Module 3

Variance (Part 1)

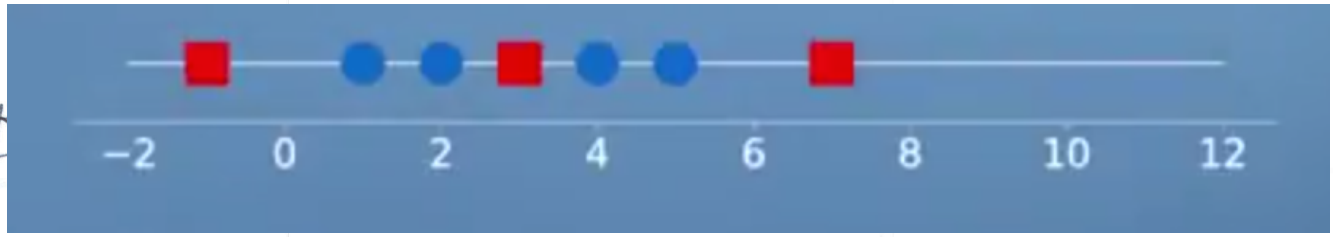
(Describes the spread of data around the mean value)

Lets look at 2 data set $D_1$ and $D_2$

$D_1$ is represented by (blue dots): 1, 3, 4, 5

$D_2$ is represented by (red squares): -1, 3, 7

(see pic.)




$D_1$ and $D_2$ has the same mean, which is 3, but the data points in $D_2$ are less concentrated around the mean than the data points in $D_1$.

Remember, the mean value is the data that you expect on average, but to describe the concentration of data around the mean value, we can use the concept of Variance

The variance is used to characterize the variability or spread of data points in dataset.

In 1Dim we can look at the average square distance of datapoint from mean value of the dataset.

Let do it for $D_1$ and $D_2$

$$D_1 = \{1, 2, 4, 5\} \quad E[D_1] = 3$$

and expected mean value was 3

$$D_2 = \{-1, 3, 7\} \quad E[D_2] = 3$$

Now we want to compute the average square distance (from $D_1$) of $D_1$ from mean and from $D_2$ from the same mean.

Let do it for $D_1$ first:

$$D_1: \frac{(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2}{4}$$

$$= \frac{4+1+1+4}{4} = \frac{10}{4} \quad (3)$$

$$D_2: \frac{(-1-3)^2+(3-3)^2+(7-3)^2}{3} = \frac{16+0+16}{3} = \frac{32}{3} \text{ (A)}$$

- (A) is now bigger than (B), which means that the average square distance of $D_2$ from mean value, is bigger than average square distance of $D_1$ from mean value.

which indicates that the spread of data is higher in $D_2$ than in $D_1$.

So how can we formalize what we have done? we can define the averge square distance as follows:

$$\{x_1, \ldots, x_N\} \Rightarrow \text{define this as our dataset } X.$$

$$\therefore \text{variance}: \mathrm{Var}[x] = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

where $\mu$ is mean value of data set $X$.

$$\therefore \mu = E[x]$$

What we have done here is exactly the same as we did before with $D_1$ and $D_2$, we computed an average square distance of data points in dataset from mean value ($\mu$) of dataset.

Now, we can also make some statement about it

First, the variance as defined here, can never be negative, as we just sum up square values that also means, we can take square root of variance, and this is called the standard deviation.

The standard deviation is expressed in the same units as the mean value, whereas the variance is unfortunately expressed in squared unit

So comparing them is quite difficult.

Before when we talk about Spread of data, we usually look at the standard deviation

So, in the session, we looked at variances of 1 Dim data set, in next session, we will (generalize) generalize it to higher Dims.