CSE 343: Machine Learning - Assignment 3

Akshat Gupta

November 13, 2023

Section A

(a) **Initialization:**
Assume:

$$w_1 = 2.0 \text{ and } b_1 = 0.5$$
$$w_2 = 1.0 \text{ and } b_2 = 0.5$$

Data:
Three data points with univariate inputs and corresponding targets:

$$\begin{aligned} &\text{Data Point 1:} \quad x_{1_1} = 1.2, \quad y_{\text{true}_1} = 3.0 \\ &\text{Data Point 2:} \quad x_{1_2} = 0.8, \quad y_{\text{true}_2} = 2.5 \\ &\text{Data Point 3:} \quad x_{1_3} = 2.0, \quad y_{\text{true}_3} = 4.0 \end{aligned}$$

Forward Pass:

$$\begin{aligned}
z_1 &= w_1 \cdot x + b_1 \\
h_1 &= ReLU(z_1) = \max(0, z_1) \\
z_2 &= w_2 \cdot h_1 + b_2 \\
h_2 &= z_2 \\
\frac{\partial L}{\partial w_2} &= \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2} = (h_2 - y_{\text{true}}) \cdot 1 \cdot h_1 \\
\frac{\partial L}{\partial b_2} &= \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial b_2} = (h_2 - y_{\text{true}}) \cdot 1 \cdot 1 \\
\frac{\partial L}{\partial w_1} &= \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1} = (h_2 - y_{\text{true}}) \cdot 1 \cdot w_2 \cdot (1 \text{ if } z_1 > 0 \text{ else } 0) \cdot x \\
\frac{\partial L}{\partial b_1} &= \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1} = (h_2 - y_{\text{true}}) \cdot 1 \cdot w_2 \cdot (1 \text{ if } z_1 > 0 \text{ else } 0) \cdot 1 \\
w_1 &= w_1 - \eta \frac{\partial L}{\partial w_1} \\
w_2 &= w_2 - \eta \frac{\partial L}{\partial w_2} \\
b_1 &= b_1 - \eta \frac{\partial L}{\partial b_1} \\
b_2 &= b_2 - \eta \frac{\partial L}{\partial b_2}
\end{aligned}$$

For Data Point 1:

$$
\begin{aligned}
z_{1_1} &= 2.0 \cdot 1.2 + 0.5 = 2.9 \\
h_{1_1} &= \max(0, 2.9) = 2.9 \\
z_{2_1} &= 1.0 \cdot 2.9 + 0.5 = 3.4 \\
h_{2_1} &= 3.4 \\
y_{\text{pred}_1} &= h_{2_1} = 3.4 \\
L_1 &= \frac{1}{2}\left(y_{\text{pred}_1} - y_{\text{true}_1}\right)^2 = \frac{1}{2}(3.4 - 3.0)^2 = 0.08 \\
\frac{\partial L}{\partial w_2} &= (3.4 - 3.0) \cdot 1 \cdot 2.9 = 1.16 \\
\frac{\partial L}{\partial b_2} &= (3.4 - 3.0) \cdot 1 \cdot 1 = 0.4 \\
\frac{\partial L}{\partial w_1} &= (3.4 - 3.0) \cdot 1 \cdot 1 \cdot 1 \cdot 1.2 = 0.48 \\
\frac{\partial L}{\partial b_1} &= (3.4 - 3.0) \cdot 1 \cdot 1 \cdot 1 \cdot 1 = 0.4 \\
w_1 &= 2.0 - 0.01 \cdot 0.48 = 1.9952 \\
b_1 &= 0.5 - 0.01 \cdot 0.4 = 0.496 \\
w_2 &= 1.0 - 0.01 \cdot 1.16 = 0.9884 \\
b_2 &= 0.5 - 0.01 \cdot 0.4 = 0.496
\end{aligned}
$$

