COMP3314 Homework 1 Wang Qifan 3035973452

1 Perceptron

The net input in Perceptron is

$$z = w_0x_0 + w_1x_1 + \ldots + w_mx_m$$

where $w_0 = -1.5, x_0 = 1$.

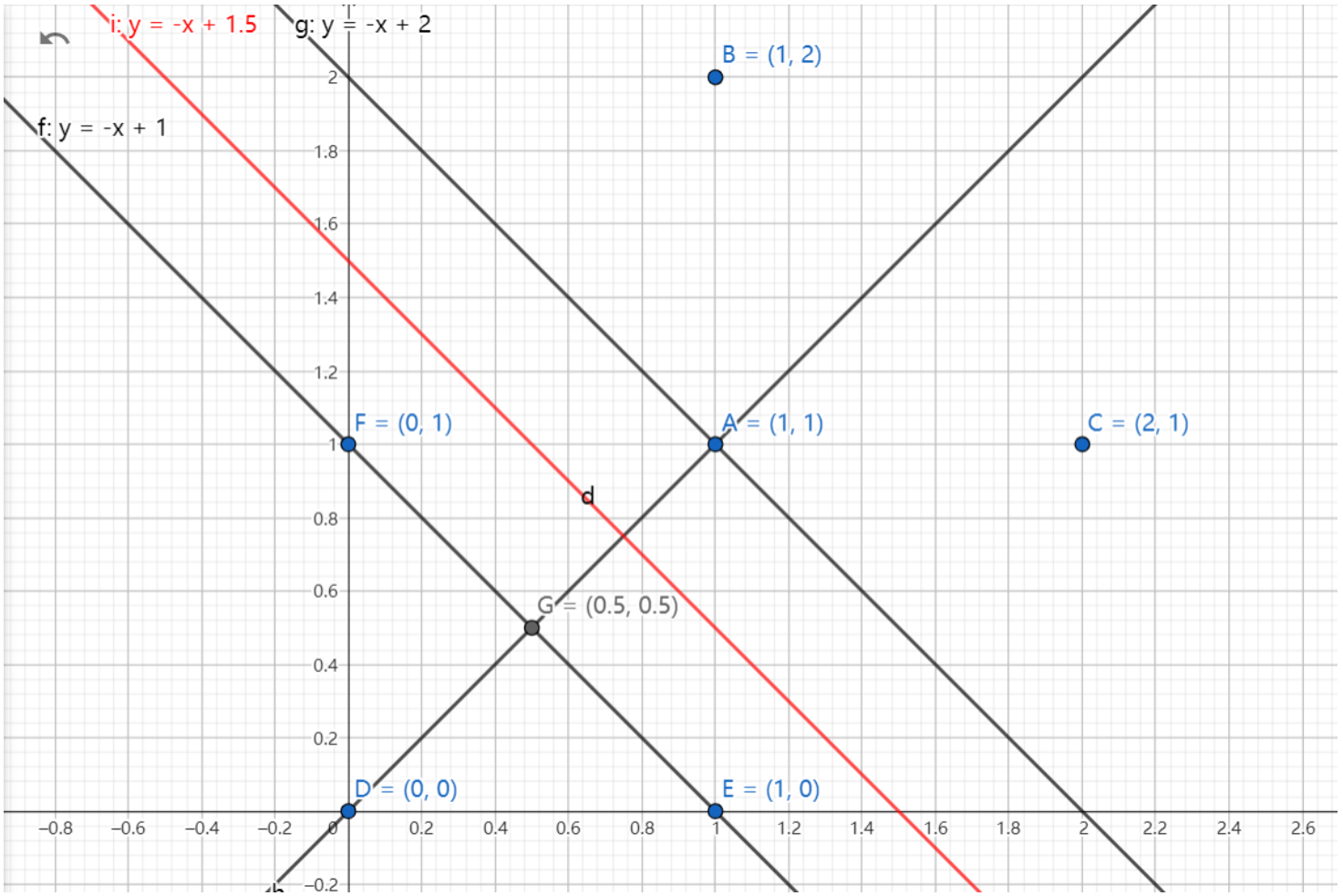
The output in Perceptron is

$$\phi(z) = \begin{cases} 1 & if\ z \geq 0 \\ -1 & otherwise \end{cases}$$

- For inputs (0,0): $z = -1.5 + 0 + 0 = -1.5 < 0$. Therefore, the **output** $\phi(z) = -1$
- For inputs (1,0): $z = -1.5 + 1 \times 1 + 0 = -0.5 < 0$. Therefore, the **output** $\phi(z) = -1$
- For inputs (0,1): $z = -1.5 + 0 + 1 \times 2 = 0.5 \geq 0$. Therefore, the **output** $\phi(z) = 1$
- For inputs (1,1): $z = -1.5 + 1 \times 1 + 1 \times 2 = -1.5 < 0$. Therefore, the **output** $\phi(z) = 1$

2 Support Vector

(1)



First, we plot the six points above.

Suppose the optimal separating line is $w_1x_1 + w_2x_2 + b = 0$. For each point from the support vector machine, we should maximize the distance between the support vectors and the separating line.

The distance from a point to a line is $d = \frac{|wx+b|}{||w||}$, and each point also should satisfy $y^{(i)}(w^T x^{(i)} + b) \geq 1$. Therefore,

- For D(0,0): $-1 \times (0 + 0 + b) = -b \geq 1$ (1)
- For F(0,1): $-1 \times (0 + 1 \times w_2 + b) = -w_2 - b \geq 1$ (2)
- For E(1,0): $-1 \times (0 + 1 \times w_1 + b) = -w_1 - b \geq 1$ (3)
- For A(1,1): $1 \times (1 \times w_1 + 1 \times w_2 + b) = w_1 + w_2 + b \geq 1$ (4)
- For B(1,2): $1 \times (1 \times w_1 + 2 \times w_2 + b) = w_1 + 2w_2 + b \geq 1$ (5)
- For C(2,1): $1 \times (2 \times w_1 + 1 \times w_2 + b) = 2w_1 + w_2 + b \geq 1$ (6)

Since the support vectors must have $d = \frac{|wx+b|}{||w||} = \frac{1}{||w||}$, we need to **maximize** this value, which means we need to **minimize** $|w|^2 = w_1^2 + w_2^2$.

From the **Lagrange Multiplier**, we have $w_1 = w_2 = 1, b = -1.5$ finally. Thus, **the optimal separating line** is $y = -x + 1.5$, which is the red line in the picture.

(2)

The support vector should satisfy $y^{(i)}(w^T x^{(i)} + b) = 1$, therefore, **A(1,1), E(1,0), F(0,1)** are the support vectors.

