

MATHEMATICS FOR INTELLIGENT SYSTEMS

PROJECT TOPIC : PREDICTING MAXIMUM TEMPERATURE OF A
GIVEN DAY USING VARIOUS PARAMETERS OF WEATHER

BATCH : AIE B

GROUP MEMBERS :

MUHAMMED SHAJAHAN (AM.EN.U4AIE21144)

AKSHAY KRISHNANT (AM.EN.U4AIE21109)

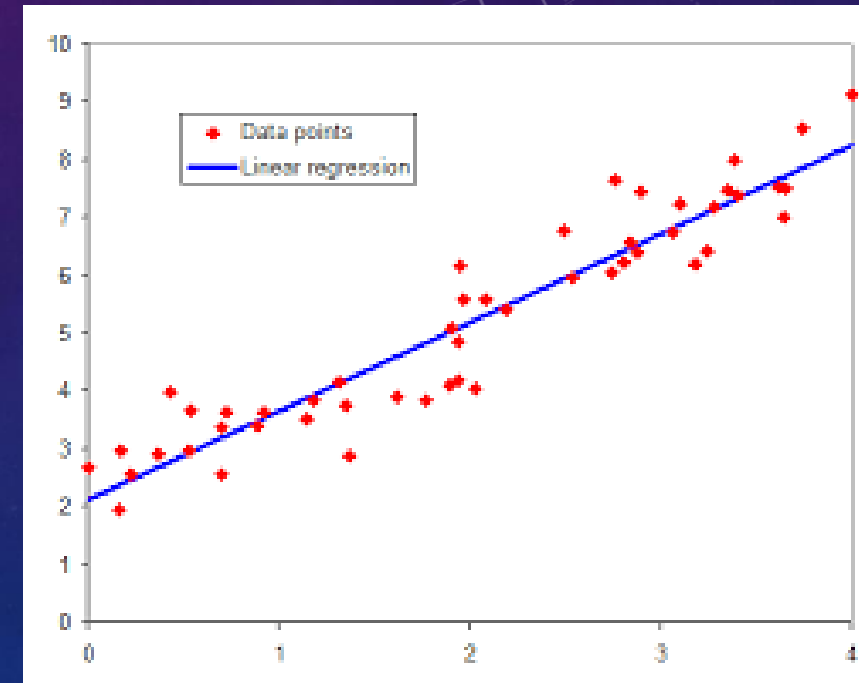
NIRANJAN PRASANTH (AM.EN.U4AIE21148)

ASWIN US (AM.EN.U4AIE21119)

NAVNEETH SURESH (AM.EN.U4AIE21147)

INTRODUCTION

THE PROBLEM IS TO DETECT THE MAXIMUM TEMPERATURE OF A GIVEN DAY USING VARIOUS PARAMETERS OF WEATHER LIKE PRECIPITATION LEVEL, TEMPERATURE, SNOWFALL ETC. THE DATASET IS TAKEN FROM KAGGLE(SUMMARY OF WEATHER). THIS DATASET IS CREATED FROM THE DATA OF THE WEATHER CONDITIONS DURING WW2. THE PROBLEM IS BE SOLVED THROUGH LINEAR REGRESSION ALGORITHM. HERE THE WEATHER CONDITIONS WILL BE THE INPUT VARIABLES AND THE MAXIMUM TEMPERATURE WILL BE THE OUTPUT VARIABLE. THE MODEL WAS BUILT ON GOOGLE COLABORATORY. WE HAVE ALSO USED PYTORCH IN THE CODING. IT WAS USED FOR CREATING THE TRAINING AND THE VALIDATION DATASET. WORKING ON THE PROJECT ALSO HELPED US TO LEARN THE BASICS OF LINEAR REGRESSION CONCEPTS.



