Lecture 2 (Lesson 6 to 10)

Types of Data:

1. Numerical data: int, float
2. Time Series data: It can be data collected during fixed and sporadic time intervals like real time performance, energy demand forecasting. It can also be multiple series , evolution of sales of a product in multiple locations.
3. Categorical data: E.g. Gender, Ethnicity. Note that some categorical variables may also have high cardinality e.g. ID's , SKU 's etc.
4. Text : News articles
5. Image : Picture , snapshots , videos etc.

Tabular Data:

This is data that is arranged in a data table, e.g. spread sheet

Each Column has a unique feature, and each row is a new entity.

Scaling Data

Scaling data means transforming it so that the values fit within some range or scale, such as 0–100 or 0–1.

Some necessary Mathematical Concepts:

- **Mean :** A mean is the simple mathematical average of a set of two or more numbers. It is denoted by **μ**
- **Standard Deviation :** The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. It is denoted by **σ**

Two common approaches to scaling data include standardization and normalization.

Feature	Standardization	Normalization
Rescaling	For **μ** =0 and **σ** =1	To make data in the range [0,1]
Formula	$\frac{(x-\mu)}{\sigma}$	$\frac{(x-x_{min})}{(x_{max}-x_{min})}$

Encoding Categorical Data

There are many ways we can encode the categorical variables as numbers and use them in an algorithm. The two approaches discussed are :

1. **Ordinal Encoding:** In this encoding, each category is assigned a value from 1 through N (here N is the number of categories for the feature. One major issue with this approach is there is no relation or order between these classes, but the algorithm might consider them as some order, or there is some implicit relationship.

2. **One –Hot Encoding:** In this method, we map each category to a vector that contains 1 and 0 denoting the presence or absence of the feature. The number of vectors depends on the number of categories for features. This method produces a lot of columns that slows down the learning significantly if the number of the category is very high for the feature.

Image Data

In ML, image can be represented in terms of pixels. A pixel can be a tiny combination of 3 colour channels, RED, GREEN, BLUE. Lots of these pixels come together to form a digital image.

- Total Number of pixels in an image = Height * Width

- Each image can be described as a 3 dimensional vector :
  Height* Width * Channel Value

Note:

1. Important to use uniform aspect ratio for all images (Squares are common)
2. Image data is typically normalised to subtract per channel mean pixel values.

Once an Image is turned into a numeric vector, it can be used, to train various Machine Learning algorithms