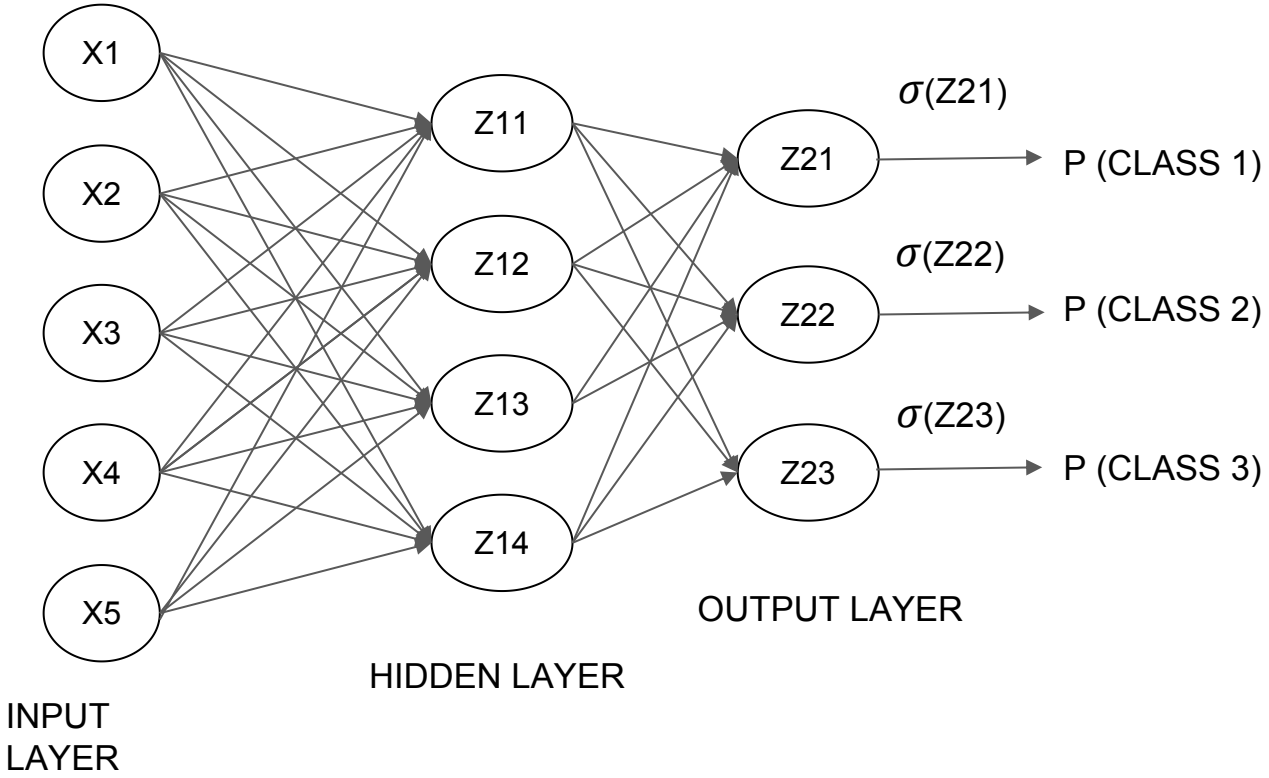
Softmax Activation function

Solved example - Multi class classification

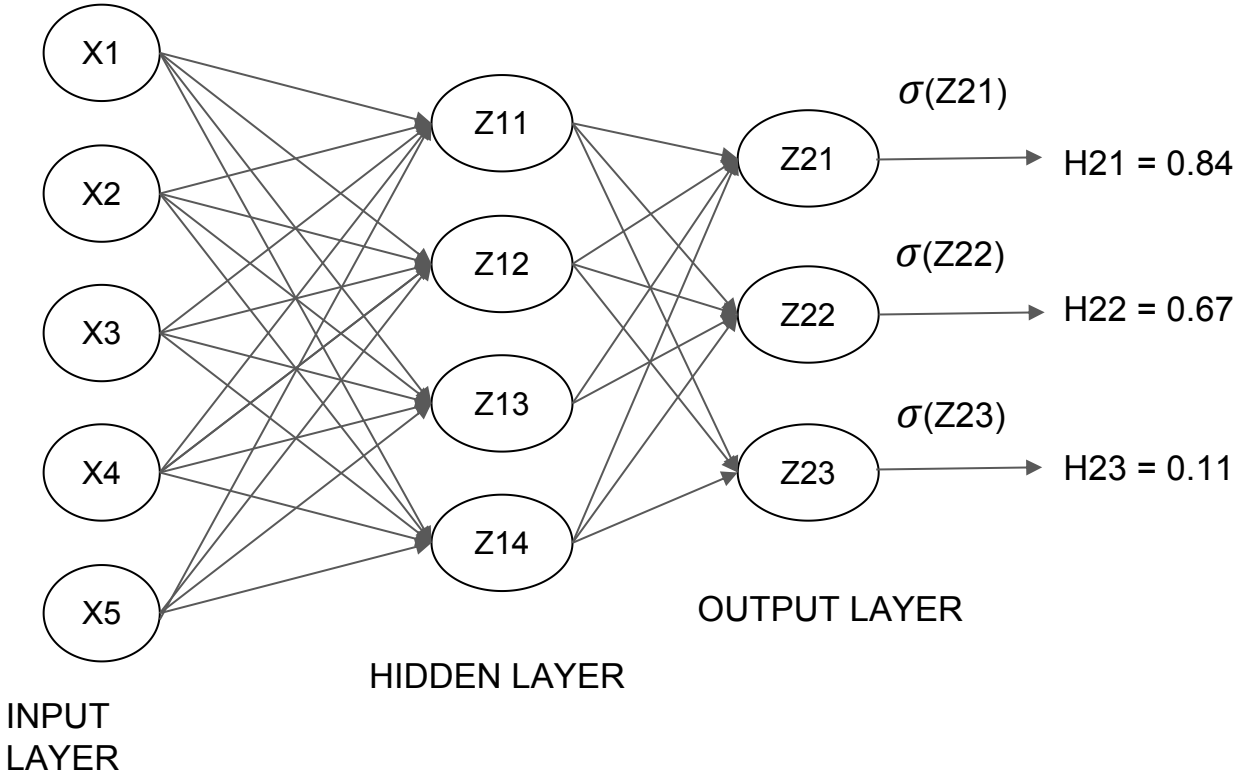
Feature X1	Feature X2	Feature X3	Feature X4	Feature X5	Target
1	22	569	35	0	Class 1
1	7	351	75	1	Class 2
1	45	451	542	1	Class 2
1	5	572	8	0	Class 1
0	22	565	44	1	Class 3
0	24	243	546	1	Class 3
1	78	953	42	0	Class 2

Simple ANN for 3-class classification



Z21 will give the probability for the data sample belongs to Class 1.

Calculate the output of the network



1. Suppose we compute the values Z21, Z22, Z23 using the weights and biases of the output layer and apply the sigmoid activation function to get the outputs H21, H22 and H23.
2. As sigmoid will give the output as a value between 0 and 1.
3. Suppose the results are as shown here.

Interpretations

1. **Problem 1**
    a. The probability for the data sample to belong to Class 1 is 0.84
    b. The probability for the data sample to belong to Class 2 is 0.67
    c. So if we apply a threshold 0.5, the network will say that the data sample belongs to both Class 1 and Class 2.
2. **Problem 2**
    a. The probability values are independent of each other.
    b. That means the probability that the data point belongs to class 1 does not take into account the probability of the other two classes.
3. Hence, s**igmoid activation function is not preferred** in the last layer **for a multi-class classification problem**.

