How ml works

- There are several stages involved

  1. Data collection

     - Gathering the data

       - Structured : spreadsheet
       - unstructured : Images, audio

  2. Data preprocessing

     - cleaning and transforming the raw data into Sustainable format, this includes handling missing values, removing noise & feature engineering

  3. model selection

     - Choosing an appropriate ML algorithm like

       - Regression
       - classification

     - model tries to predict

4. Training

- Feeding the prepared data to the choosen algorithm allowing it to recognize patterns and algorithm

5. Evaluation

- Assesing the model performance on unseen data to ensure that it isn't memorizing the data

- we compare prediction and reality
  - The difference = error

6. Hyper parametre tuning

- Adjusting the external configurational parametres of the model to optimise its performance

- Repeating the loop Eventually, the model becomes good at prediction

Keywords

1. Attribute :- datatype ("Mileage")

2. feature :- feature = attribute + Value (Mileage = 45)

3. Residual Error: Actual - prediction

performance measurement :-

- In ML we measure the performce, because model makes error and we measure the error using
  - RMSE
  - MSE
  - MAE
  - accuracy
  - precision
  - R^2 score
  - F1 score

1. Accuracy

- How many total predictions were correct?

$$\text{Accuracy} = \frac{\text{correct prediction}}{\text{total prediction}} \%$$

Ex:-

Suppose, there are 100 emails

- 90 correctly classified
- 10 wrong

$$\text{accuracy} = \frac{\text{Total} - \text{wrong} = \text{correct}}{\text{total}} \%$$

$$= \frac{100-10}{100}$$

$$= \frac{90}{100}$$

$$= 0.9 * 100$$

$$= 90\%$$

accuracy fails if

- 95% of emails are not spam
- model always says not spam

2. precision

- Of the items predicted positive, how many were truly positive

$$\text{precision} = \frac{TP}{TP + FP}$$

where:

TP = True positive

FP = False positive

Ex:-

model says 20 emails are spam

→ 15 really Spam

→ 5 not spam

$$\text{precision} = \frac{15}{15+5} = \frac{15}{20}$$

$$= 75\% \quad (75\% \text{ Correct})$$

3. Root mean square error (RMSE)

- used to measure how much error does it make during prediction

$$rmse(x, h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((h(x_i)) - y_i)^2}$$

n = no. of Instances

x_i = feature vector of ith instance

y_i = label

h = System prediction, called hypothesis

$$h(x_i) = \hat{y}$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}(residual_i)^2$$

code:-

```
rmse = (1/n) * ((prediction - actual) ** 2)
```

```
• for i in range(len(data)):
    pred = predict(x, y, z)
    rmse = (1/len(data)) * ((pred(i) - actual(i)) * * 2)
    rmse = √ mse
```

4- mean square error

- looks at average magnitude of error

$$\frac{1}{n}\sum_{i=1}^{n}\left[h(x_i) - y_i\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\text{residual}\right]^2$$

5- mean absolute error

- Look at average magnitude of error

$$= \frac{1}{n}\sum_{i=1}^{n}\left|h(x_i) - y_i\right|$$

$$h(x_i) = \hat{y}$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}\left|\text{residual}\right|$$

6. Recall

- Of all actual positives, how many did we catch correctly

$$R = \frac{TP}{TP + FN}$$

TP = True positive

FN = False negative

Ex:-

100 sick patients, model finds 80

$$R = \frac{80}{100} = 0.8$$

$$= 80\%$$

7) F1 Score

- A Single score balancing false ~~alq~~ alarms and misses

$$F1 = \frac{2PR}{P+R}$$

P = precision

r = recall

$$F1 = \frac{2 \cdot \left(\frac{TP}{TP+FP}\right) \cdot \left(\frac{TP}{TP+FN}\right)}{\left(\frac{TP}{TP+FP}\right) + \left(\frac{TP}{TP+FN}\right)}$$

Ex:-

precision = 0.5

recal = 1.0

$$F1 = \frac{2 \times 0.5 * 1}{0.5 + 1}$$

$$= \frac{2 \times 0.5}{1.5}$$

$$= 0.67$$

8) R^2 score

- It measure how well the model explain the variance in the data

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res} = residual sum of squares

$$SS_{res} = \sum (\hat{y}_t - y)^2$$

SS_{tot} = total sum of squares

$$SS_{tot} = \sum (y_t - \hat{y})^2$$

Outlier:-

data that behaves differently from the rest of the data