Machine Learning CS60050
Instructor: Dr. Aritra Hazra

Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Scribed by: Ankita Saha, 20CS92R01
18 February 2021

1 Previously

The theoretical aspect of machine learning will be studied in this scribe. Previously we have learned how machine learning algorithms work. Taking into consideration a set of training data $[(x_1, y_1)\ldots (x_n, y_n)]$ and the designed algorithm we have a hypothesis set(H) which may be infinite. For example, in neural network algorithms we have to adjust the weights which are in continuous space. Therefore the number of hypothesis that a neural network can produce is infinite as it will take an infinite number of values for any of the weights from continuous real values. Those values converge into a final hypothesis $(g: x \rightarrow \hat{y})$ depending on which it can minimize the total error.

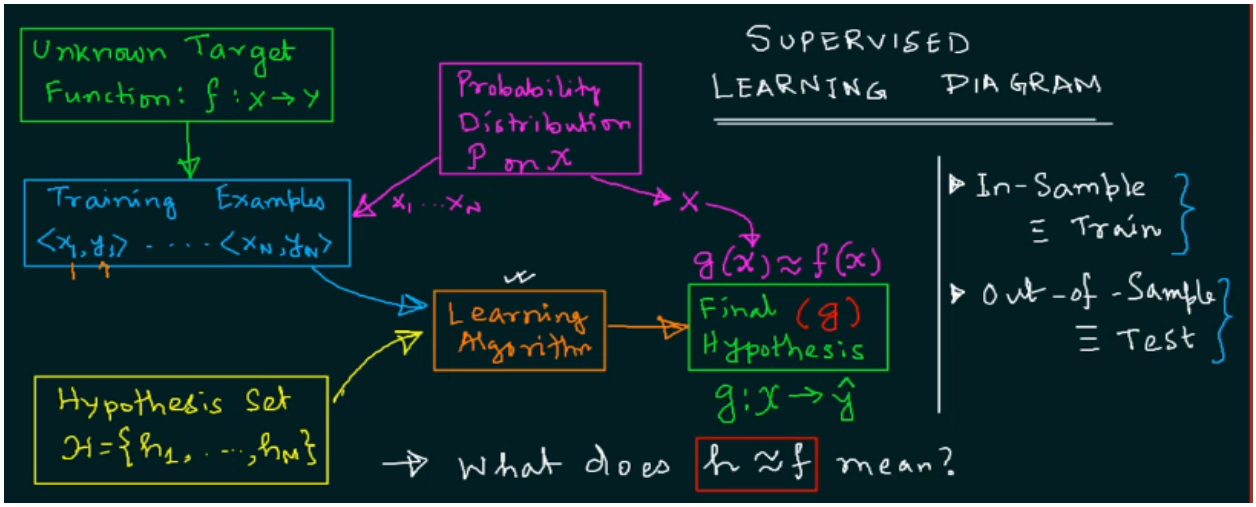

Figure1: Schematic representation of learning diagram

Many iterative solutions like stochastic gradient descent and batch gradient descent are studied which we utilize to minimize the error. The two other terminologies we mostly use are called in-sample which is given for training and out-of-sample which is given for testing. Interchangeably they are termed as training error and test error.

2 Learning from Data

2.1 Learning diagram

The learning diagram shown in Figure1 are described as follows:
(i) unknown target function ($f: x -> y$)which we want to learn,
(ii) set of training examples $[(x_1, y_1)... (x_n, y_n)]$ which are the simulation of unknown target function,
(iii) a hypothesis set (H) for the learning algorithm which tries to minimize the overall error,
(iv) function converges to find out the final hypothesis ($g: x -> \hat{y}$),
(v) probability distribution to help to learn the feasibility of training and testing instances.

2.2 Error

Error with respect to hypothesis over data x and an unknown target function, a pointwise error estimation is performed.

$$E(h, f) = average\ of\ pointwise\ error\ e(h(x), f(x))$$

We formulate error as
(i) squared error i.e. the square of the difference between the predicted and actual target variables,
(ii) binary error i.e. for two class binary problems stating either it is matching or not matching if not matching error is 1 otherwise 0.

