6/7/21

~~Linear Neural Networks~~

Basics of Machine Learning

- Supervised learning
- Unsupervised learning
- Semi-Supervised Learning.

Supervised Learning Algorithms

~ Trained using labeled examples
- Input where the desired output is known.

eg

~ Spam Vs Legimate Email

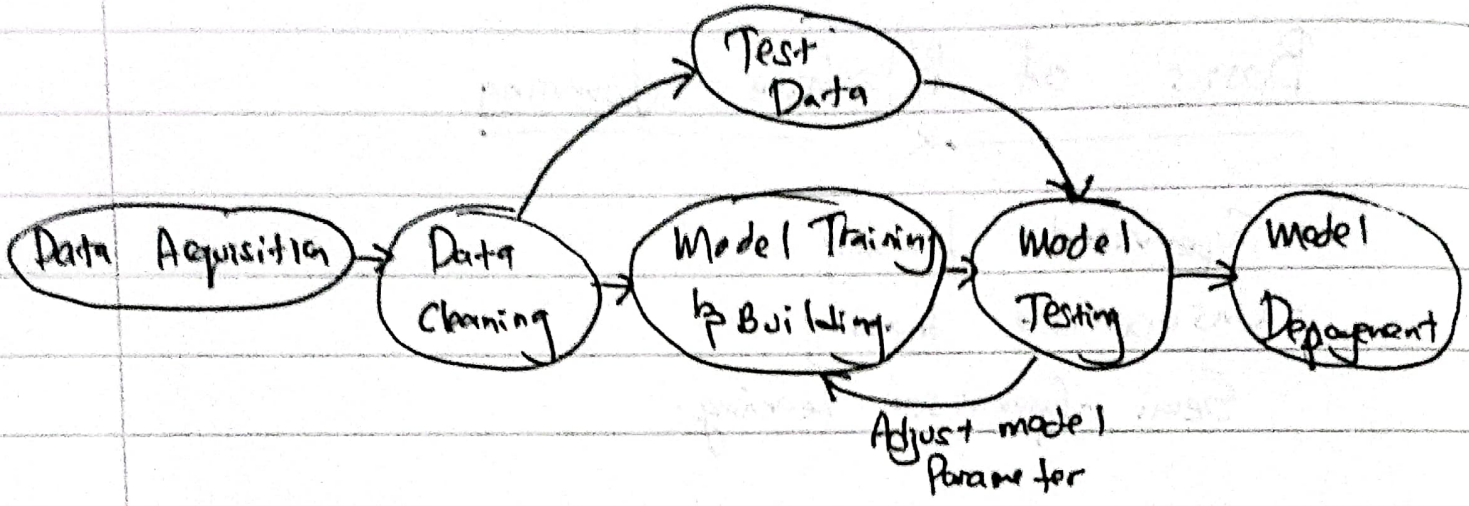
~ Positive vs Negative Movie Review

~ Receives a set of inputs along with correct outputs.

~ The algorithm learns by comparing its actual output with correct output to find errors.

~ It's then modifies accordingly.

Machine Learning Process for Supervised learning:



Data Acquisition

Get your data from Customers, Sensors.

Data Cleaning

Clean & format data (Using Pandas).

Issue

Is it fair to use a single split of the data to evaluate our models performance.

Fix Issue

Data split into 3 sets.

1) Training Data
- Used to train model parameters.

2) Validation Data
- Used to determine what model hyperparameter to adjust

3) Test Data.
- Used to get some final performance metric
- Performance in real world.

X cannot go back

Evaluating Performance Classification.

