Deep Learning: Which Loss and Activation Functions should I use?

Credits:Stacey Ronaghan

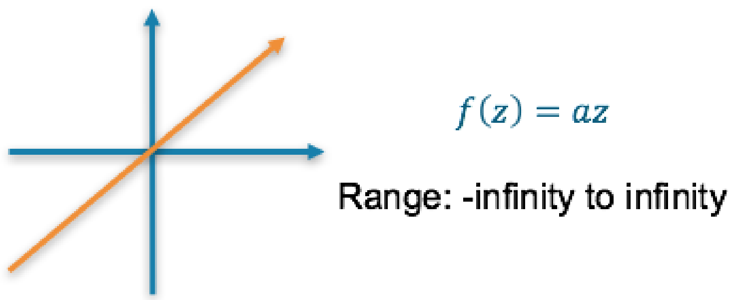
Regression: Predicting a numerical value

The **final layer of the neural network** will have **one neuron and the value it returns is a continuous numerical value.**

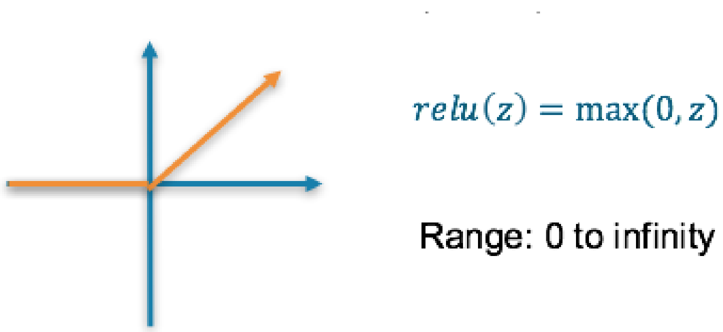
To **understand the accuracy of the prediction**, it is **compared with the true value which is also a continuous number.**

Final Activation Function

Linear — This results in a **numerical value which we require**



ReLU — This results in a **numerical value greater than 0**



Loss Function

Mean squared error (MSE) — This **finds the average squared difference between** the **predicted value and the true value**

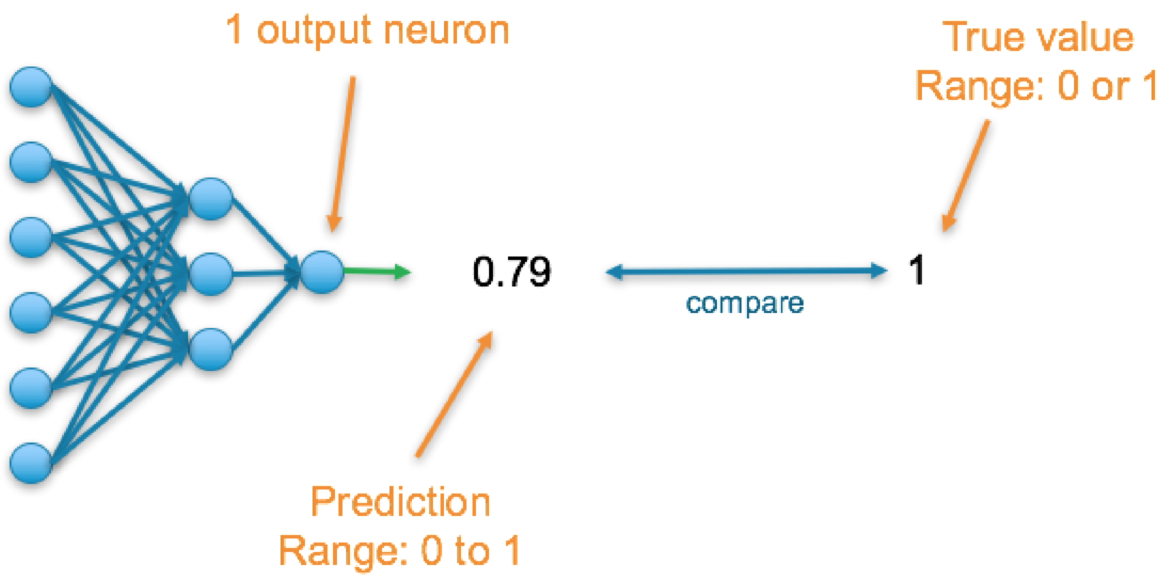
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where $\hat{y}$ is the predicted value and y is the true value

Categorical: Predicting a binary outcome

The **final layer of the neural network will have one neuron** and **will return a value between 0 and 1**, which can be inferred as a probably.

