50.007 Machine Learning

Simriti Bundhoo 1006281

Homework 1

1

(a) From the description given, assume that the positive data points should lie inside the circle while the negative data points lie outside the circle. (i.e., the distance from the centre (0,0) to the point should less or equal to the radius r of the circle for that data point to be classified as positive).

Thus, the decision boundary of the data is:

$$x : ||x|| = r$$

where x is an input data and r is the radius of the circle.

The classifier $h(x)$ is assumed to be:

$$h(x) = \begin{cases} +1 & \text{if } ||x|| \leq r \\ -1 & \text{otherwise} \end{cases}$$

Positive examples:

- (1, 1) is labeled as +1 because it belongs to the positive class.
- (2, 2) is labeled as +1 because it belongs to the positive class.

Negative examples:

- (-1, 1) is labeled as -1 because it belongs to the negative class.
- (1, -1) is labeled as -1 because it belongs to the negative class.

Given the positive examples (1,1) and (2,2), their magnitudes are:

$$||(1,1)|| = \sqrt{(1)^2 + (1)^2} = \sqrt{2}$$

$$||(2,2)|| = \sqrt{(2)^2 + (2)^2} = 2\sqrt{2}$$

Given the negative examples (-1,1) and (1,-1), their magnitudes are both $\sqrt{2}$.

$$||(-1,1)|| = \sqrt{(-1)^2 + (1)^2} = \sqrt{2}$$

$$||(1,1)|| = \sqrt{(1)^2 + (1)^2} = \sqrt{2}$$

Since the distances for (1,1), (-1,1) and (1, -1) are still the same despite having different signs, the classifier is deemed unable to classify the data points in the given dataset using an origin-centred circle hypothesis correctly. The decision boundary based solely on the distance from the origin cannot differentiate between the positive and negative examples in this case, resulting in misclassifications.

(b) In a two-dimensional space, a line passing through the origin (0, 0) can be represented by the equation:

$$y = mx$$

To consider the equation in terms of θ, consider x_1 as the x-coordinate and x_2 as the y-coordinate . The whole equation can be re-written as:

$$\theta_1 * x_1 + \theta_2 * x_2 = 0$$

where θ_1 and θ_2 are the parameters representing the coefficients of x_1 and x_2, respectively. Let $x = (x_1, x_2)$ and $\theta = (\theta_1, \theta_2)$. Therefore, the decision boundary is:

$$x : \theta.x = 0$$

The decision boundary $h(x)$ is:

$$h(x) = \begin{cases} +1 & \text{if } \theta.x > 0 \\ -1 & \text{otherwise} \end{cases}$$

From the $h(x)$ equation, a data point which dot product between x and θ is greater than 0 will be classified as positive. Otherwise, it will be classified as negative.

Recalling that the examples given and their labels:

- (1, 1) is labeled as +1 because it belongs to the positive class.
- (2, 2) is labeled as +1 because it belongs to the positive class.
- (-1, 1) is labeled as -1 because it belongs to the negative class.
- (1, -1) is labeled as -1 because it belongs to the negative class.

Assuming the θ values to be +1, the positive examples (1,1) and (2,2) each give a dot product of:

$$(1 \times 1) + (1 \times 1) = 2$$

$$(2 \times 2) + (2 \times 2) = 8$$

Both dot products are greater than 0. Thus these two data points are classified as positive.

The negative examples (-1,1) and (1,-1) each give a dot product of:

$$(1 \times -1) + (1 \times 1) = 0$$

$$(1 \times 1) + (1 \times -1) = 0$$

Both dot products are equal to 0. Thus, they are classified as negative.

As the positive examples are correctly classified as positive and the negative examples are correctly classified as negative, it can be concluded that, for the given data set, the data points are correctly classified using the hypothesis space of a line through the origin with a normal vector $\theta = (\theta_1, \theta_2)$.

2

Refer to the attached .ipynb file.

3

Refer to the attached .ipynb file.

4

Refer to the attached .ipynb file.