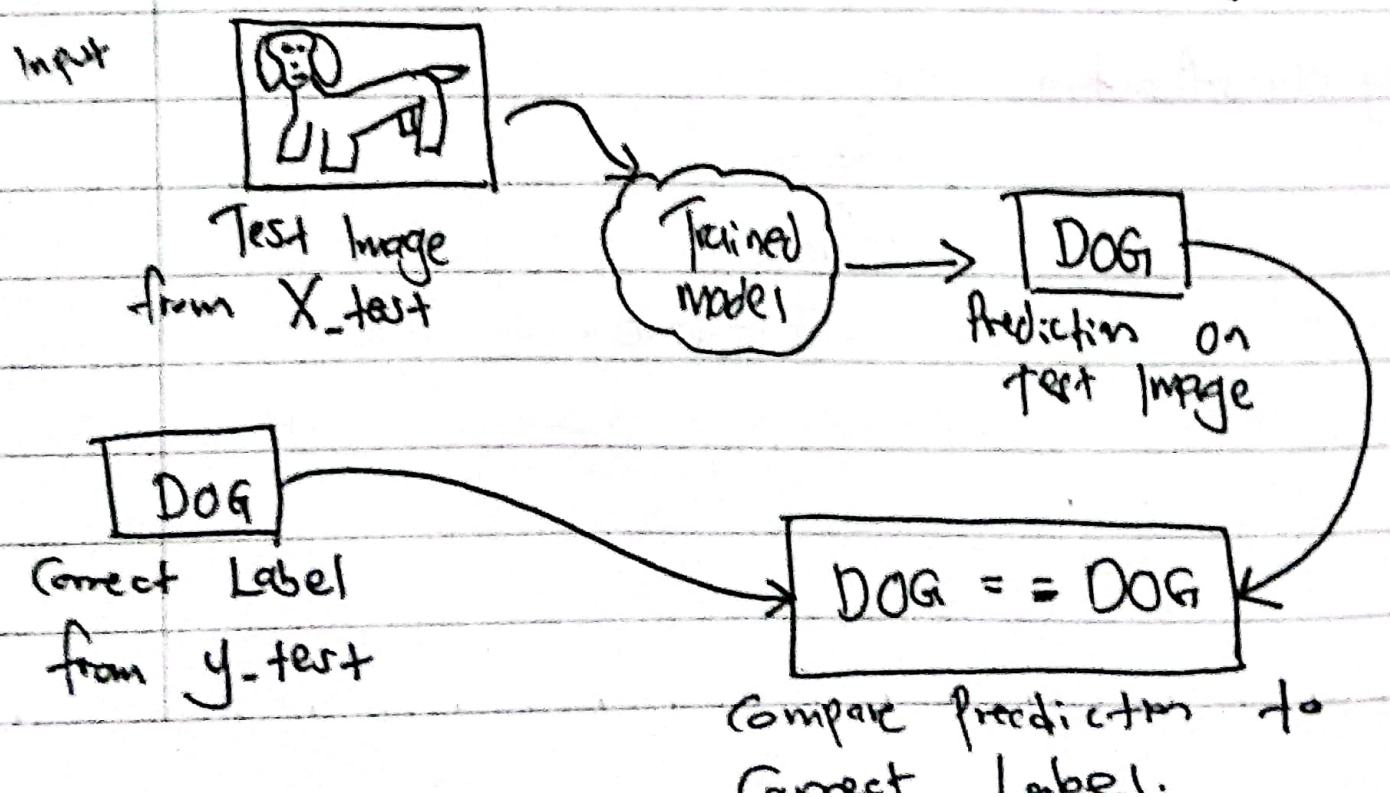
Key classification
- ~ Accuracy
- ~ Recall
- ~ Precision
- ~ F1-Score.

- Any classification task model can achieve :
  - Correct in its prediction
  - Incorrect in its prediction.

~~text~~ Let's imagine a binary classification :
- binary = 2.

Example Predict Image ~~Cat~~ Cat or Dog :
- Supervised Learning ~ fit / train a model on training data.
  ~ test the model on testing data.

~ We compare X_test data with true y values (the correct labels).



- This process is repeated for all images in our X test data.

- We will have count of correct & a count of incorrect matches.

- Not all correct & incorrect matches hold equal values in the real world.

Accuracy

- Useful when target classes are well balanced.

- if we have image (50 dogs & 50 cats)
  The accuracy will be 50%
  Hence we can use that as a good metric to use.

- Not a good choice with unbalanced classes
- If we have (99 dogs & 1 cat).
- The accuracy will be 99% because it only missed 1.
- Hence it is not a good metric to use.

