03/05/23.

Comparing algorithm - tests.

* t-test - The accuracy of the models is compared.
* It is used to compare different classifiers such as SVM, DT, RF, etc.
* First take mean.
  - * But mean will not always give the accurate results.
  - * Because mean is always highly affected by the datapoints. Whenever new data point is added the change in mean occur.

* So we will go to t-test.

Why t-test?

* : A The choice of applying test depends on the sample size.
* Z-test is used when there is a large sample. t-test is used for small sample.
* In machine learning, ~~work~~ we use real time data, where we need to make assumption normality. t-test is more robust in this case.

Hypothesis testing :-

- It is a statistical technique which is used to make decisions based on sample data.

Steps :-

1. Define null hypothesis and alternate hypothesis.

H_0 :- There is no difference in the mean accuracy.

H_1 :- There is difference in the accuracy of the model.

* Choose significance level -
  It is the probability of rejecting the null hypothesis.
  * Usually 0.05 is taken as the level of significance - 95% confidence H_0 is true
    5% confidence H_0 is false.

* Collect the data.
* Calculate t - statistic

⇒ If we are comparing same classifiers then it is known as related t-test.

⇒ If we are comparing different classifiers then it is known as paired independent t-test.

* p-value is determined.
* p-value is the probability of obtaining a test statistic as extreme or more extreme than observed
  - assuming H_0 is true
    ↳ area under t-distribution curve that is more extreme than observed value of t

formula of test test:

$$t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$$

=> $\bar{x}$ - sample mean
μ - population mean
s - standard deviation
n - sample size.

$$t = (\bar{x}_1 - \bar{x}_2) / (s_p * \sqrt{1/n_1 + 1/n_2})$$

s_p = pooled standard deviation

~~$A_1 \& A_2$~~

$$s_p = \sqrt{((s_1^2 + s_2^2)/2}$$

s_1 & s_2 are standard deviations of two samples.

n_1 & n_2 are sample size