Types of errors

In-sample error : In-sample errors are the error rate found from the training data, i.e., the data used to build predictive models.

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} e(h(x_n), f(x_n))$$

Out-of-sample error : Out-of-sample errors are the error rates found on a new data set, and are the most important since they represent the potential performance of a given predictive model on new and unseen data.

$$E_{out}(h) = Exp_x[e(h(x_n), f(x_n))]$$

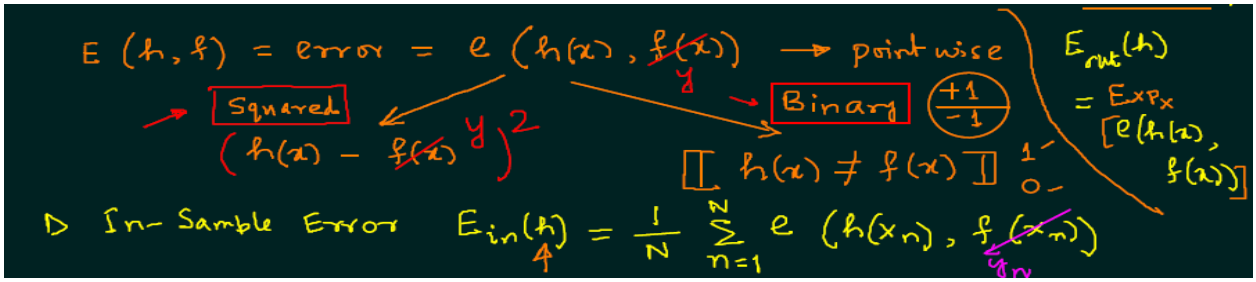

Figure2: Types of error

Now the question arises, how do we choose our error measures?
Let us understand by an example, say we want to build a fingerprint recognition device. The device takes the fingerprint of our hand and passes it through our classifier resulting in 1 (fingerprint to be passed), otherwise 0 (fingerprint not to be passed). In confusion matrix false accept (false positive) and false reject (false negative) should be penalised.

Why to penalise?
Suppose an normal employee is being rejected by the device but there is a chance of being allowed after a number of trials in such cases we penalise less.
But if an intruder is being allowed to enter i.e false negative case the algorithm should be penalised more as it may be dangerous.

