

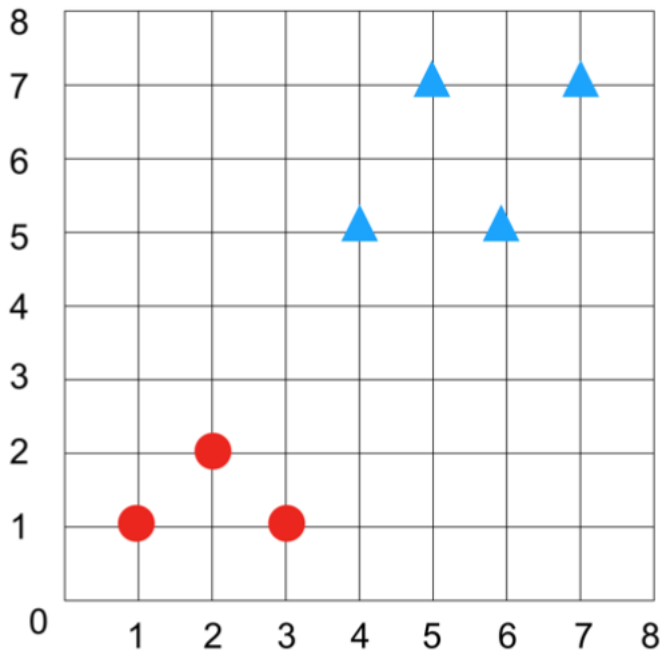
Supervised learning Questions

Question 1. SVM

- A hard-margin support vector machine (SVM), takes n training points $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ with labels $y_1, y_2, \dots, y_n \in \{+1, -1\}$, and finds parameters $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ that satisfy a certain objective function subject to the constraints

$$y_i(X_i \cdot w + \alpha) \geq 1, \forall i \in \{1, \dots, n\}.$$

For parts (a) and (b), consider the following training points. Circles are classified as positive examples with label +1 and triangles are classified as negative examples with label -1.



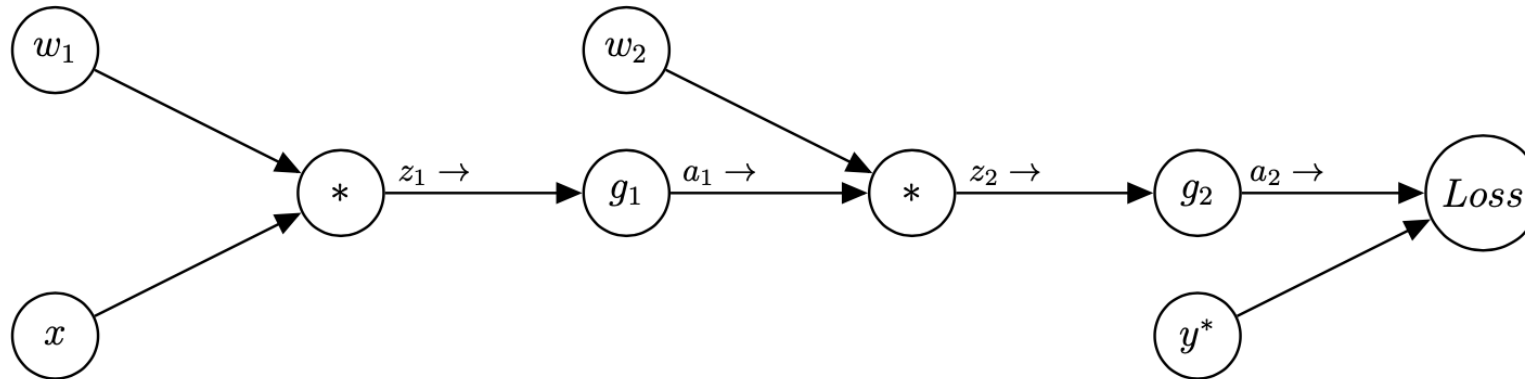
- (a) Which points are the support vectors? Write it as $\begin{bmatrix} \text{horizontal} \\ \text{vertical} \end{bmatrix}$. E.g., the bottom right circle is $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$.
- (b) If we add the sample point $x = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ with label -1 (triangle) to the training set, which points are the support vectors?