Support vectors are data points that lie closest to the decision surface (or hyperplane). They are critical elements of the training set because the orientation and position of the decision hyperplane are completely determined by these points. If these points were to be removed or changed, the position of the decision boundary would also change. In other words, they "support" the hyperplane, hence the term "support vectors". The `SVM` optimization problem ensures that the margin, or distance between the decision boundary and the closest points from each class, is maximized. These closest points are the support vectors.

3 Entropy

We will use the following formula to calculate the entropy for the given probabilities of the events.

$$H = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

Since we know that P(first) = 0.5, the other four probabilities are all 0.125. We have

$$H = -0.5 \times log_2 0.5 - 4 \times 0.125 \times log_2 0.125 = 0.5 + 1.5 = 2$$

Therefore, the entropy is 2.

The entropy measures the uncertainty or randomness of the information. The higher the information entropy, the greater the uncertainty of the information. The lower the information entropy, the smaller the uncertainty of the information.

The higher the information entropy, the harder it is for us to predict the outcome of the next event. Conversely, if the probability of some outcomes occurring in an event is high and the probability of others occurring is low, the information entropy will be low. This means that we can predict the outcome of events relatively easily, because some outcomes are more likely to occur than others.

In our problem, the probability of the first event occurring is 0.5, which is much higher than other events, which reduces the uncertainty of the system. Therefore, although there are five possibilities, the information entropy is only 2 bits, indicating that the overall uncertainty is not maximum. An entropy of 2.0 bits means that, on average, we would need 2 bits of information to describe the outcome of one event from this set.