PROBLEM

- THE PROBLEM IS TO PREDICT THE MAXIMUM TEMPERATURE OF A GIVEN PROVIDED WE HAVE THE DATASET THAT CONTAINS VARIOUS WEATHER PARAMETERS LIKE TEMPERATURE, PRECIPITATION LEVEL, SNOWFALL, WINDSPEED ETC. THE CHALLENGE IN THE PROBLEM IS TO IDENTIFY THE REQUIRED PARAMETERS AND EXCLUDE THE PARAMETERS NOT REQUIRED. UNDERSTANDING THE DATASET IS ALSO A KEY CHALLENGE AS IT IS THE MOST IMPORTANT PART OF THE PROBLEM. AFTER GENERATING THE REQUIRED DATASET WE HAVE TO BUILT THE LINEAR REGRESSION MODEL. ANOTHER CHALLENGE IN THE PROBLEM IS TO REDUCE THE DIFFERENCE BETWEEN THE ACTUAL VALUE AND THE PREDICTED VALUE. THIS PROBLEM ALSO HELPED US IN UNDERSTANDING VARIOUS CONCEPTS OF LINEAR REGRESSION MODEL.
- THE LINK TO KAGGLE DATASET :
<https://www.kaggle.com/smid80/weatherww2/data?select=Summary+of+Weather.csv>



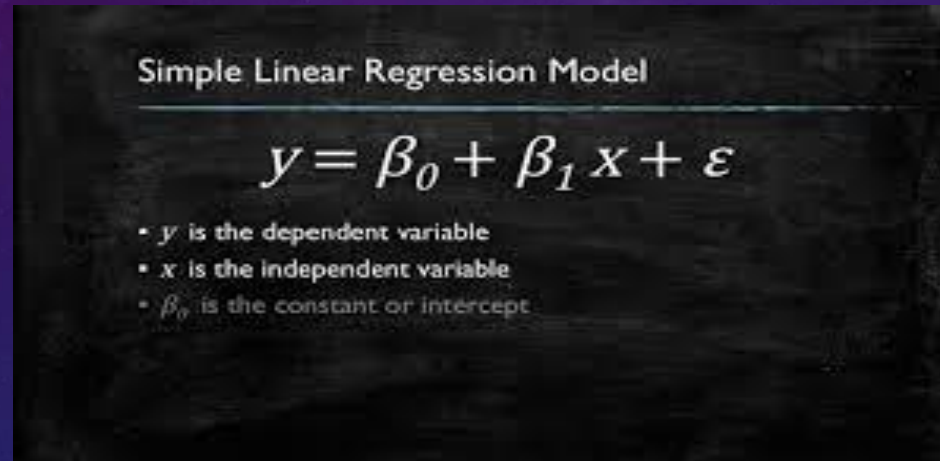
SOLUTION APPROACH

- WE SOLVED THE PROBLEM USING LINEAR REGRESSION.
- LINEAR REGRESSION IS A LINEAR MODEL THAT ASSUMES A LINEAR RELATIONSHIP BETWEEN THE INPUT VARIABLES (X) AND THE SINGLE OUTPUT VARIABLE(Y).
- WE GOT THE SOLUTION BY IMPORTING THE REQUIRED LIBRARIES AND THEN READING THE DATASET INTO A DATAFRAME.
- WE HAVE TO UNDERSTAND THE DATASET, WE USE THIS METHOD AS IT DOES NOT HAVE MUCH COMPLICATIONS BUT WE REALIZE THIS WHEN WE TRY TO DETERMINE INPUT PARAMETERS AND THE OUTPUT DATA.
- NOW WE HAVE TO VISUALIZE THE DATASET AND THEN GENERATE THE DATASET AND SPLIT THE DATASET.
- NEXT WE HAVE TO CREATE A LINEAR REGRESSION MODEL AND TRAIN THE MODEL.
- AFTER CREATING THE LINEAR REGRESSION MODEL WE HAVE TO FIND THE DIFFERENCE BETWEEN THE TARGET VARIABLE AND THE PREDICTED VARIABLE IN ORDER TO REDUCE THE ERRORS.
- FINALLY THE PREDICTION WAS MADE BY THE MODEL
- PYTORCH WAS VERY USEFUL IN THE IMPLEMENTATION OF THE DATASETS SUCH AS GENERATING DATASETS, SPLITTING IT INTO TRAINING SET AND VALIDATION SET AND CREATING TRAINING AND VALIDATION DATA LOADERS.



