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.

**Why 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 CE loss, the model 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:

Based on the derived result, 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, the model training involves the gradient descent algorithm in every epoch. 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:

Picture

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, it may easily stuck at the local minima using non-CE loss function. In other loss functions, we can observe that it takes longer time for training in the earlier epochs owing to the non-convex nature of the loss generated. However, CE is in a different scenario, gradient descent algorithm works well on it to find the global minimum in the curve. The curve after the softmax /sigmoid activation is a convex function. Therefore, CE is the best loss function on the classification task.

Picture

In the gradient descend algorithm, the general goal is that with the larger the loss of one epoch, the gradient should descent in a faster rate. In MSE, the gradient will not have such effect on the sigmoid activation function. In a larger value of the MSE loss, the gradient is relatively flat, so the rate of gradient descent is slow, vice versa. This is the typical problem of non-CE loss functions: greater error, learn faster in every epochs.   
Picture  
  
 However, in CE, the above constraints can be solved with its formula structure, deriving the gradient descent algorithm:

Derve

In conclusion, CE have the advantages on the gradient descent algorithm in the classification task owing to its formula structure. CE focuses on the difference between true label and the predicted label. In the formula structure, we can observe that the first coefficient yi is in 0/1 value, the formula dose not put emphasis on the incorrect class value, it focuses on the correctness of the class. This would not be a problem for two class of images. However, if the number of classes increases, the accuracy may drop as the inter class information is not focused by CE.

Experiment