4 Logistic Regression

(1)

The logistic regression prediction function is given by:

$$p = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2)}}$$

Thus, we put $x_1 = 40, x_2 = 3.5, w_1 = 0.05, w_2 = 1, w_0 = -6$ in the above function

$$p = \frac{1}{1+e^{-(-6+0.05\times 40+1\times 3.5)}} = \frac{1}{1+e^{0.5}} = 37.75\%$$

Therefore, the probability for the student to get A is **37.75%**.

(2)

This time, we have $p = 0.5, w_0, w_1, w_2, x_2$ are remain the same as in (1). Thus,

$$0.5 = \frac{1}{1 + e^{-(-6+0.05\times x_1+1\times 3.5)}} = \frac{1}{1 + e^{2.5-0.05x_1}}$$

which equals to

$$e^{2.5-0.05x_1} = 1 \Rightarrow 0.05x_1 = 2.5 \Rightarrow x_1 = 50$$

Therefore, the student needs to study **50 hours** for this course so that he can get an A.

5 Decision Tree

(1)

We will use the following formula to calculate the `Gini impurity` and the `Information Gain`

$$IG(D_P) = I_G(D_P) - \sum_{j=1}^{m} \frac{N_j}{N_p} I_G(D_j)$$

$$I_G(t) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

The Gini impurity is initially

$$I_G(0) = 1 - 0,6 \times 0.6 - 0.4 \times 0.4 = 0.48 \qquad (\text{N=5})$$

- The Gini impurity **after V** is separated is

$$I_G(1)_1 = 1 - 0.5 \times 0.5 - 0.5 \times 0.5 = 0.5 \quad \text{(N=2)}$$

$$I_G(1)_2 = 1 - \frac{1}{3} \times \frac{1}{3} - \frac{2}{3} \times \frac{2}{3} = \frac{4}{9} \quad \text{(N=3)}$$

Therefore,

$$IG(Y|V) = I_G(0) - \frac{2}{5} I_G(1)_1 - \frac{3}{5} I_G(1)_2 = 0.48 - 0.2 - \frac{4}{15} = 0.013$$

- the Gini impurity **after W** is separated is

$$I_G(1)_1 = 1 - 0.5 \times 0.5 - 0.5 \times 0.5 = 0.5 \quad \text{(N=2)}$$

$$I_G(1)_2 = 1 - \frac{1}{3} \times \frac{1}{3} - \frac{2}{3} \times \frac{2}{3} = \frac{4}{9} \quad \text{(N=3)}$$

Therefore,

$$IG(Y|W) = I_G(0) - \frac{2}{5} I_G(1)_1 - \frac{3}{5} I_G(1)_2 = 0.48 - 0.2 - \frac{4}{15} = 0.013$$

