Loss functions in machine Learning

Mainly, loss functions can be classified into two major categories considering the training task, that are:

1. Regression losses

There are various factors involved in choosing a loss function for specific problem such as type of machine learning algorithm chosen, ease of calculating the derivatives and to some degree the percentage of outliers in the data set.

1.1 Mean Square Error Loss

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

The Mean square error is measured as the average of squared difference between predictions and actual observations.

1.2 Mean Absolute Error

Mean absolute error is measured as the average of sum of absolute differences between predictions and actual observations

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

2. Classification Losses

2.1 Multi class SVM Loss

The score of the correct category should be greater than sum of scores of all incorrect categories by some safety margin (usually one). And hence this loss is used for maximum-margin classification, most notably for support vector machines. Although not differentiable, it's a convex function which makes it easy to work with usual convex optimizers used in machine learning domain.

$$SVM\ Loss = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

Example: Consider an example where we have three training examples and three classes to predict: 1) Dog, 2) cat; and 3) Horse. Below the values predicted by our algorithm for each of the classes [2]:



	Image #1	Image #2	Image #3
Dog	-0.39	-4.61	1.03
Cat	1.49	3.28	-2.37
Horse	4.21	1.46	-2.27

* This table and pictures are brough from [2]

2.2 Negative Log Likelihood Function (Cross Entropy)

The value of this loss function increases as the predicted probability diverges from the actual label.

$$NLL(Cross\ Entropy\ Loss) = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

Notice that when actual label is 1 ($y(i) = 1$), second half of function disappears whereas in case actual label is 0 ($y(i) = 0$) first half is dropped off. In short, we are just multiplying the log of the actual predicted probability for the ground truth class.

References:
1] Bishop, Christopher M., and Nasser M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4, no. 4. New York: springer, 2006.
2] https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23, last visited Feb 25, 2022.