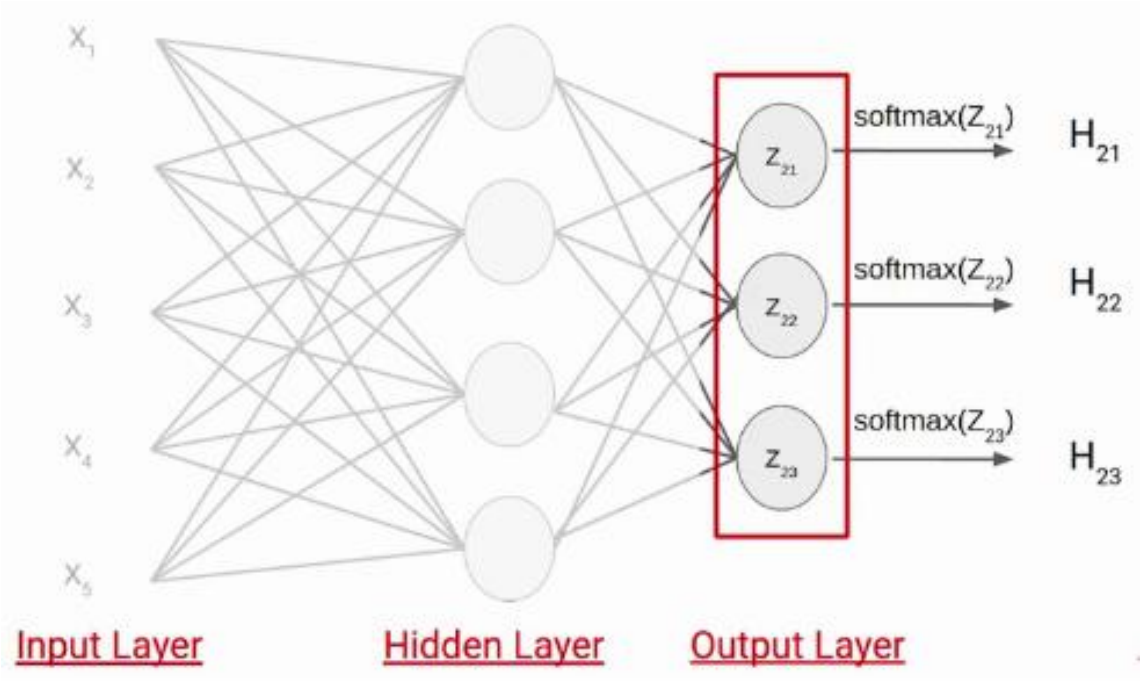
Softmax activation function

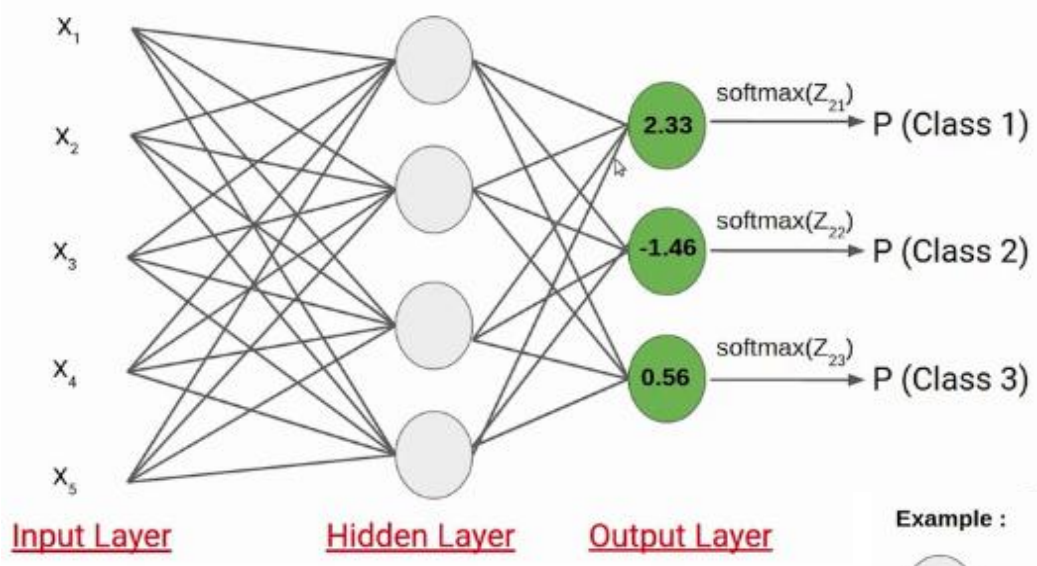
- The **Softmax activation function calculates the relative probabilities**.
- That means it uses the value of Z21, Z22, Z23 to determine the final probability value.

$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

- Here, the *z* represents the values from the neurons of the output layer. The exponential acts as the non-linear function. Later these values are divided by the sum of exponential values in order to normalize and then convert them into probabilities.
- Note that, when the number of classes is two, it becomes the same as the sigmoid activation function. In other words, sigmoid is simply a variant of the Softmax function.
- Any multi-class classification problem can be reduced to multiple binary classification problems using "one-vs-all" method, i.e. having C sigmoids (when C is the number of classes) and interpreting every sigmoid to be the probability of being in that specific class or not, and taking the max probability.

Example Network





Suppose the value of Z21, Z22, Z23 comes out to be 2.33, -1.46, and 0.56 respectively.

Now the SoftMax activation function is applied to each of these neurons and the following values are generated.

- These are the probability values that a data point belonging to the respective classes.
- The sum of the probabilities, in this case, is equal to 1.
- It is clear that the input belongs to class 1.
- So if the probability of any of these classes is changed, the probability value of the first class would also change.

Example :

2.33 → $P(\text{Class 1}) = \frac{\exp(2.33)}{\exp(2.33) + \exp(-1.46) + \exp(0.56)} = 0.83827314$

-1.46 → $P(\text{Class 2}) = \frac{\exp(-1.46)}{\exp(2.33) + \exp(-1.46) + \exp(0.56)} = 0.01894129$

0.56 → $P(\text{Class 3}) = \frac{\exp(0.56)}{\exp(2.33) + \exp(-1.46) + \exp(0.56)} = 0.14278557$

Summary

- Softmax allows CNNs to output a probability distribution over the possible classes. This is important because it allows the CNN to make more accurate predictions.
- Softmax works by first normalizing the input vector so that all of the numbers in the vector sum to 1. Then, it exponentiates each number in the vector and divides by the sum of all of the exponentiated numbers. This results in a vector of probabilities, where each probability is between 0 and 1 and represents the probability that the input belongs to a particular class.
- The probability distribution output by the softmax function can then be used to make a more accurate prediction about the class of an input image. For example, if the CNN is predicting whether an image contains a cat or a dog, the probability distribution can indicate how likely it is that the image contains a cat and how likely it is that the image contains a dog.
- Softmax is typically used in the last layer of a neural network to predict the class of an input image. It is also used in other applications, such as natural language processing and machine translation.