a_4, a_5

Solution: The lagrange function is:

$$\begin{aligned}L(w,b,a) =&\frac{1}{2}(w_1^2+w_2^2) - a_1(-2w_1+b-1) - a_2(2w_2+b-1) - a_3(-2w_1-2w_2-b-1)\\ &- a_4(-2w_1-b-1) - a_5(-3w_1+w_2-b-1)\end{aligned}$$

where the Lagrange multipliers a_1, a_2, a_3 and a_4 all have to be non-negtive.
To get the dual function, we take partial derivatives and set to zero:

$$\begin{aligned}\frac{\partial L}{\partial w_1} &= w_1 + 2a_1 + 2a_3 + 2a_4 + 3a_5 = 0\\ \frac{\partial L}{\partial w_2} &= w_2 - 2a_2 + 2a_3 - a_5 = 0\\ \frac{\partial L}{\partial b} &= -a_1 - a_2 + a_3 + a_4 + a_5 = 0\end{aligned}$$

Substituting into the Lagrange function, we get the dual function:

$$\begin{aligned}L_D(a) =&\sum_{i=1}^{5} a_i - \frac{1}{2}\sum_{i=1}^{5}\sum_{j=1}^{5} a_i a_j y_i y_j x_i^T x_j\\ =&a_1 + a_2 + a_3 + a_4 + a_5 - \frac{1}{2}(4a_1^2 + 4a_2^2 + 8a_3^2 + 4a_4^2 + 10a_5^2\\ &+ 8a_1a_3 + 8a_1a_4 + 12a_1a_5 - 8a_2a_3 + 4a_2a_5 + 8a_3a_4 + 8a_3a_5 + 12a_4a_5)\end{aligned}$$

Hence the dual optimal problem is:

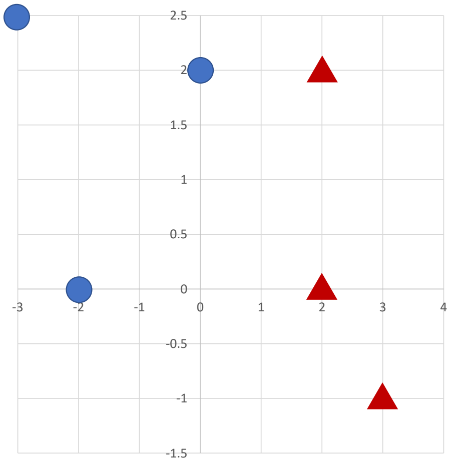
$$\begin{aligned}&\text{max } L_D(a)\\ &\text{s.t. } \ a_i \geq 0,\ i \in \{1,2,3,4,5\}\\ &\qquad a_1 + a_2 - a_3 - a_4 - a_5 = 0\end{aligned}$$

c) (6pt) What is the decision boundary of the SVM classifier? *Hint: You may make a geometric argument using **figure**. You are not required to solve the optimization problem.*

Solution: The vertical line x = 1. Any non-vertical line would have a smaller margin due to points $(0, 2)$ and $(2, 2)$.

d) (6pt) Suppose we add a new data point at $(-3, 2.5)$ (as shown in the figure below). How will the decision boundary change in this case?

Solution: The decision boundary doesn't change. The decision boundary is only decided by the support vectors, in this case $(0, 2)$ and $(2, 2)$. Adding the new data point does not change the assignment of the support vectors, therefore won't change the decision boundary.



Problem 2

Figure 2 plots decision boundaries of SVM classifiers using different kernels and/or different slack penalty C. The data points labeled $+1$ and -1 are represented by circles and triangles, respectively. The support vectors for each decision boundary is illustrated as solid circles and triangles. For each of the following scenarios, find the matching plot from Figure 2 (there is a one-to-one match). Justify each choice briefly in 1–2 sentences.

a) (5pt) A hard-margin linear SVM.

Solution: (a). A hard-margin linear SVM ensures a correct classification for every data point and the distances are larger than or equal to the margin.

b) (5pt) A soft-margin linear SVM with $C = 0.1$.

