← Go Back to Model Tuning

Course Content

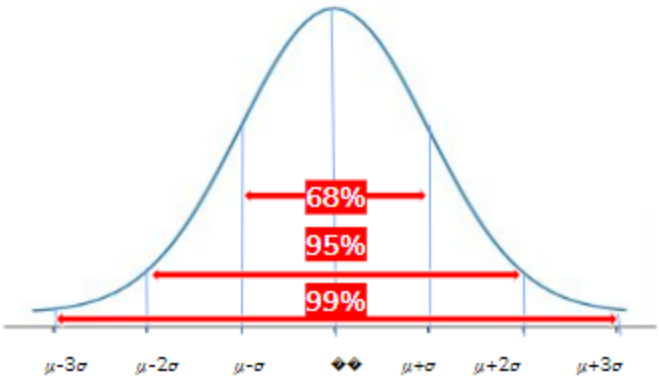
Confidence Interval [Confidence Interval will be introduced by the professor in the following lecture, let's have a quick look!]

We all certainly make claims in our day-to-day life and we commit so many statements with confidence. For eg, I am 90% sure that I can finish the assessment by this weekend. Similar claims can be made with respect to the machine learning models as well. Let's get familiar with a few terminologies before we understand confidence intervals.

Mean: It is the expected/average value of the data which is calculated by adding the data points and dividing them by the total number of data points. It is denoted by ω

Standard deviation: It is a measure of dispersion that tells you how dispersed is the data with respect to the mean. It is denoted by σ.

Normal Distribution: When we plot the data points, it forms a bell curve and in simple terms, the data points lie around the mean value. The curve is symmetric around the mean.



Properties of the normal distribution:

1. The total area under the curve is 1
2. About 68% of the data fall within 1 standard deviation from the mean i.e (Ɯ +- σ)
3. About 95% of the data fall within 2 standard deviations from the mean i.e (Ɯ +- 2 σ)
4. About 99.7% of the data fall within 3 standard deviations from the mean i.e (Ɯ +- 3 σ)

Confidence interval: Confidence interval measures the certainty or uncertainty of an event in an interval that indicates the probability of occurrence of an event will fall within the pair of values about the mean.

Seems confusing?

Don't Worry! Let's break down the definition of the confidence interval and understand it with an example.

Let's say we are building a model to classify if a person will buy a product or not. When we are ready with our model and we pitch this to a business team that they can use the model to make business decisions. The team will ask how confident are you that your model will give x% accuracy?

To answer the business team, we validate our model on 10 different test sets and we come up with 10 accuracy scores i.e [0.77, 0.81, 0.81, 0.83, 0.72, 0.85, 0.74, 0.66, 0.77, 0.79].

Now we know from the normal distribution that 95% of the data is covered within 2 Standard deviations from the mean. From the 10 accuracy scores obtained on training the model, we can be 95% sure that our accuracy will lie within 2 standard deviations from the mean.

Mean = $\sum \frac{Xi}{n}$ = $\frac{0.77 + 0.81 + 0.81 + 0.83 + 0.72 + 0.85 + 0.74 + 0.66 + 0.77 + 0.79}{10}$ = 77.82%

Standard deviation = $\sqrt{\sum \frac{(Xi - mean)^2}{N}}$ = $\sqrt{\sum \frac{(0.77 - 0.7782)^2 + (0.81 - 0.7782)^2 + ... + (0.79 - 0.7782)^2}{10}}$ = 5.46%

The mean accuracy obtained from the model is 77.82 and the standard deviation of accuracy is 5.46.

We can confidently pitch our model to the business team by claiming that we are 95% confident that the accuracy of our model will lie within the range $[77.82 \pm 2 * 5.46]$ = $[66.9,\ 88.74]$

Note - We encourage you to go through the self-paced course “Statistical Learning” for detailed understanding. You can refer to the “Estimation” video under the “Inferential Statistics”.

Previous Next