Nile University

# Machine Learning

# Assignment 3

Eman Moustafa Ismail

17th December, 2023

❖ Exploratory Data Analysis (EDA):

➢ Heteroscedastic Data Analysis

- The analysis showed that the data has no outliers—additionally, the features needed to be normalized.
- The Summary statistics along with the visualization showed that the first feature of the data has an almost uniform distribution, whereas the second is nearly normal.

➢ Heteroscedastic Data Preprocessing:

- Given that the data has no null values and no outliers, only data normalization is needed.

➢ Monotonic Data Analysis

- The first intuition was that the data has almost perfect linear data with no scattering point.
- The features have no outliers and no null values
- The summary statistic along the visualization showed that the two features have a uniform distribution

➢ Monotonic Data preprocessing

- The data has no null values and no outliers, so only data normalization is needed.

## ❖ Polynomial regression implementation:

- ➢ As shown in applying the algorithms results, increasing the degree will result in overfitting the model, thus having bad results on the test data and increases the error.
- ➢ These findings are due to the fact that increasing the degree will be fitting perfectly to the training data, but shall behave badly with any other data due to lack of generilzation.
- ➢ In the implementation, the model fitting got applied on the training data. The validation was implemented on the test data and both were plotted against each others.

## ❖ Regulization

- ➢ One possible solution for data overfitting is doing somesort of regulization, this shall decrease the overfitting and improve the generlization.
- ➢ I used Lasso technique
  - With Hetero data, my testing scenario included applying polynomial regression with degree =15. This resulted in 78% error. Whereas after applying the regulization this error decreased to 68%
  - With monotonic data, my testing scenario included applying polynomial regression with degree =15. This resulted in 37% error. Whereas after applying the regulization this error

decreased to 2% . Almost eliminating any overfitting.

- ❖ Finding optimal hyperparameters
  - ➢ Another thing to be taking into consideration is determining the hyperparameters, which could be tricky sometimes.
  - ➢ Applying this technique did not make any difference with the hetero data, but reduced the error with the the monotonic data from 1.4 to 1.3

- ❖ Handing heteroscedasticity:
  - ➢ Having heteroscedasticitic data could be tricky as it gives an equal weight to all of the features/data points. This makes the outliers contribute hardly to the model.
  - ➢ Handling this could be done by giving a different weigh to the features. This could be done by giving the features different weight based on the variance of the error.

## ❖ Handle heteroscedasticity Data:

- ➢ Heteroscedasticity presence makes the variance in the data not constant across all independent.
- ➢ In normal least squares algorithms, each variable is given an equal weight.
- ➢ In Heteroscedasticity data, giving all of the variables equal weights leads to inefficiency in parameter determination.
- ➢ In the function “apply_pol_reg_handle_heteroscedasticity”, each variable is given a certain weight. This weight is given based on the estimated variance of the residuals.