Solution: (d). A soft-margin SVM with a small penalty term allows more data points to be misclassified or have distances smaller than the margin.

c) (5pt) A soft-margin linear SVM with $C = 10$.

Solution: (b). A soft-margin SVM with a large penalty term results in fewer misclassified samples.

d) (5pt) A hard-margin kernel SVM with $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^2$.

Solution: (e). Quadratic kernels can have elliptical or hyperbolic decision boundaries.

e) (5pt) A hard-margin kernel SVM with $k(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2)$.

Solution: (c). Gaussian kernel may contain monomials of arbitrary degree, corresponding to an infinite dimensional feature mapping. Therefore the contour is irregular.

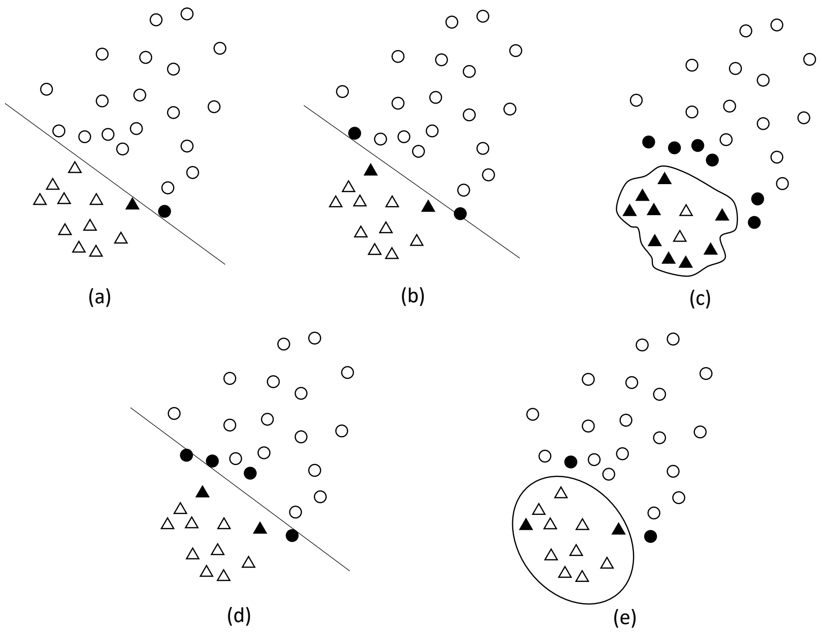


Figure 2: SVM decision boundaries

Problem 3

Recall the perceptron $y(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + b)$ where $f(a)$ is the step function:

$$f(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

In this question, you will construct a multilayer perceptron with a single hidden layer to represent the XNOR function.

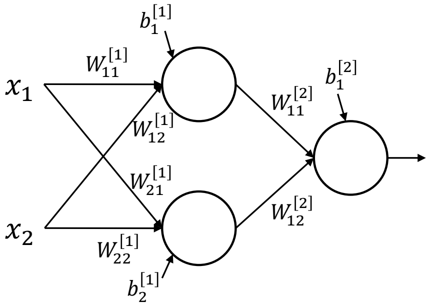
a) (8pt) Consider a perceptron with a two-dimensional input: $y(x_1, x_2) = f(w_1x_1 + w_2x_2 + b)$. What is the value of w_1, w_2, b such that this perceptron represents the logical conjunction (AND) between x_1 and x_2? Assume that the inputs are Boolean: $x_1, x_2 \in \{0, 1\}$.

Solution: There are multiple valid combinations of (w_1, w_2, b). One example is $w_1 = 1$, $w_2 = 1$, $b = -2$.

b) (8pt) Again, consider a perceptron with two Boolean inputs: $y(x_1, x_2) = f(w_1x_1 + w_2x_2 + b)$. What is the value of w_1, w_2, b such that this perceptron represents the logical NOR between x_1 and x_2? In other words, the function should output 1 if both inputs are 0, and output 0 otherwise.

