Introduction to machine learning

* Supervised Learning
  + Learn an input to map output
    - Classification: categorical output
    - Regression: continuous output
* Unsupervised Learning
  + Discover patterns in the data
    - Clustering: cohesive grouping
    - Association: frequent cooccurrence
* Reinforcement Learning
  + Learning Control

**Machine Learning Task**

|  |  |
| --- | --- |
| Task | Measure |
| Classification | Classification error |
| Regression | Prediction error |
| Clustering | Scatter / purity |
| Associations | Support / Confidence |

**Challenges**

* How good is a model?
* How do I choose a model?
* Do I have enough data?
* Is the data of sufficient quality?
  + Errors in data (Eg; Age:225, or moise in low resolution images)
  + Missing values
* How confident can I be of the results?
* Am I describing the data correctly?
  + Are age and income feature enough? Or addition of other features like gender can giver better output.
  + How age can be represented? As a number or range like (young, middle, old…)

Supervised Learning

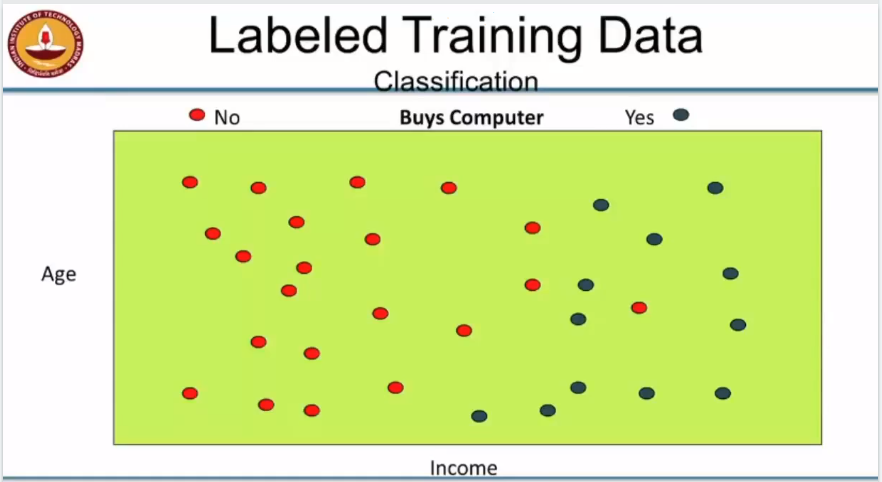
Let’s look for the classification problem in supervised learning. Here are two types of labels; one buying computer and other not buying computer.

Buy 🡪 Blue color

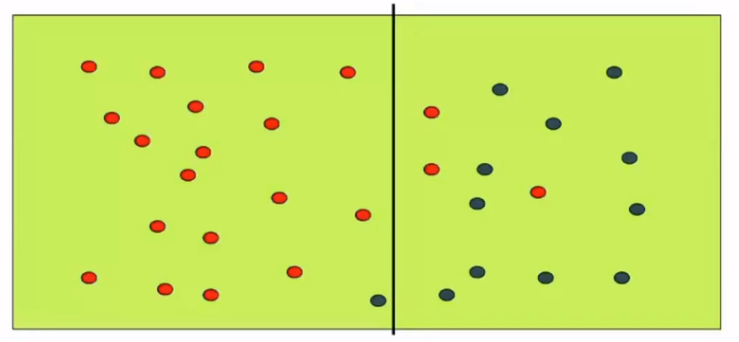
Not Buy 🡪 Red Color

X-axis shows income

Y-axis shows age



Simply classification lets see:

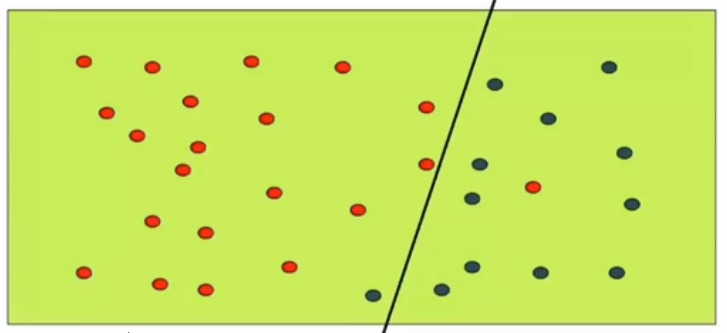


Here, vertical line is drawn to classify buying and not buying.

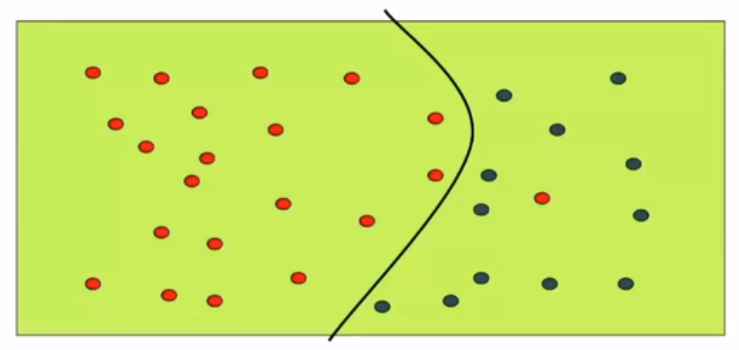
1 Buying classified as not buying

3 not buying are classified as buying

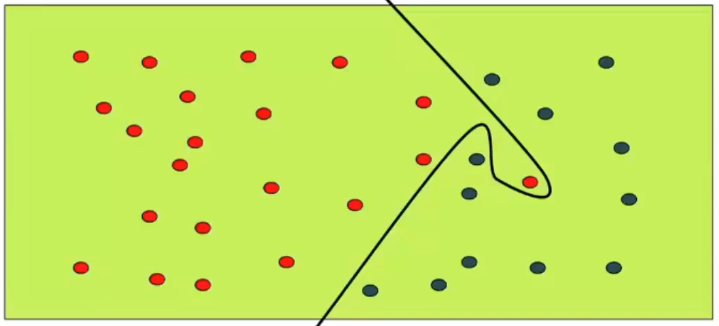
Analyzing this error, let’s make it better:



It is improvised by bending the straight. Now one error is still there



While explicitly trying to classify the data, it becomes so complex and it may also hamper further classification. This may misclassify some data. Therefore, the up to above curve line classification can be good choice.



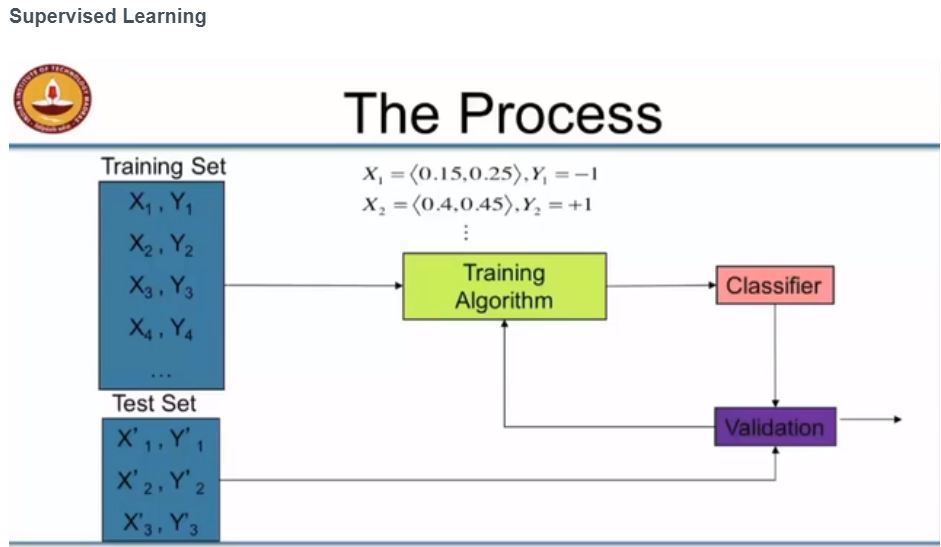
The red point at between is a noisy point which should have bought computer but some reason he/she did not buy.

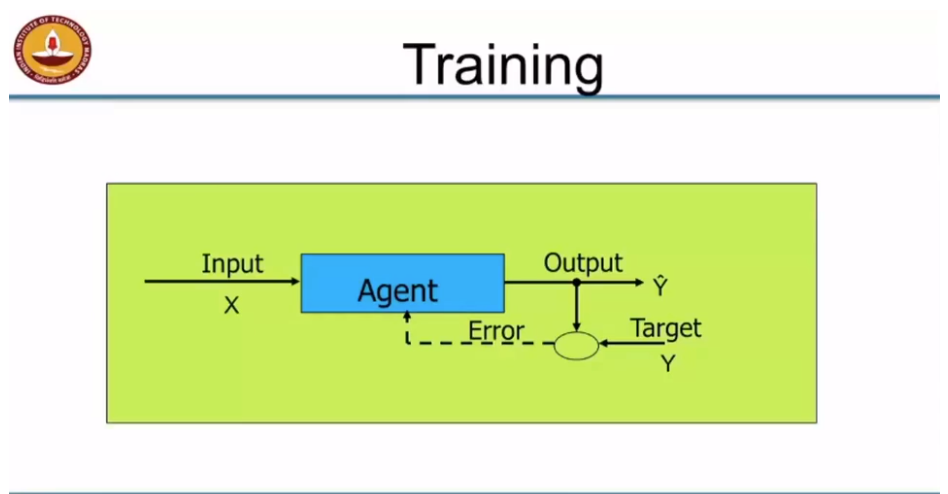
Parabolic curve is the best choice which makes the balance between the classifier and the accuracy on the training data.

Inductive Bias:

* Language bias (St. line on X or Y axis, or only parabola will be selected, ..)
* Search bias

The Process:





Application of Classification:

* Credit Card fraud detection
* Sentiment Analysis
* Churn prediction
* Medical diagnoses

Supervised Learning Architecture:

* Artificial Neural Network
* Support Vector Machine
* Decision Trees
* Set of Rules
* Nearest Neighbors
* Bayesian Networks

Linear Regression

* Minimize sum squared error
* With sufficient data simple enough
* With many dimensions, challenge is to avoid over fitting
  + Regularization
* Higher order function?
  + Basis transformations
  + Ex; (X1, X2) -> (X12,X22, X1, X2)

Application Area

* Time series predictions
  + Rainfall in a certain region
  + Spend on voice calls
* Classification
* Data reduction
* Trend analysis
  + Linear or exponential
* Risk factor analysis
  + Factors contributing most to output

Unsupervised Learning

Applications of clustering

* Customer Data
  + Discover classes of customers
* Image pixels
  + Discover regions
* Words
  + Synonyms
* Documents
  + Topics