For parts (c)–(f), forget about the figure above, but assume that there is at least one sample point in each class and that the sample points are linearly separable.

- (c) Describe the geometric relationship between w and the decision boundary.
- (d) Describe the relationship between w and the margin. (For the purposes of this question, the margin is just a number.)
- (e) Knowing what you know about the hard-margin SVM objective function, explain why for the optimal (w, α) , there must be at least one sample point for which $X_i \cdot w + \alpha = 1$ and one sample point for which $X_i \cdot w + \alpha = -1$.
- (f) If we add new features to the sample points (while retaining all the original features), can the optimal $\|w_{new}\|$ in the enlarged SVM be greater than the optimal $\|w_{old}\|$ in the original SVM? Can it be smaller? Can it be the same? Explain why! (Most of the points will be for your explanation.)

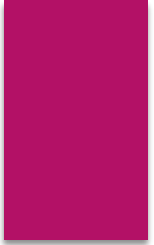
Question 2: NN

- Consider the following computation graph for a simple neural network for binary classification. Here x is a single real-valued input feature with an associated class y^* (0 or 1). There are two weight parameters w_1 and w_2 , and non-linearity functions g_1 and g_2 (to be defined later, below). The network will output a value a_2 between 0 and 1, representing the probability of being in class 1. We will be using a loss function $Loss$ (to be defined later, below), to compare the prediction a_2 with the true class y^* .



- Perform the forward pass on this network, writing the output values for each node z_1, a_1, z_2 and a_2 in terms of the node's input values.
- Compute the loss $Loss(a_2, y^*)$ in terms of the input x , weights w_i , and activation functions g_i .
- Now we will work through parts of the backward pass, incrementally. Use the chain rule to derive $\frac{\partial Loss}{\partial w_2}$.

Write your expression as a product of partial derivatives at each node: i.e. the partial derivative of the node's output with respect to its inputs. (Hint: the series of expressions you wrote in part (a) will be helpful; you may use any of those variables.)

- 
- (d) Suppose the loss function is quadratic, $Loss(a_2, y^*) = \frac{1}{2}(a_2 - y^*)^2$, and g_1 and g_2 are both sigmoid functions $g(z) = \frac{1}{1+e^{-z}}$ (note: it's typically better to use a different type of loss, cross-entropy, for classification problems, but we'll use this to make the math easier). Using the chain rule from Part (c), and the fact that $\frac{\partial g(z)}{\partial z} = g(z)(1 - g(z))$ for the sigmoid function, write $\frac{\partial Loss}{\partial w_2}$ in terms of the values from the forward pass, y^*, a_1 , and a_2 .
- (e) Now use the chain rule to derive $\frac{\partial Loss}{\partial w_1}$ as a product of partial derivatives at each node used in the chain rule.
- (f) Finally, write $\frac{\partial Loss}{\partial w_1}$ in terms of x, y^*, w_i, a_i, z_i .
- (g) What is the gradient descent update for w_1 with step-size α in terms of the values computed above?

Question 3. DT

► Here is a table which records some data about whether a student will go out to play. Use decision tree to analysis the following questions:

(a) Using entropy to analysis which attribute should be root node among outlook, temperature, humidity and windy? Write your analysis process in question .

(b) Draw the decision tree.

| Outlook | Temperature | Humidity | Windy | Play? |
|----------|-------------|----------|-------|-------|
| sunny | hot | high | false | No |
| sunny | hot | high | true | No |
| overcast | hot | high | false | Yes |
| rain | mild | high | false | Yes |
| rain | cool | normal | false | Yes |
| rain | cool | normal | true | No |
| overcast | cool | normal | true | Yes |
| sunny | mild | high | false | No |
| sunny | cool | normal | false | Yes |
| rain | mild | normal | false | Yes |
| sunny | mild | normal | true | Yes |
| overcast | mild | high | true | Yes |
| overcast | hot | normal | false | Yes |
| rain | mild | high | true | No |