Solution: Again, there are multiple valid combinations of (w_1, w_2, b). One example is $w_1 = -1$, $w_2 = -1$, $b = 0$.

c) (10pt) The logical exclusive nor (XNOR) between x_1 and x_2 is 1 if the inputs are equal (both 0 or both 1); and 0 otherwise. **(i)** Show why XNOR cannot be represented by a single perceptron. You may use a geometric argument with a simple figure (the figure may be a hand-drawn image). **(ii)** Derive the



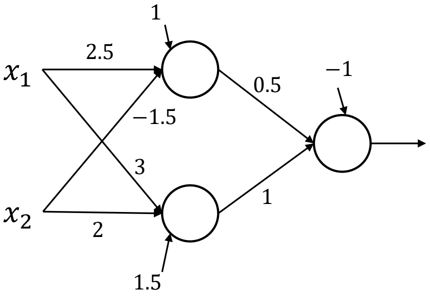
weights for the following multilayer perceptron such that it outputs the XNOR between x_1 and x_2. *Hint: can you write XNOR using AND, NOR, and OR?*

Solution:
(i) The XNOR can be considered as a classification problem where class 0 contains $(0,1)$ and $(1,0)$, and class 1 contains $(0,0)$ and $(1,1)$. The distribution of the data points are not linearly separable.
(ii) The XNOR can be expressed as x_1 XNOR x_2 = (x_1 AND x_2) OR (x_1 NOR x_2). One valid combination of (w_1, w_2, b) for OR is $(1,1,-1)$. Taking the parameters from AND and NOR and combining with OR, we can set the parameters as follows:

$$w_{11}^{[1]} = 1, w_{12}^{[1]} = 1, b_1^{[1]} = -2$$
$$w_{21}^{[1]} = -1, w_{22}^{[1]} = -1, b_2^{[1]} = 0$$
$$w_{11}^{[2]} = 1, w_{12}^{[2]} = 1, b_1^{[2]} = -1$$

Problem 4

The figure below shows a feedforward neural network with a single two-unit hidden layer and an output unit. Assume all neurons use the sigmoid activation function. Let $\mathbf{x} = (x_1, x_2) = (0.5, 0.1)$ be an example with label 1.



a) (8pt) By forward propagation, calculate the output values of the neurons for the input $\mathbf{x}$. In other words, what are the values of $z_1^{[1]}, z_2^{[1]}, z_1^{[2]}$?

Solution:

$$\begin{aligned}
a_1^{[1]} &= 0.5 \times 2.5 + 0.1 \times -1.5 + 1 = 2.1 \\
a_2^{[1]} &= 0.5 \times 3 + 0.1 \times 2 + 1.5 = 3.2 \\
z_1^{[1]} &= \sigma(a_1^{[1]}) = 0.891 \\
z_2^{[1]} &= \sigma(a_2^{[1]}) = 0.961 \\
a_1^{[2]} &= 0.5 \times z_1^{[1]} + 1 \times z_2^{[1]} - 0.1 = 0.406 \\
z_1^{[2]} &= \sigma(a_1^{[2]}) = 0.600
\end{aligned}$$

b) (10pt) Suppose the error function is the squared error: $E = \frac{1}{2}(y(\mathbf{x}) - t)^2$ where t is the label for $\mathbf{x}$ and $y(\mathbf{x})$ the output of the neural network given $\mathbf{x}$. By backpropagation, calculate the following partial derivatives (note that $W_{ij}^{[l]}$ denotes the weight associated with the j-th input of the i-th node in layer l):

$$\frac{\partial E}{\partial W_{11}^{[1]}}, \quad \frac{\partial E}{\partial W_{12}^{[1]}}, \quad \frac{\partial E}{\partial W_{21}^{[1]}}, \quad \frac{\partial E}{\partial W_{22}^{[1]}}, \quad \frac{\partial E}{\partial W_{11}^{[2]}}, \quad \frac{\partial E}{\partial W_{12}^{[2]}}$$