DESIGN ASPECTS

- LINEAR REGRESSION : LINEAR REGRESSION IS A MECHANISM IN WHICH WE BUILD A LINEAR MODEL WHICH USES ONE OR MORE THAN ONE INPUT PARAMETERS($x_1, x_2, x_3..$) AND USES THEM TO DETERMINE AN OUTPUT(y). IN MATHEMATICAL TERMS, y IS A LINEAR COMBINATION OF THE INPUT PARAMETERS $x_1, x_2, x_3 ..$
- FORM OF THE MODEL MATHEMATICALLY :



Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y is the dependent variable
- x is the independent variable
- β_0 is the constant or intercept

- THE LINEAR MODEL IS DEFINED USING `nn.linear()` METHOD WHICH TAKES 5 INPUTS AND PRODUCES ONE OUTPUT. THIS MODULE IS DESIGNED TO CALCULATE THE LINEAR EQUATION $AX = B$ WHERE x IS INPUT, B IS OUTPUT, A IS VALUE.

- **LOSS FUNCTION** : A LOSS FUNCTION IS A MEASURE OF HOW GOOD A PREDICTION MODEL DOES IN TERMS OF BEING ABLE TO PREDICT THE EXPECTED OUTCOME. A MOST COMMONLY USED METHOD OF FINDING THE MINIMUM POINT OF FUNCTION IS “GRADIENT DESCENT”.
- **MEAN ABSOLUTE ERROR** : MAE IS THE SUM OF ABSOLUTE DIFFERENCES BETWEEN OUR TARGET AND PREDICTED VARIABLES. SO IT MEASURES THE AVERAGE MAGNITUDE OF ERRORS IN A SET OF PREDICTIONS, WITHOUT CONSIDERING THEIR DIRECTIONS.
- **MATHEMATICAL FORMULA OF MEAN ABSOLUTE ERROR IS :**

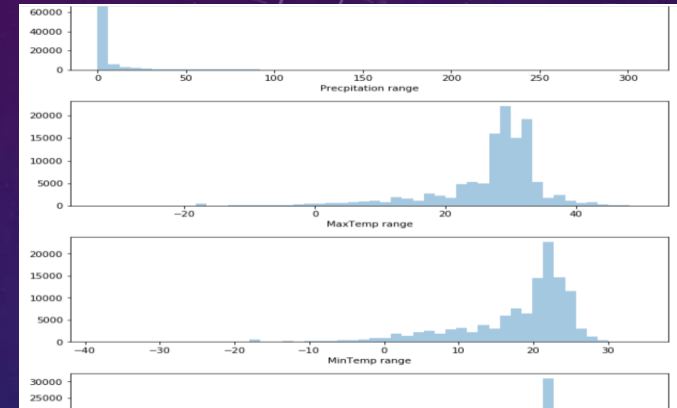
The diagram illustrates the Mean Absolute Error (MAE) formula with the following components and annotations:

- Formula:** $MAE = \frac{1}{n} \sum |y - \hat{y}|$
- Annotations:**
 - Divide by the total number of data points:** Points to the $\frac{1}{n}$ term.
 - Sum of:** Points to the summation symbol \sum .
 - Actual output value:** Points to the y term inside the absolute value.
 - Predicted output value:** Points to the \hat{y} term inside the absolute value.
 - The absolute value of the residual:** Points to the entire absolute value expression $|y - \hat{y}|$.