To **understand the accuracy of the prediction**, it is **compared with the true value. If the data is that class, the true value is a 1, else it is a 0.**



Final Activation Function

Sigmoid — **This results in a value between 0 and 1** which we can infer to be how confident the model is of the example being in the class

Loss Function

Binary Cross Entropy — **Cross entropy quantifies the difference between two probability distribution.**

Our **model predicts a model distribution of {p, 1-p}** as we have a binary distribution. We use binary cross-entropy to compare this with the true distribution {y, 1-y}

Categorical: Predicting a single label from multiple classes

The **final layer of the neural network will have one neuron** for **each of the classes** and **they will return a value between 0 and 1**, which can be inferred as a probably. **The output then results in a probability distribution as it sums to 1.**

To **understand the accuracy of the prediction**, each **output is compared with its corresponding true value. True values have been one-hot-encoded meaning a 1 appears in the column corresponding to the correct category**, else a 0 appears

Final Activation Function

Softmax — **This results in values between 0 and 1** for **each of the outputs which all sum up to 1.** Consequently, this can be inferred as a probability distribution

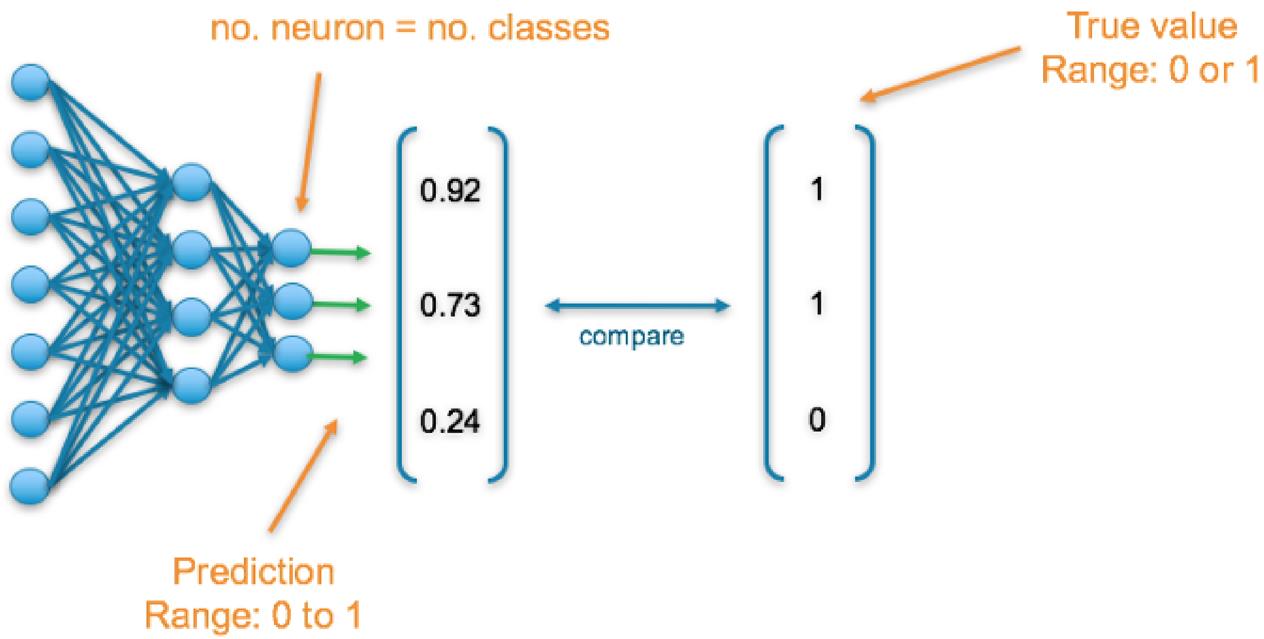
Loss Function

Cross Entropy — Cross entropy quantifies the difference between two probability distribution. Our model predicts a model distribution of {p1, p2, p3} (where p1+p2+p3 = 1). We use cross-entropy to compare this with the true distribution {y1, y2, y3}

Categorical: Predicting multiple labels from multiple classes

The final layer of the neural network will have one neuron for each of the classes and they will return a value between 0 and 1, which can be inferred as a probably.

To understand the accuracy of the prediction, each output is compared with its corresponding true value. If 1 appears in the true value column, the category it corresponds to is present in the data, else a 0 appears.



Final Activation Function

Sigmoid — This **results in a value between 0 and 1 which we can infer to be how confident it is of it being in the class**

Loss Function

Binary Cross Entropy — Cross entropy quantifies the difference between two probability distribution. Our model predicts a model distribution of {p, 1-p} (binary distribution) for each of the classes. We use binary cross-entropy to compare these with the true distributions {y, 1-y} for each class and sum up their results