1. **Diff between classification and clustering**

**Clsutering: Grouping data wrt their similarities. It is unsupervised learning where grouping is based on a natural or inheritance characteristics.Here we work on full data and no splitting require.**

**Classification: Classify data wrt labels to existing classes. It’s a supervised learning. Here we divide data into training and test.**

**b) What is feature space in classification?**

Ans: Feature space just refers to the collections of features that are used to characterize your data. For example, if your data is about people, your feature space might be (Gender, Height, Weight, Age)

Feature Space refers to n-dimensions where variables live but it doesn’t include target variable if its available.

Suppose Y= target and X1,X2,X2 = Variables

So here feature space is R^3. If we include new features then feature space will increase.

https://stats.stackexchange.com/questions/46425/what-is-feature-space

**c) Can we use 2 dependent variables in clustering?**

**d) Working/methodology of clustering**

**e) Homogeneous point**

**f) Euclidean distance**

Calculate distance between two vector points.

**g) Centroid vs Medoid**

**h) Mean (Sum of all value divided by no of values) vs Median(Middle value if odd or sum of two middle value divided by 2 if even) vs Mode (most frequent number in data)**

**i) Assumptions of Linear Regression**

**j) Confusion Metrix, recall and precision**

<https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>

**Confusion Matrix: It’s a summary of prediction results of a classification problem**

|  |  |  |
| --- | --- | --- |
|  | **Class 1 Predicted** | **Class 2 Predicted** |
| **Class 1 Actual** | **TP** | **FN** |
| **Class 2 Actual** | **FP** | **TN** |

**Here Class 1 is Positive and Class 2 is negative**

**TP = Actual and predicted both are positive value**

**TN: Actual and predicted both are negative value**

**FN: Actual is positive and predicted negative**

**FP: Actual is negative and predicted positive (Most dangerous)**

**Classification Accuracy: TP+TN / TN+TP+FN+FP**

**Recall: It checks how many values from actual positive are predicted positive**

**TP / TP + FN**

**Precision: It checks how many predicted values are actual positive**

**TP/ TP + FP**

**High recall low precision: Most of the positive values are predicted positive (low FN) but there are more FP**

**High precision low recall: We missed lots of positive example but whatever we have got are actual positive (low FP)**

**F-measure: It’s a Harmonic relation between recall and precision**

**2 \* Recall \* Precision / Recall + Precision**

**A better model should have high precision and high recall**