Recall

- Ability of a model to find all the relevant cases within a dataset.
- definition :

Number of true positive / the number of true positives + the number of false negatives.

Precision

- Ability of classification model to identify only the relevant data points.
- definition :

Number of true positives / Number of true positives + Number of false positives.

Recall & Precision

~ Recall express the ability to find all relevant instances in a dataset.

~ Precision expresses the proportion of the data points our model says relevant was actually relevant (Mate sure).

F1 - Score

- To find optimal blend of precision & recall we combine the two ~~metric~~ metric (Precision & Recall) using F1-Score.

- F1 - Score, harmonic mean of precision & recall.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- We use harmonic mean instead of simple average because it punishes extreme values.

- A classifier with a precision of 1.0 & a recall of 0.00 has a simple average of $\frac{1.0}{2} = 0.5$.
- But in F1 Score $\frac{1 \times 0}{1 + 0} = 0$ ✱.

more accurate.

Confusion Matrix.

What our model thinks.

True Positive (TP) - eg. A patient have the disease & Our model predicted positive.

True Negative (TN) - eg. A patient does not have disease & Our model predicted negative.

False Positive - eg. A patient does not have disease. Our model predicted positive.

False Negative - eg. A patient do have the disease & Our model predicted negative.

$$\text{Accuracy} = \frac{\sum TP + \sum TN}{\sum \text{total population}}$$

Example.

In case of patients with covid-19.

- It better to increase False positive to reduce false Negative.

- We can prevent patients with covid identified.

- Even-lough False ~~Negati~~ Positive is higher we can run some more test to reduce The false positive.

- This makes us to detect patient with covid-19 ~~more~~ to classify as many cases of disease as possible.

Evaluation Performance

Regression

- Regression is a task when a model attempts to predict continuous values. (unlike categorial values, which is classification).

Metric for Regression:

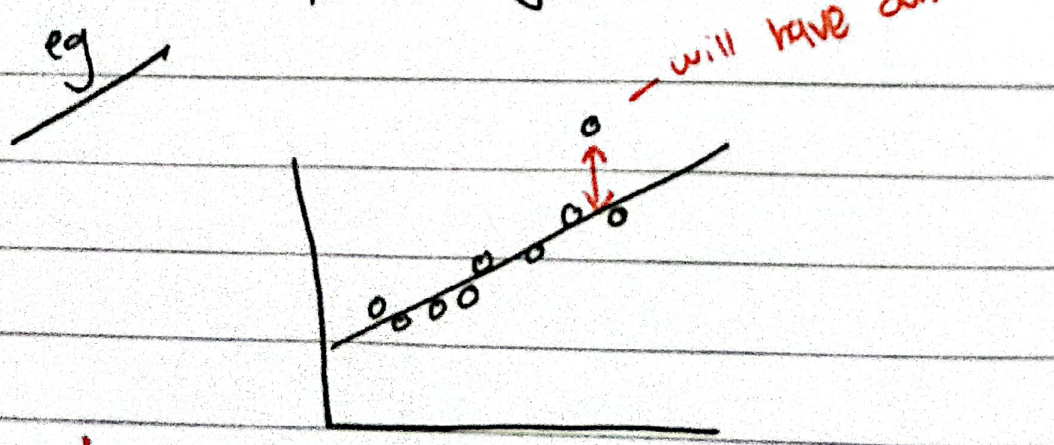
- Mean Absolute Error
- Mean squared Error
- Root Mean Square Error.

* Mean Absolute Error (MAE)

$$\frac{1}{n}\sum_{i}^{n} = \left| \underset{\text{true}}{y_i} - \overset{\text{predicted}}{\hat{y}_i} \right|$$

- Mae won't punish large error

eg



- Larger errors won't show aggresive enough.

* Mean squared Error (MSE).

- Larger errors are noted more than with MAE

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

- Sometime when using MSE if we use currency.
  - $(RM)^2$ is difficult to interpret.

* Root Mean Square (RMSE). * Popular.

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

Compare error metric to the average value of label in a data set.