**SEMI-SUPERVISED LEARNING**

Mondeep Das(40016937)

horizontal line

# 

# 1. Introduction

The most basic disadvantage of any supervised learning algorithm is that the dataset has to be hand-labelled either by machine Learning engineer or a data scientist.This is a very costly process especially when dealing with large volumes of data.

Similarly for unsupervised learning, its problem space is very confined. So to tackle the same a newer method is coming into picture that is **SEMI-SUPERVISED** learning.

## 1. Deciding factors on algorithms:

**Supervised:**

1. Domain includes well defined solutions.
2. Small dataset.
3. Labelled dataset.

**Unsupervised:**

1. Larger dataset.
2. Real-time application

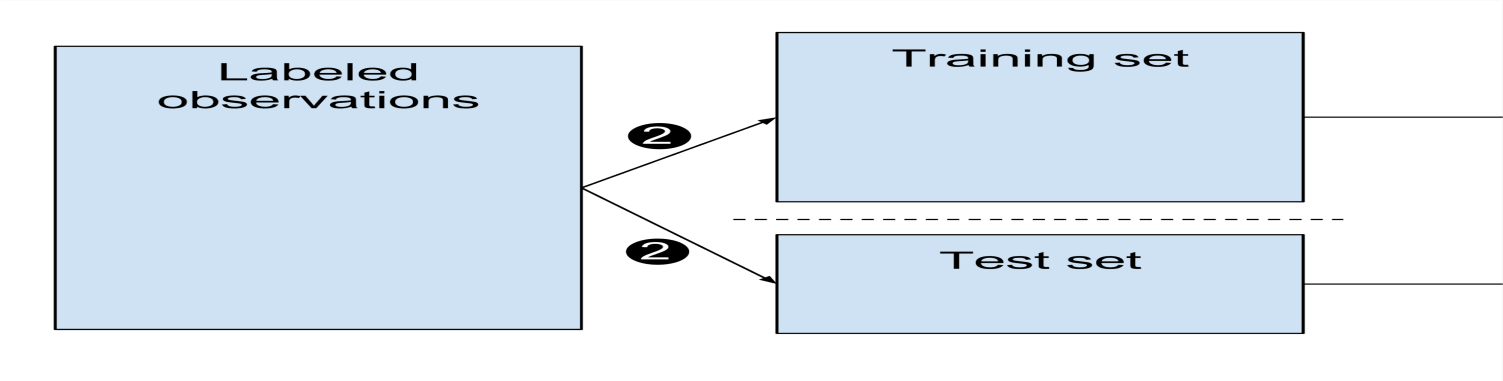
**Semi-Supervised:**

1. Dataset is a mixture of labelled and unlabelled data.
2. Existing algorithm unable to improve the results.

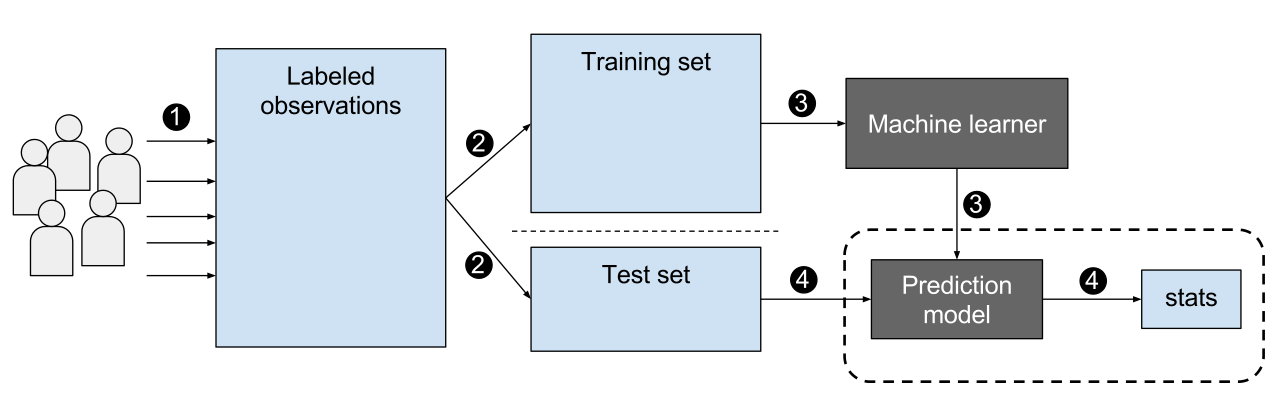
**2.SEMI-SUPERVISED ALGORITHM**

This class includes problem domains which consist of both labelled and unlabelled data. Programmer wants to develop a model which is trained upon on datasets which have very little labelled data and a large number of unlabelled data.

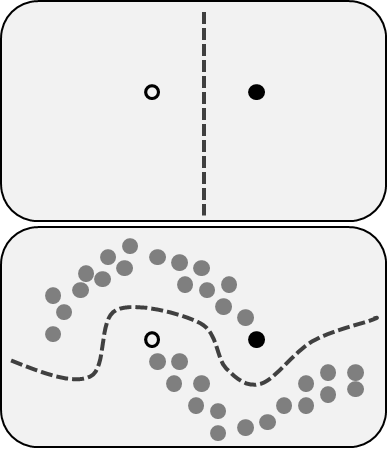
In the same note, this is the same as ” Teacher teaches a few concepts in class and gives questions as homework which are based on similar concepts.”

**3. Representation:**

**Supervised learning**



**Unsupervised learning**

****

**Semi-Supervised learning**

### 

**4. AlGORITHM:**

**Clustering for classification:**

| **Idea: Use naive Bayes on labelled example and then apply:**   1. Build naive Bayes model on labelled data. 2. Label unlabelled data based on class probabilities. 3. Train new naive Bayes model based on all data 4. Repeat 2nd and 3rd step until convergence. |
| --- |

**Semi-Supervised Assumptions:**

| **This algo is based on some assumptions:**   1. **Continuity Assumption:** The above assumption is based on the fact that points which are closer to each other are more likely to have the same output label. 2. **Cluster Assumption**: The data can be divided into discrete clusters and points in the same cluster are more likely to share an output label. 3. **Manifold Assumption:** The data lie approximately on a manifold of much lower dimension than the input space. This assumption allows the use of distance and densities which are defined on a manifold. |
| --- |

**5. Applications of semi-Supervised learning:**

Below are some fields where we can use semi-supervised algorithm:

1. **Speech Analysis:** Since labelling of audio files is a very intensive task,semi-supervised learning is a very natural approach to solve this problem.
2. **Internet content classification:** Labelling each webpage is an impractical and unfeasible process.
3. **Protein-Sequence classification:** Since DNA is typically very large in size, the risk of semi-supervised learning has been imminent in this field.

**References:**

1. <https://www.slideshare.net/lukastencer/semisupervised-learning-42075774>
2. <https://www.geeksforgeeks.org/ml-semi-supervised-learning/>
3. <https://link.springer.com/article/10.1007/s10994-019-05855-6>

eh webpage is an impractical and unfeasible process.**Pr**