- **AFTER THE FOLLOWING DESIGN ASPECTS TRAINING WILL BE PERFORMED.**

IMPLEMENTATION

- **VISUALIZING THE DATA:** WE USED THE SEABORN LIBRARY OFFERED BY PYTHON TO VISUALIZE THE DATA AND WE DISPLAYED THE DATA THROUGH HISTOGRAMS.
- **GENERATING THE DATASET, SPLITTING INTO TRAINING AND VALIDATION SET AND CREATING TRAINING AND VALIDATION DATA LOADER:**
- WE HAVE FIRST FORMED THE INPUT PARAMETERS DATAFRAME AND THE TARGETS DATAFRAME
- WE THEN CREATED TENSORDATASET WHICH HELPS US TO ACCESS TO ROWS FROM INPUTS AND TARGETS AS TUPLES.
- IN THIS TUPLE FORM THE FIRST ELEMENT THE FIRST ELEMENT CONTAINS THE INPUT VARIABLES FOR THE SELECTED ROWS, AND THE SECOND CONTAINS THE TARGETS.
- WE ALSO CREATED TWO DATALOADERS *train_loader* AND *val_loader* WHICH IS USED TO SPLIT THE DATA INTO TRAINING DATA AND VALIDATION DATA. WE HAVE ALSO USED *random.split()* FOR RANDOMLY SPLITTING THE DATA FOR UNBIASED EVALUATION. THIS SPLITTING IS DONE BY TAKING 10% OF THE DATA AS VALIDATION SET AND THE REMAINING 90% FOR TRAINING.



model.parameters() is passed as an argument to optim.sgd, so that the optimizer knows which matrices should be modified during the update step.

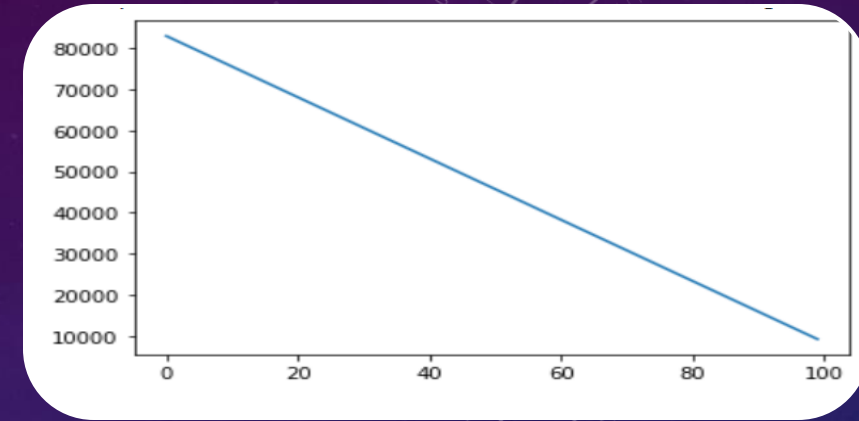
- **CREATING THE REGRESSION MODEL :**

- WE HAVE DEFINED A LINEAR MODEL USING `nn.linear()` WHICH TAKES 5 INPUTS AND PRODUCES ONE OUTPUT.
- THE WEIGHTS OF EACH OF THE PARAMETERS ARE RANDOMLY INITIALIZED BY PYTORCH WHEN WE DEFINE THE LINEAR MODEL.
- THE LOSS FUNCTION WAS IMPLEMENTED ALONG WITH THE MODEL CODE. LOSS FUNCTION HAS BEEN APPLIED IN BOTH TRAINING AND VALIDATION MODELS.
- **OPTIMIZER** : INSTEAD MANUALLY MANIPULATING THE LEARNABLE PARAMETERS WE CAN USE `optim.SGD`(SGD STANDS FOR STOCHASTIC GRADIENT DESCENT) IT IS CALLED STOCHASTIC BECAUSE SAMPLES ARE SELECTED IN BATCHES (OFTEN WITH RANDOM SHUFFLING).
- IS PASSED AS AN ARGUMENT TO SO THAT THE OPTIMIZER KNOWS WHICH MATRICES SHOULD BE MODIFIED DURING THE UPDATE STEP. THE LEARNING RATE IS ALSO PASSED TO THE OPTIMIZER AS A PARAMETER.

- **TRAINING THE MODEL:**

- AFTER IMPLEMENTING THE ABOVE STEPS WE ARE NOW READY TO TRAIN THE MODEL. THE BASIC STEPS TO TRAIN THE MODEL ARE :
 1. GENERATE PREDICTIONS
 2. CALCULATE THE LOSS
 3. COMPUTE GRADIENTS W.R.T THE WEIGHTS AND BIASES
 4. ADJUST THE WEIGHTS BY SUBTRACTING A SMALL QUANTITY PROPORTIONAL TO THE GRADIENT
 5. RESET THE GRADIENTS TO ZERO
- THE TRAINING PROCESS IS FOR 100 EPOCHS AT A LEARNING RATE OF $1e-10$. EPOCHS INDICATES THE NUMBER OF PASSES OF THE ENTIRE TRAINING DATASET THE MACHINE LEARNING ALGORITHM HAS COMPLETED.

