Class Notes: Basics of Machine Learning and NLP

IISC- CCE (Jan- May 2018)

**Date Notes Taken**: 10th Feb 2018

**Notes Taken by**: Nihar Kanungo

**For any corrections Please drop a mail to** [**Nihar**.kanungo@gmail.com](mailto:Nihar.kanungo@gmail.com)

**Define Machine Learning**

In general we can say that Machine Learning is a set of programs which continuously learn based on Experience

**Mitchell’s Definition**

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*

**E- Experience**: Based on the Data being collected

**T- Decision** to be taken based on the data

**P –** How to **Evaluate** the Decision

Machine Learning

Normal Program

Data

COMPUTE

Result

COMPUTE

Program

Data

Program

Output

Every Machine Learning typically involves three major steps

1. **Representation** of the Knowledge
2. **Evaluation** based on the rule on the data
3. **Optimization** of the Output

Let’s look at the graph. We found that the starting data points form a curve

And then it became straight. We have 5 data points at this time and based

On the data we are expected to predict the result of the 6th data point. So

Do we need to (i) Exclude the Highlighted one or (ii) Include it for training?

The Answer to this is based on the Perfection that we are assuming to see.

5 6

So that means if we are fine with some degree of approximation then we may ignore the initial ones which makes the graph almost straight and based on the same we need to predict the result of the future data points. However if we need a higher degree of approximation then we may not be able to ignore it. So that means for a given set of data there can be multiple Representation. We need to take all the representations and select the optimal one as per our need.

Machine Learning can be broadly classifies as two types of learning.

* + **Supervised Learning ( Point of desired output)**
  + **Unsupervised Learning**

Before we get into the above lets understand what’s called **Reinforced Learning**.

This is based on a Mechanism called Reward Mechanism. We assume to know the Way to learn given a set of data and use the Reward mechanism to improve the Optimization /Speed or Quality of learning.

So with the given set of data the final decision is being evaluated. Each Correct Decision is being rewarded and Wrong decision gets penalized. Weights are attached to each of the Correct and wrong result.

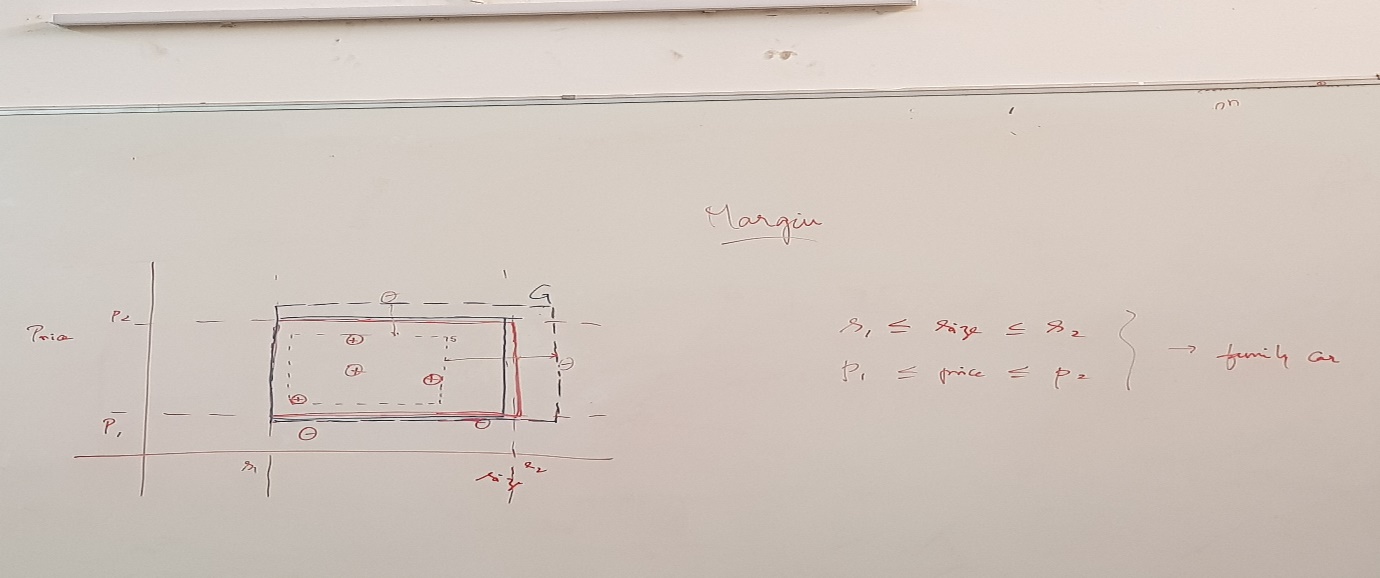
Finally the One with Maximum correct ones is chosen

**Supervised Learning**

**Classification**

Here for the sake of completeness we are taking the same example which is mentioned in our text book.

