**HW1 – Question 1 & 2**

ID: eo9232

Name: Md Reza

CSC 5825 – Fall 2021

**Question 1.** (10 points) Describe one problem in your study area that can be solved by

machine learning techniques. Classify your problem in terms of supervised or unsupervised, classification or regression? Explain the unique challenges to standard machine

learning methods?

**Use Case:** Develop a metadata framework with AI/ML model to automate data tagging/classification processes that would:

* Standardized enterprise complex data
* Minimize intensive manual efforts/resources
* maximize data discoverability
* maximize governance of enterprise data

**Problem Classification:**This is a classification problem with the supervised learning category. The objective was to build a machine learning model that includes:

* Prepare the training dataset
* Choose the best model
* Train the model
* Perform cross-validation
* Evaluate the model
* Tune the parameter
* Put the model in production to predict and classify/annotate the newly ingested data.

**Challenges:**

* Poor quality of Data.
* Manually annotating/tagging the training data (Thousands of rows).
* Selecting the best model since the initial accuracy derived from leveraging different models was so close.
* Train the models with more than a hundred tags/classifiers.
* Models’ complexity and bias & variance trade-offs.
* Issues with Overfitting and & Underfitting.
* It was tricky and challenging to integrate the confusion matrix into the model.

**Question 2.** (10 points) In equation 2.13 (textbook page 35), we summed up the squares

of the differences between the actual value and the estimated value. This loss functions

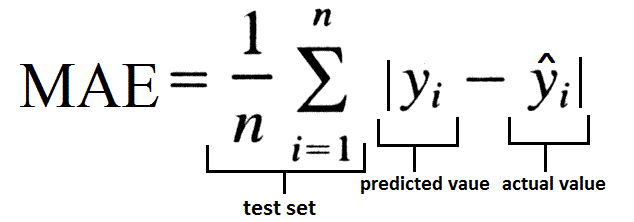
is the one most frequently used for continuous output, but it is one of several possible

loss functions. Because it sums up the squares of the differences, it is not robust to

outliers. Propose a better error function to enable robust regression? Please define all

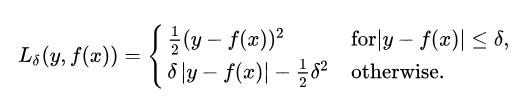
mathematical notations clearly.

**Mean Absolute Error (MEA):** Unlike the square of the difference, the mean absolute error takes the difference between the actual and estimated values and divides it with the number of values. Instead of square the difference, it takes an absolutes value to measure the accuracy of continuous variables. The mathematical notation is as follows:



The MAE is robust to outliers as compared to square difference approach.

**Huber Loss:** Huber loss is combination of mean squared and mean absolute error. It quadratic MSE when error is smaller otherwise MAE. It is identified by its delta parameter that define the range for MAE & MSE which can be tuned to improve it robustness on outlies.



Huber loss is defined as:

* *error 2/2*, if *error < delta* (if it is a small error)
* *delta \* ( |error| - delta/2)*, otherwise ( *|error|* means the absolute value *error*)