- WE ALSO CREATED A LOSS FUNCTION GRAPH FOR BETTER UNDERSTANDING OF LOSS FUNCTION IN THE MODEL WHICH IS GIVEN ON THE RIGHT



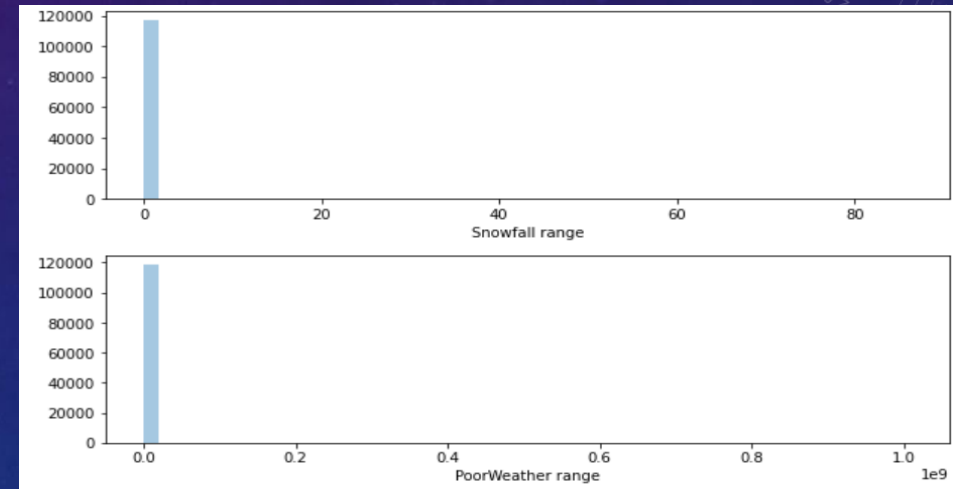
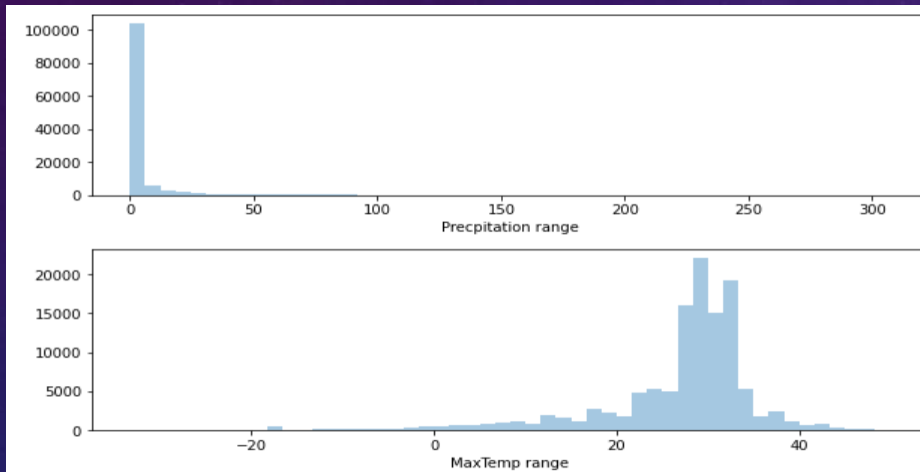
LOSS FUNCTION GRAPH

- THE LAST PART OF THE IMPLEMENTATION IS THE PREDICTION OF THE MAX TEMPERATURE.
- THE THREE COLUMNS IN THE PREDICTION ARE INPUT DATA AND PREDICTION COLUMNS. THE THREE COLUMNS ARE SELECTED RANDOMLY DUE TO THE RANDOM METHOD AND HENCE THE OUTPUTS ARE RANDOM.
- **THE IMPLEMENTATION CODE CAN BE VIEWED HERE**