Solution:

$$\delta_1^{[2]} = \frac{\partial E}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial a_1^{[2]}} = (z_1^{[2]} - t) z_1^{[2]}(1 - z_1^{[2]}) = -0.400 \times 0.600 \times 0.400 = -0.096$$

$$\frac{\partial E}{\partial W_{11}^{[2]}} = \delta_1^{[2]} \cdot \frac{\partial a_1^{[2]}}{\partial W_{11}^{[2]}} = \delta_1^{[2]} \cdot z_1^{[1]} = -0.096 \times 0.891 = -0.0855$$

$$\frac{\partial E}{\partial W_{12}^{[2]}} = \delta_1^{[2]} \cdot \frac{\partial a_1^{[2]}}{\partial W_{12}^{[2]}} = \delta_1^{[2]} \cdot z_2^{[1]} = -0.096 \times 0.961 = -0.0923$$

$$\delta_1^{[1]} = \frac{\partial E}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial a_1^{[2]}} \cdot \frac{\partial a_1^{[2]}}{\partial z_1^{[1]}} \cdot \frac{\partial z_1^{[1]}}{\partial a_1^{[1]}} = \delta_1^{[2]} \cdot W_{11}^{[2]} \cdot z_1^{[1]}(1 - z_1^{[1]}) = -0.096 \times 0.500 \times 0.891(1 - 0.891) = -0.0047$$

$$\frac{\partial E}{\partial W_{11}^{[1]}} = \delta_1^{[1]} \cdot \frac{\partial a_1^{[1]}}{\partial W_{11}^{[1]}} = \delta_1^{[1]} \cdot x_1 = -0.0047 \times 0.500 = -0.00235$$

$$\frac{\partial E}{\partial W_{12}^{[1]}} = \delta_1^{[1]} \cdot \frac{\partial a_1^{[1]}}{\partial W_{12}^{[1]}} = \delta_1^{[1]} \cdot x_2 = -0.0047 \times 0.1 = -0.00047$$

$$\delta_2^{[1]} = \frac{\partial E}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial a_1^{[2]}} \cdot \frac{\partial a_1^{[2]}}{\partial z_2^{[1]}} \cdot \frac{\partial z_2^{[1]}}{\partial a_2^{[1]}} = \delta_1^{[2]} \cdot W_{12}^{[2]} \cdot z_2^{[1]}(1 - z_2^{[1]}) = -0.096 \times 1 \times 0.961(1 - 0.961) = -0.00360$$

$$\frac{\partial E}{\partial W_{21}^{[1]}} = \delta_2^{[1]} \cdot \frac{\partial a_2^{[1]}}{\partial W_{21}^{[1]}} = \delta_2^{[1]} \cdot x_1 = -0.00360 \times 0.5 = -0.00180$$

$$\frac{\partial E}{\partial W_{22}^{[1]}} = \delta_2^{[1]} \cdot \frac{\partial a_2^{[1]}}{\partial W_{22}^{[1]}} = \delta_2^{[1]} \cdot x_2 = -0.00360 \times 0.1 = -0.000360$$

c) (7pt) What will be the weights after a *single* step of gradient descent using the example $\mathbf{x}$ with a learning rate of $\eta = 0.1$?

Problem 4 continued on next page...

Solution:

$$
\begin{aligned}
W_{11}^{[1]} &= 2.5 - 0.1 \times (-0.00235) = 2.500235 \\
W_{12}^{[1]} &= -1.5 - 0.1 \times (-0.00047) = -1.49995 \\
W_{21}^{[1]} &= 3 - 0.1 \times (-0.00180) = 3.00018 \\
W_{22}^{[1]} &= 2 - 0.1 \times (-0.000360) = 2.000036 \\
W_{11}^{[2]} &= 0.5 - 0.1 \times (-0.0855) = 0.50855 \\
W_{12}^{[2]} &= 1 - 0.1 \times (-0.0923) = 1.00923
\end{aligned}
$$