BASIC Terminologies

What is the reason need to use dimensionality techniques like PCA, t-SNE and clustering algorithms like K-Means clustering, Gaussian Mixture Models, Hierachial clustering and DBSCAN for **unsupervised learning** - This is due to lack of **Labeled data**

Description	What	Example/Formula/Limitations	Reference/Comments
Supervised Learning	Classified into Regression and Classification Problems		
Continuous variable	Takes **infinite number** of distinct numerical values possibly in a given range of numbers	**Monthly Income** of employees in a certain firm	
Categorical Variables	Can take only a **limited (finite) number** of distinct values	Image dataset of single handwritten digits (as number of distinct values is 0 to 9)	
Dependent Variables	In Data Science, for a given set of variables we need to establish the relationship between one variable and others. The variable to be estimated is **dependent** on the rest of the variables	Independent Variables - Age, Education,Level and Work Experience **Dependent Variable** - Salary	
Independent Variables	In Data Science, for a given set of variables we need to establish the relationship between one variable and others. The variable that affects the dependent variables are called **independent variables.**	**Independent Variables** - Age, Education,Level and Work Experience Dependent Variable - Salary	
Magnitude of the observed data from the Mea	**Variance** and **standard deviation** are two quantities that address this concept		
Variance	Take the difference between each number in the dataset and the mean of the data , square this difference to make it positive (independent of sign) and finally divide the sum of squares by the total number of values in the dataset	The major drawback of using variance, to understand the spread of data is its interpretability. The unit of variance is the square of the original unit of the data.	
Standard Deviation	The **square root** of the variance is also called the **standard deviation of the population**.	The standard deviation is able to give a sense for the measure of spread of the dataset around its mean	
Infrential Statistics	is associated with estimating the population parameters by extracting samples from the same population		

Point Estimate	When we make an estimate about some quantity of the population (for example, mean) we come up with a single number. This single number is called a point estimate	If we take a sample from a population and the sample mean is 35, then our most reasonable estimate of the population mean is also 35. The drawback of the point estimates is that we do not know **how sure** we can be that population mean is close to 35.	
Confidence Interval	To increase the informativeness of our estimate, we associate it with another concept known as Confidence Interval. A confidence interval is a **range of values around the point estimate**, constructed so that this range will contain the population parameter with a **certain degree of confidence**, expressed in percentage terms. **The higher the confidence, the greater the width of the confidence interval.**	Extract 100 samples (data sets) from a group of students ina. university, where each sample has a certain number of records. Calculate mean age of students from each sample. Construct confidence intervals (95% confidence level) using the method that will be covered in the lecture 1. 95% of samples (data sets) are expected to result in confidence intervals that contain the true population mean.	

Supervised Learning - Linear Regression

Description	What	Example/Formula	Reference/Comments
Labeled Data	Data is already tagged with the correct output		
Regression Algorithm (Supervised Learning)	Output variable is a **continous variable**	Price of a house	
Classification Algorithm (Supervised Learning)	Output variable is **categorical variable**	Loan approval - yes or no categories	
Linear Regression	Useful for for finding the **linear relationship** between the **independent and dependent variables** of a dataset	$Y=O^*_0+O^*_1X+W$ Where O^*_0 is the contstant term O^*_1 is the coefficient of the variable X W is the difference between the actual value and predicted value ($O^*_0+O^*_1$) X is the independent variable Y is the dependent variable	
Error	The **difference** between the actual value and the predicted value is called the **error or residual**		
Linear Regression Model		$Y=O^*_0'+O^*_1'X$ Where O^*_0 is the contstant term O^*_1 is the coefficient of the variable X W is the difference between the actual value and predicted value ($O^*_0+O^*_1$) X is the independent variable Y is the dependent variable	
R-squared	Is a useful performance metric to understand **how well** the regression model has **fitted over the training data**. A **higer R-squared** value indicates a **better fit for the model.** The R-squared always increases when a new independent variable is added to the model, irrespective of whether that variables adds value to the model or not. Misleading when we have multiple independent variables and cannot identify unnecessary variables included in the variable	R-squared of 80% reveals that the model is able to capture 80% of the variation in the dependent variable	

Adjusted R-squared	The **adjusted R-squared** is a modified version of R-squared that takes into the account the number of **independent variables present in the model.** The adjusted r-squared increases if that variable adds value to the model, and decreases if it does not. Adjusted R-squared is a better choice of metric than R-squared to evaluate the quality of a regression model with independent variables (Reason - adjusted R-squared remains high when all those independent variables are required to predict the value of the dependent variable well. It decreases if there are any independent variables which don't have a significant effect on the predicted variable)		
RSME	Root Mean Squared Error - Square root of the mean of the squared differences between the actual output and the predictions. Lower the RSME the better performance of the model		

To establish between variables it is required to calculated the parameters of the model			

Maximum Likelihood and Empirical Risk Estimation

Description	What	Example/Formula	Reference/Comments
Likelihood of the event	Probability of occurence of the event		
Multiplication rule of probability	According to the multiplication rule of probability, the probability of occurence of both the events A and B is equal to the product of the proability of B occuring and the conditional probability that event A occuring given that B occurs. Likelihood of entire sample of events where the events in the sample are independ of each other we can apply multiplication rule of probability		
Maximum Likelihood	is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed		
Empirical Risk	Collective error made by the model over all the training records.		
Principle of Empircal Risk minimization	Empirical risk is minimized to get the model parameters. A higher empirical risk is quivalent to a lower likelihood and viceversa		