: https://github.com/niranjanprasanth/TASK_2/blob/main/MaxTemp.ipynb

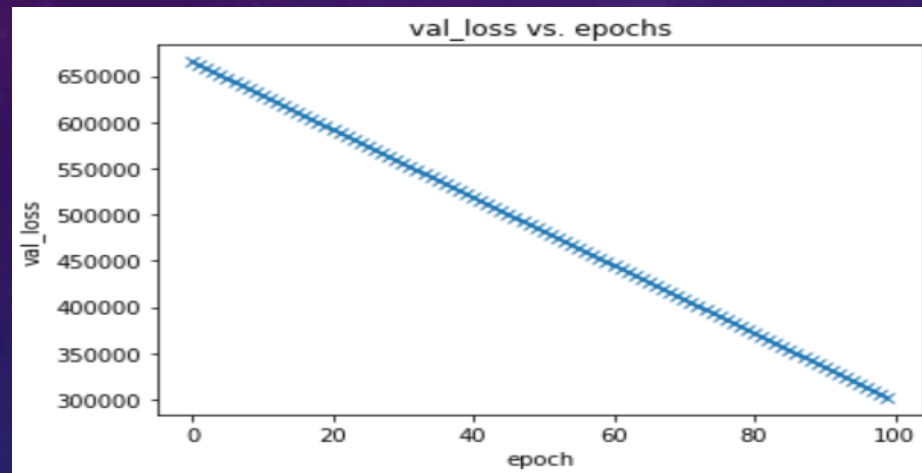
RESULTS

- WE OBSERVED THAT VARIOUS TYPES OF DATA WERE OBSERVED TO IMPLEMENT TO THE APPROACHES FOR PREDICTION. WE USED PRERECORDED DATA. THIS MODEL CAN BE MADE BETTER BY NORMALIZATION, TRYING OUT DIFFERENT LEARNING RATES.
- WE ALSO OBSERVE MULTIPLE HISTOGRAMS CREATED BASED ON PRECIPITATION RANGE, MAX TEMPERATURE RANGE, MIN TEMPERATURE RANGE, MEAN TEMPERATURE RANGE, SNOWFALL RANGE AND POOR WEATHER RANGE. THERE ARE ALSO MULTIPLE GRAPHS.



VISUALIZATION OF DATA SHOWN THROUGH HISTOGRAMS

- IN THIS MODEL WE USE A LOSS FUNCTION WHICH IS BASICALLY A MEASURE OF HOW GOOD A PREDICTION
- MODEL DOES IN ORDER TO PREDICT THE EXPECTED OUTPUT. HENCE LOWER THE LOSS FUNCTION BETTER THE PREDICTION , SO THE MINIMAL LOSS VALUE WILL GIVE A MORE ACCURATE PREDICTION
- THE COMMON METHODS USED ARE GRADIENT DESCENT AND OTHER MEAN ABSOLUTE ERROR.
- OTHER MAJOR OBSERVATION IS THAT OF THE RELATIONSHIP BETWEEN EPOCH AND VAL LOSS(VALUE OF THE COST FUNCTION IN VALIDATION DATA). THIS IS GIVEN THROUGH THE GRAPH BELOW



- IN THE MODEL IN PREDICTION THERE ARE THREE ITEMS INPUT TARGET AND PREDICTION, IN THIS CASE PREDICTION WILL BE RANDOM AS THE ITEMS ARE RANDOM. HERE THE INPUT WILL BE A TUPLE.
- **RESEARCH GAP** : A BETTER PROCESSED DATA IS REQUIRED FOR MORE ACCURATE PREDICTIONS AND THE MODELS CAN PERFORM BETTER WHEN MULTIPLE LINEAR REGRESSION METHODS CAN BE USED HENCE ALLOWING MORE WEATHER FACTORS TO AFFECT THE MODEL.

REFERENCES

- <https://medium.com/swlh/using-linear-regression-to-predict-max-temperature-based-on-weather-conditions-2d776947cc2d>
- https://www.researchgate.net/figure/Research-gap-for-existing-predictive-analytics-methods_tbl1_326071541
- <https://heartbeat.comet.ml/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0>
- https://www.youtube.com/watch?v=vo_fUOk-IKk&t=58s



THANK YOU