Let’s say we the data for a set of cars and our job is to find out if the car is **a Family Car** or **Not a Family car** based on the parameters given. For this example let’s assume that the parameters which decides the same are **Size of the Car** and the **Price of the car**



X axis represents the Size of the Car and Y axis represents the Price of the car.

The Points with + represents the Family Car and – says Not a Family car.

So for to satisfy as a Family car the Car should be within S1 and S2. Also the Price between P1 and P2.

Now we don’t know what that S1, S2 is and P1, P2. So we need to find out the best values for P1, P2, S1 and S2 from our training datasets in such a way that none of the training data and future data does fails. That is called the ideal one.

Now looking at the above picture

S - Space which exactly covers all the Positive cases (This is the Decision based on expert’s comments) – Tightest Fit

G – Space exactly after which the negative cases starts cases (Again this is the Decision based on expert’s comments) – Largest Area excluding all –ve

Now if any new data point comes which is in between S and G then what should be our Decision? Positive or Negative?

Now assuming that **h € H** (H – Overall Hypothesis and h – Our hypothesis)

SO our candidate hypothesis is the area between S & G

**Margin –** The Distance between S & G

Now if we take the one half of the distance between S & G on both the sides then that’s called our Candidate Hypothesis which is equidistance from +ve and –ve. This is the rectangular area as highlighted in red color in the above picture **(Area C)** .This is the Evaluation part.

Now if there is any data point which the Area S then the algorithm determines it as Positive but the expert conforms this as –ve. **This case is called as False Positive**

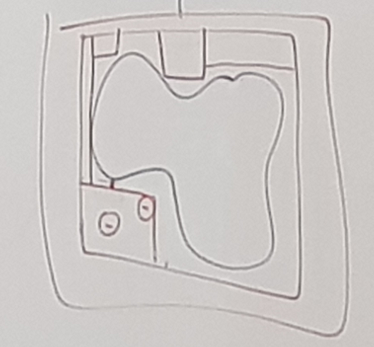
Similarly the case which is actually positive but algorithm determines as –ve being in that area is called **False Negative**

Now Area S, G and C changes many times based on the new data points. The Program optimizes itself by learning from the new data points till a scenario comes where all data points are clearly falls in their respective areas. That we call as the Optimized model.

However if the model finds a point which confuses the learning then Manual intervention of the experts are required.  
At times to come up to a best model the Extreme events are ignored for a large dataset. The larger the Dataset the better the program performs in real time.

The Highlighted one represents the scenario when decision is taken

based on two or more parameters and when it satisfies both

Even though the Rectangular model is the best and easy to determine but not always the data points can be determined using the rectangular form. The above picture shows one example which is again a classification one but not the rectangular one.

**Minimize the Error**

It’s obvious that the model with minimum error is the best one.

H

C

+

+

+

So we need to ensure that P (Error) <= ᵟ where ᵟ is the margin/Tolerance of error that would be acceptable for the scenario

Say **ệ** is the Probability of error.

Now from the above picture if a data point is outside of the area of C then there is a possibility of error **ệ**

So there are 4 rectangles outside of Area of C and within C and H

So possibility of error in each Rectangle is **ệ/4**

Hence Possibility of Correct answer is **(1- ệ/4)**

For N such data points it becomes **(1- ệ/4) ^ N**

Possibility that it misses all 4 rectangles is **4 \* {(1- ệ/4) ^ N)**

and the above should be less than the delta value . That means

**4 \* {(1- ệ/4) ^ N) <=** ᵟ

So by transforming we got the final equation to be

N >= 4/ **ệ log (**4/ **ệ)**

So that means if this satisfies then error will definitely be the minimum.

Read: **PAC Learning**: Probably Approximately Correct Learning

**Type of Errors**

* **Measurement Error** : This error occurs if measurement is not taken properly
* **Labeling Error** - If Expert gives the wrong answer
* **Unknown Parameter Error** - Error due to any unknown Parameter