For Data Point 2:

$$
\begin{aligned}
z_1 &= 1.9952 \cdot 0.8 + 0.496 = 2.09 \\
h_1 &= \max(0, 2.09) = 2.09 \\
z_2 &= 0.9884 \cdot 2.09 + 0.496 = 2.56 \\
h_2 &= 2.56 \\
y_{\text{pred}} &= h_2 = 2.56 \\
L_2 &= \frac{1}{2}\left(y_{\text{pred}_1} - y_{\text{true}_1}\right)^2 = \frac{1}{2}(2.56 - 2.5)^2 = 0.0018 \\
\frac{\partial L}{\partial w_2} &= (2.56 - 2.5) \cdot 1 \cdot 2.09 = 0.1254 \\
\frac{\partial L}{\partial b_2} &= (2.56 - 2.5) \cdot 1 \cdot 1 = 0.06 \\
\frac{\partial L}{\partial w_1} &= (2.56 - 2.5) \cdot 1 \cdot 0.9884 \cdot 1 \cdot 0.8 = 0.0474 \\
\frac{\partial L}{\partial b_1} &= (2.56 - 2.5) \cdot 1 \cdot 0.9884 \cdot 1 \cdot 1 = 0.0593 \\
w_1 &= 1.9952 - 0.01 \cdot 0.0474 = 1.9947 \\
b_1 &= 0.496 - 0.01 \cdot 0.0593 = 0.4954 \\
w_2 &= 0.9884 - 0.01 \cdot 0.0018 = 0.9883 \\
b_2 &= 0.496 - 0.01 \cdot 0.06 = 0.4954
\end{aligned}
$$

For Data Point 3:

$$
\begin{aligned}
z_1 &= 1.9947 \cdot 2.0 + 0.4954 = 4.4848 \\
h_1 &= \max(0, 4.4848) = 4.4848 \\
z_2 &= 0.9883 \cdot 4.4848 + 0.4954 = 4.9277 \\
h_2 &= 4.9277 \\
y_{\text{pred}} &= h_2 = 4.9277 \\
L_3 &= \frac{1}{2}\left(y_{\text{pred}_1} - y_{\text{true}_1}\right)^2 = \frac{1}{2}(4.9277 - 4.0)^2 = 0.4303 \\
\frac{\partial L}{\partial w_2} &= (4.9277 - 4.0) \cdot 1 \cdot 4.4842 = 4.16 \\
\frac{\partial L}{\partial b_2} &= (4.9277 - 4.0) \cdot 1 \cdot 1 = 0.9277 \\
\frac{\partial L}{\partial w_1} &= (4.9277 - 4.0) \cdot 1 \cdot 0.9883 \cdot 1 \cdot 2.0 = 1.8633 \\
\frac{\partial L}{\partial b_1} &= (4.9277 - 4.0) \cdot 1 \cdot 0.9883 \cdot 1 \cdot 1 = 0.9168 \\
w_1 &= 1.9947 - 0.01 \cdot 1.8633 = 1.9760 \\
b_1 &= 0.4954 - 0.01 \cdot 0.9168 = 0.4862 \\
w_2 &= 0.9883 - 0.01 \cdot 4.16 = 0.9467 \\
b_2 &= 0.4954 - 0.01 \cdot 0.9277 = 0.4861
\end{aligned}
$$

Final Parameters:

$$
\begin{aligned}
w_1 &= 1.976 \\
w_2 &= 0.9467 \\
b_1 &= 0.4862 \\
b_2 &= 0.4861
\end{aligned}
$$

(b) For an SVM trained with a Gaussian kernel, the kernel function $K(\cdot, \cdot)$ is defined as:

$$K(x_i, x) = exp(-\frac{||x_i - x||^2}{2\sigma^2})$$

where σ is the parameter of the kernel

For an unseen test sample (say x), the SVM would apply the following operation to classify x:

$$f(x) = sign(\sum_{i=1}^{N_s} \alpha_{i,O} y^{(i)} K(x^{(i)}, x) + (1 - \sum_{i=1}^{N_s} \alpha_{i,O} y^{(i)} K(x^{(i)}, x)))$$

where:

N_s is the number of support vectors

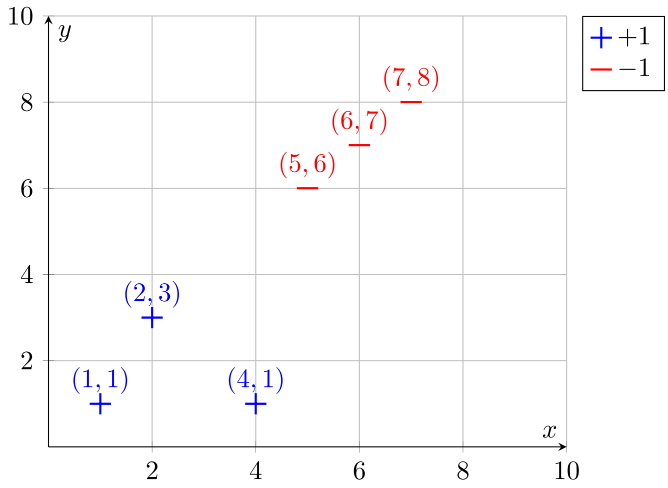
$\alpha_{i,O}$ are the optimal Langrangian multipliers

