Assignment 02
Learning from Data, Related Challenges, Classification

submitted for

EN3150 - Pattern Recognition
Department of Electronic and Telecommunication Engineering
University of Moratuwa

Udugamasooriya P. H. J.
220658U
Progress on GitHub ⌁

20 August 2025

1   Linear Regression

Question 01 The main reason is the presence of outliers.
Based on the inliers, we see that the desired line is an almost horizontal one. To see why a slanted line has been produced instead of a horizontal line, we look at the behavior of the squaring function;
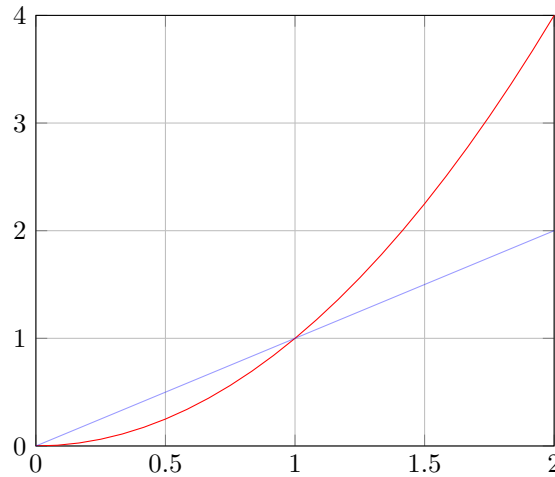

Figure 1: $y = x^2$

Note that for $|x| < 1$, $x^2 < |x| < 1$, i.e., the squares of numbers less than 1 are less than the numbers themselves (in absolute value). This means that the squared loss function can remain small enough for prediction errors that are just slightly below 1, but increases rapidly as it increases beyond 1.

In the given situation, even though a horizontal line would give near-zero error for the inliers, the error due to outliers would be very large. It is reasonable then, that a more optimal, i.e., with a lesser mean squared error, could be achieved with a slanted line, with lesser distance between outliers and the corresponding predictions.

Such a line would increase the gap between the inliers and the corresponding predictions and therefore increase the loss, but *not by too much*, due to the behavior of the square function. Instead,

the reduction in the already-large gap between the outliers and the corresponding predictions, *will* cause a *rapid* decrease in the loss, *exceeding* the increase in the loss due to the inliers.

Question 02 Scheme 1.
The issue we wish to mitigate is the rapid increase in loss due to outliers, while keeping the loss due to inliers significant, with the intention that a model that better fits the inliers and is not strongly affected by the outliers as described above would be preferred.

With Scheme 1, we do not completely neglect the effect of outliers, but significantly scale down their contribution to the loss function, so they do not dominate the loss when the loss due to the inliers is small.

Scheme 2 is obviously not suitable, as it amplifies the loss from outliers over inliers; minimizing such a function would require the error due to outliers to be significantly reduced, causing more slanted line to be preferred, as described in the previous problem.

Question 03 why not linear regression

Question 05

2 Logistic Regression

Question 02 The following error was encountered.

Question 03 Why sagasolver bad.

Question 04 Classification accuracy

Question 05 Why liblinear better

Question 06 Sagasolver accuracy

Question 07 Sagasolver with feature scaling

Question 08 Categorical

3 First and Second Order Methods for Logistic Regression

Question 02 Method to initialize weights

Question 03 Loss function

Question 04 Newton's Method

Question 05 Plots

Question 06 Iterations

Question 07 Convergence