Based on all the experiments result of the multiclass image classifications and binary image classifications, the cross entropy loss function is the best loss function in the classification task. Using cross entropy loss function, the overall accuracy will guarantee to be the highest among other built-in loss functions.

**Cross Entropy**

Now, I will discuss the power of cross entropy on classification task in theoretical view. Imagine there are two models that make prediction to four class of images, (cats, dogs, cows and wolfs). In the first model, the prediction given will be:

|  |  |  |
| --- | --- | --- |
| Predicted label | Correctness | True label |
| [0.2, 0.1, 0.2, 0.5] | Yes | [0, 0, 0, 1] cats |
| [0.3, 0.2, 0.4, 0.1] | Yes | [0, 0, 1, 0] dogs |
| [0.2, 0.3, 0.4, 0.1] | No | [0, 1, 0, 0] cows |
| [0.2, 0.5, 0.2, 0.1] | No | [1, 0, 0, 0] wolfs |

In second model, the prediction given will be:

|  |  |  |
| --- | --- | --- |
| Predicted label | Correctness | True label |
| [0.2, 0.1, 0.2, 0.5] | Yes | [0, 0, 0, 1] cats |
| [0.3, 0.2, 0.4, 0.1] | Yes | [0, 0, 1, 0] dogs |
| [0.2, 0.3, 0.4, 0.1] | No | [0, 1, 0, 0] cows |
| [0.4, 0.3, 0.2, 0.1] | Yes | [1, 0, 0, 0] wolfs |

Based on the formula of cross entropy, the loss will be evaluated by the formula of logarithm:

In the formula, yi is either in value 1/0, 1 if the sample i matched with the true label, vice versa.  
For the binary classification problem, the formula is similar. But the value of p changed in the formula, as there is only two classes of image. So, there will only be p or 1-p for the probability outputted:

The loss of model 1:

The loss of model 2:

Through the cross entropy, we can identify which model can output better prediction on the classification task. However, in the case of other loss functions, like MSE, the loss function can also evaluate which models is better based on its formula:

MSE loss of model 1:

MSE loss of model 2:

The advantages of using cross entropy on machine learning is not about how well the function can evaluate the models. Machine learning is an optimization problem, it involves the gradient descent algorithm during the training process. With cross entropy, gradient descent algorithm works better, the model can learn more efficiently: In the process of generating the final prediction, the last layer of model will receive the logits/scores of each class of images. The logits/scores will then pass to the sigmoid/ softmax activation function to output probability. Finally, the model will make use of the cross entropy function to output the prediction in one-hot encoded format:

The problem of other loss function is that after passing the sigmoid/softmax activation function, the output becomes non-convex. For example, MSE after the softmax activation will become a non-convex curve, meaning that the gradient descent algorithm can hardly find the global minimum. However, CE is a different scenario, gradient descent algorithm works well on finding its global minimum. Therefore, CE will be the best