$y^{(i)}$ is the label for the i^{th}support vector

$x^{(i)}$ is the i^{th}support vector

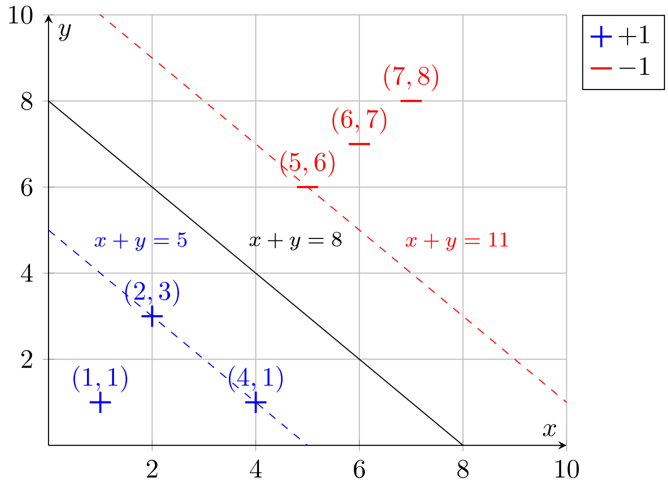
Based on the sign of $\mathbf{w}_O^T\mathbf{x} + \mathrm{b}_O$ calculated as shown above, the point $\mathbf{x}$ is classified.

(ca) **Plotting the Points:**
Consider the label $(+1)$ represented by blue color and label (-1) represented by red color.



Clearly, the two classes are linearly separable.

(cb) **Optimal Decision Boundary:**
Clearly, for an optimal decision boundary, as can be seen from the graph, the support vectors are $(2,3)$,$(4,1)$ (corresponding to Class -1) and $(5,6)$ (corresponding to Class $+1$). The decision boundary must be orthogonal to the line joining the support vectors for each individual class. The line joining $(1,1)$ and $(4,1)$ is $x+y=5$. The optimal decision boundary must be parallel to $x+y=5$ since an SVM decision boundary minimizes the distance of the support vectors from the boundary while maximizing the inter-class distance. This is only possible when the decision boundary is parallel to $x+y=5$.
Since $(5,6)$ is the support vector corresponding to Class -1, a line parallel to $x+y-5=0$ passing through $(5,6)$ would be equidistant from the decision boundary as $x+y-5=0$. Let this line be $x+y-c=0$ (since it is parallel to $x+y-5=0$ the slope remains the same). Since it passes through $(5,6)$, $c=11$.
The optimal decision boundary must, thus, be a line parallel to these 2 lines and at the midpoint of these lines which can be determined as $x+y-\frac{5+11}{2}=0 \Rightarrow x+y-8=0$.



(cc) The support vectors for this decision boundary are, as stated above, $(2,3)$, $(4,1)$ and $(5,6)$.

(cd) The margin for this decision boundary will be the perpendicular distance of any of the support vectors from the line $x+y=8$. Consider the support vector $(5,6)$:

$$l := Ax+By+C=0$$
$$d_{(a,b),l} = \frac{|A\cdot a+B\cdot b+C|}{\sqrt{A^2+B^2}}$$
$$\text{For } l := x+y-8=0$$
$$d_{(5,6)} = \frac{|1\cdot 5+1\cdot 6-8|}{\sqrt{1^2+1^2}} = \frac{3}{\sqrt{2}}$$

(ce) The decision boundary may change if any of the support vectors are removed. The optimal decision boundary is characterized by the support vectors and removing any of them may change the optimal decision boundary in this case:

- If $(5, 6)$ is removed, the decision boundary will shift upwards (towards the point $(6, 7)$) and will remain parallel to the current decision boundary i.e. the equation will be of the form $x+y-c = 0$.
- If $(2, 3)$ is removed, it will be replaced by $(1, 1)$ as the support vector. This will cause the decision boundary to change and the new decision boundary will be parallel to the x-axis.
- If $(4, 1)$ is removed, the support vectors will still be $(2, 3)$ and $(5, 6)$ since they clearly separate the 2 classes. The optimal decision boundary in this case will be the perpendicular bisector of the line joining $(2, 3)$ and $(5, 6)$ which in this case will be $x + y - 8 = 0$. The decision boundary will thus remain unchanged.