- The Gini impurity **after X** is separated is

$$I_G(1)_1 = 1 - 0.5 \times 0.5 - 0.5 \times 0.5 = 0.5 \quad \text{(N=4)}$$

$$I_G(1)_2 = 1 - 1 \times 1 = 0 \quad \text{(N=1)}$$

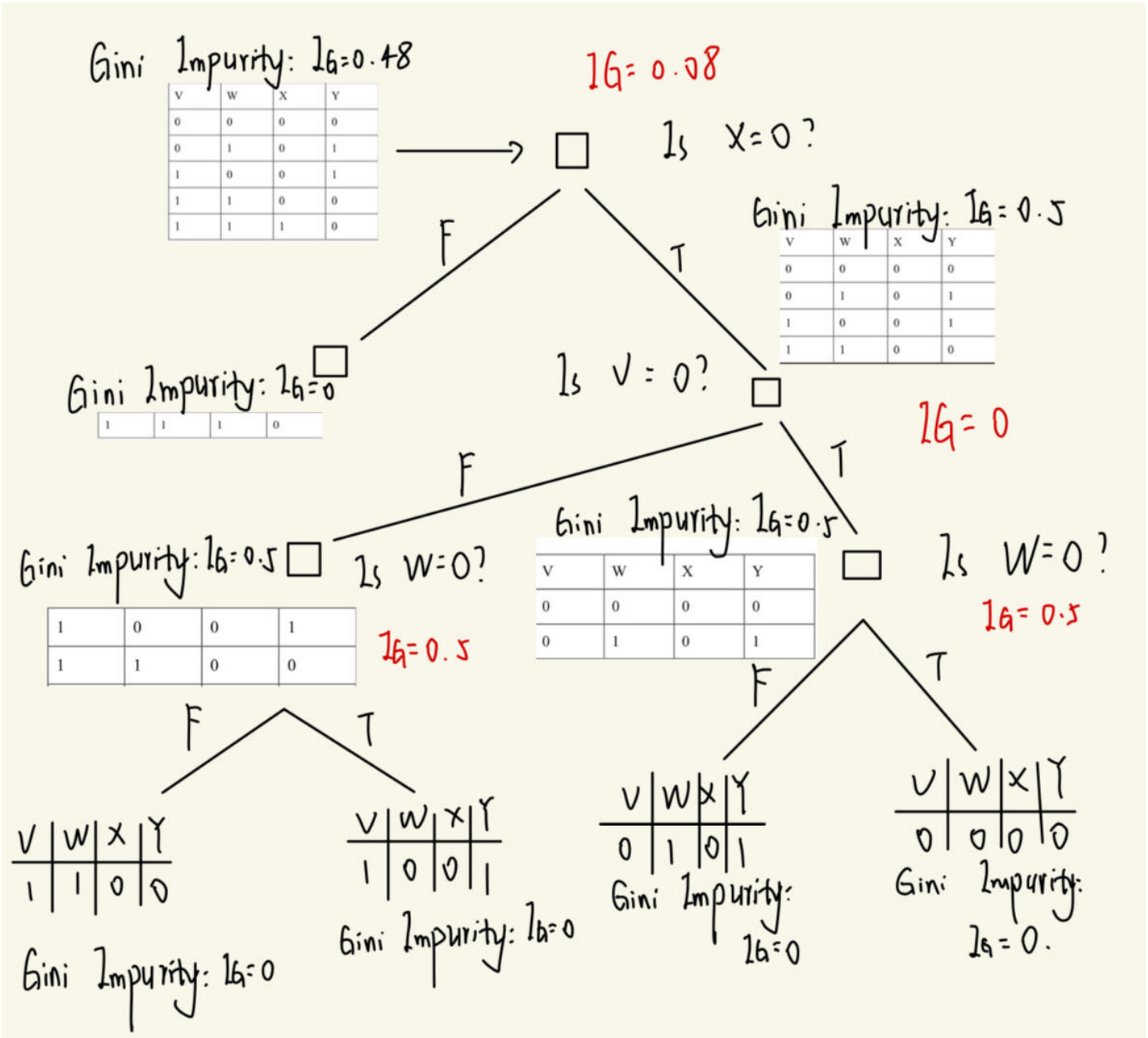
Therefore,

$$IG(Y|W) = I_G(0) - \frac{4}{5} I_G(1)_1 - \frac{1}{5} I_G(1)_2 = 0.48 - 0.4 - 0 = 0.08$$

Above all, **IG(Y|V)=IG(Y|W)=0.013, IG(Y|X)=0.08**.

(2)

Below is the entire decision tree.


Gini Impurity: IG=0.48
V W X Y
0 0 0 0
0 1 0 1
1 0 0 1
1 1 0 0
1 1 1 0
IG= 0.08
Is X=0?
F
T
Gini Impurity: IG=0
1 1 1 0
Gini Impurity: IG=0.5
V W X Y
0 0 0 0
0 1 0 1
1 0 0 1
1 1 0 0
Is V=0?
IG=0
F
T
Gini Impurity: IG=0.5
Is W=0?
1 0 0 1
1 1 0 0
IG=0.5
Gini Impurity: IG=0.5
V W X Y
0 0 0 0
0 1 0 1
Is W=0?
IG=0.5
F
T
V W X Y
1 1 0 0
Gini Impurity: IG=0
V W X Y
1 0 0 1
Gini Impurity: IG=0
V W X Y
0 1 0 1
Gini Impurity:
IG=0
V W X Y
0 0 0 0
Gini Impurity:
IG=0.