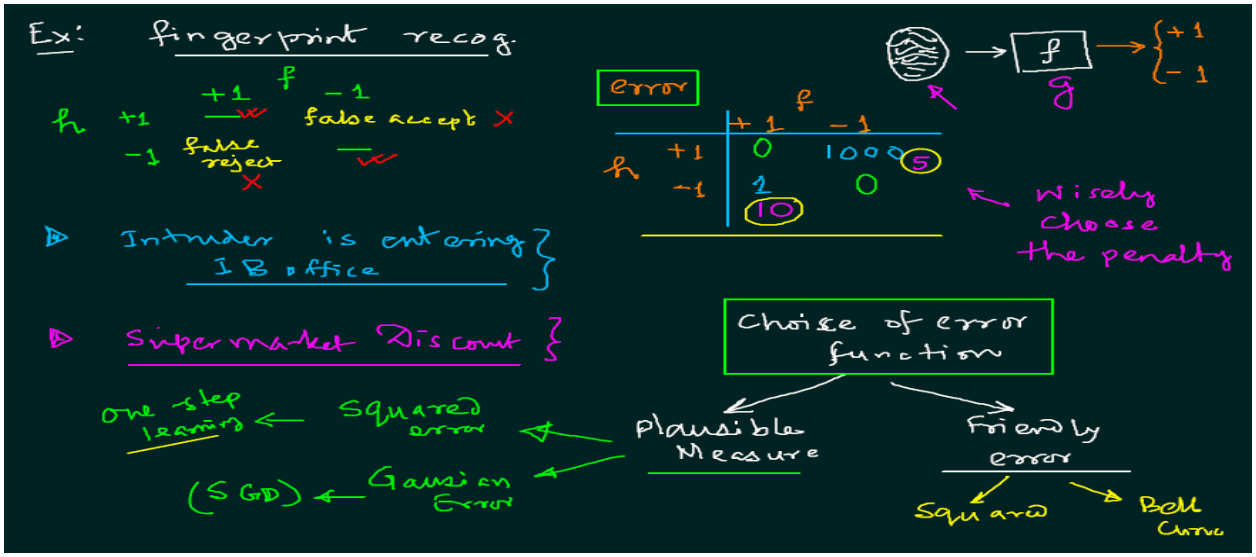

Figure 3: Example described to estimate the penalise rate

Another example, say a supermarket is offering a discounted rate to its daily customer. In such a case if the fingerprint matches but the customer does not get the discount there is a potential chance of losing the customer. Whereas a customer not allowed for a discount but the supermarket gave a discount it won't have much effect on the outcome.

So, firstly depending upon the type of application we should wisely choose the penalty. Secondly, the choice of error function is a plausible error (computationally easy) which helps in one step learning, friendly error (intuitively easy) measure where we use error measure or bell curve. These two govern the choice when we go for theoretical analysis.

2.3 Noise

Noise refers to the irrelevant information or randomness in a dataset. So when we try to learn a training data the unknown function we try to learn is no more a function because of its ambiguity. Example, for the same attribute in training data somebody says yes and somebody says no in result.

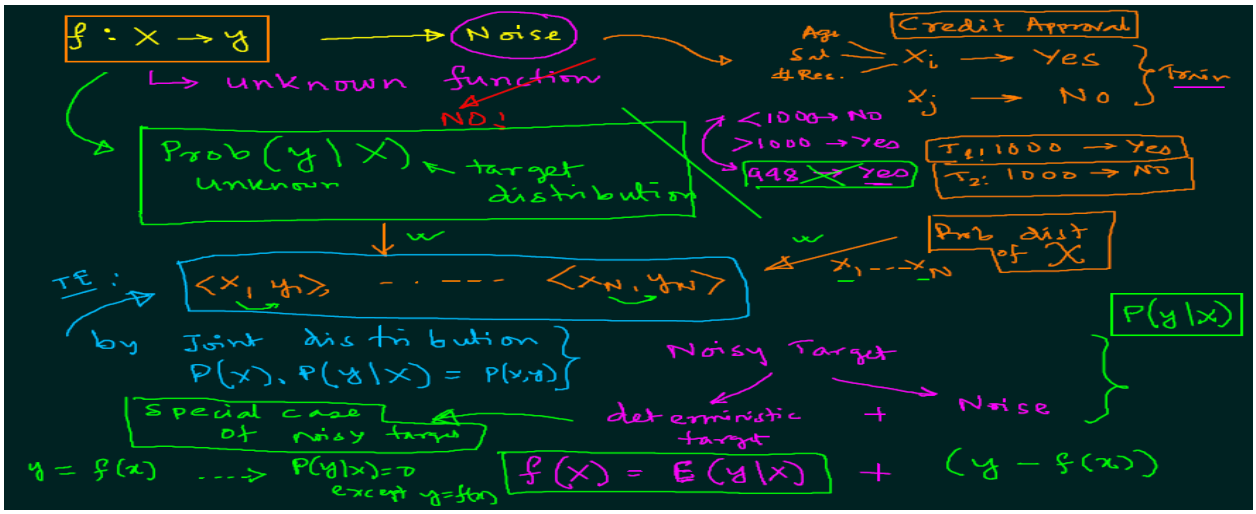

Figure 4: Noise estimation and target distribution

- Therefore we are trying to learn a target distribution rather than a function i.e. $Prob(y|x)$.
- We are generating the points ($< x_1, y_1 >, < x_2, y_2 >$..... $< x_n, y_n >$) by a joint distribution of $P(x), P(y|x) = p(x, y)$.
- Noisy target = deterministic target $f(x) = \mathbb{E}(y|x) + (y - f(x))$.
- Deterministic target is a special case of noisy target where,

$$P(y|x) = 0 \; except \; for \; y = f(x)$$

Note: The training data is not only generated from only the target distribution. The input distribution quantifies the relative importance of each of x i.e. the probability distribution on X. The modified learning diagram including noisy target is shown below in Figure 5:

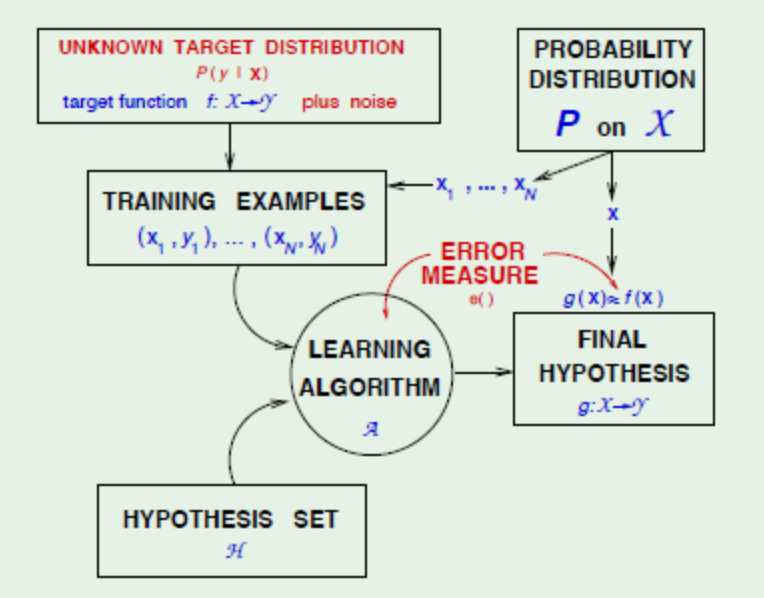

Figure 5: Learning diagram including noisy target

3 Preamble to the theory

- In inductive learning principle we assumed when $E_{out}(g) \approx E_{in}(g)$ (sufficiently approximate) and tried to prove by $P[|E_{out} - E_{in}| > \varepsilon] \leq 2Me^{-2\epsilon^2 N}$ which is feasible. Here M is the number of hypothesis and could be infinity. We try to minimize the error in E_{in} which follows E_{out}.
- We plot an error versus model complexity curve as shown in Figure 6. As model complexity increases the in-sample error decreases because all the samples fit too well. So the sample is increasing as it is reducing the chance of generalisation. The point beyond which the model have no chance of generalisation is termed as d_{vc}.

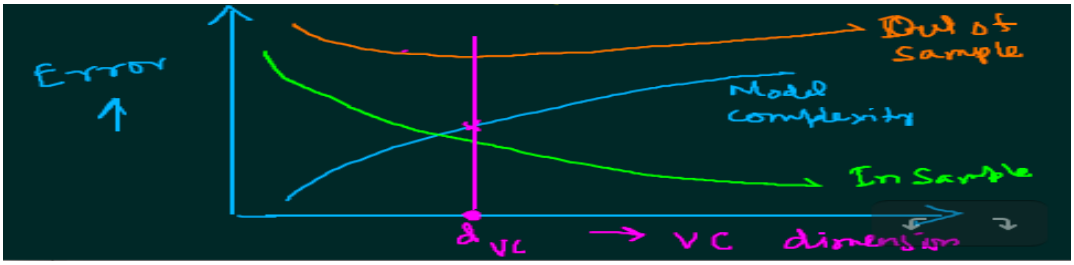

Figure 6: Plotting of error curve based on model complexity

- Suppose the goal is to learn the machine learning subject and some sample questions are provided. Based on this learning and certain estimation is performed to appear for the

test.

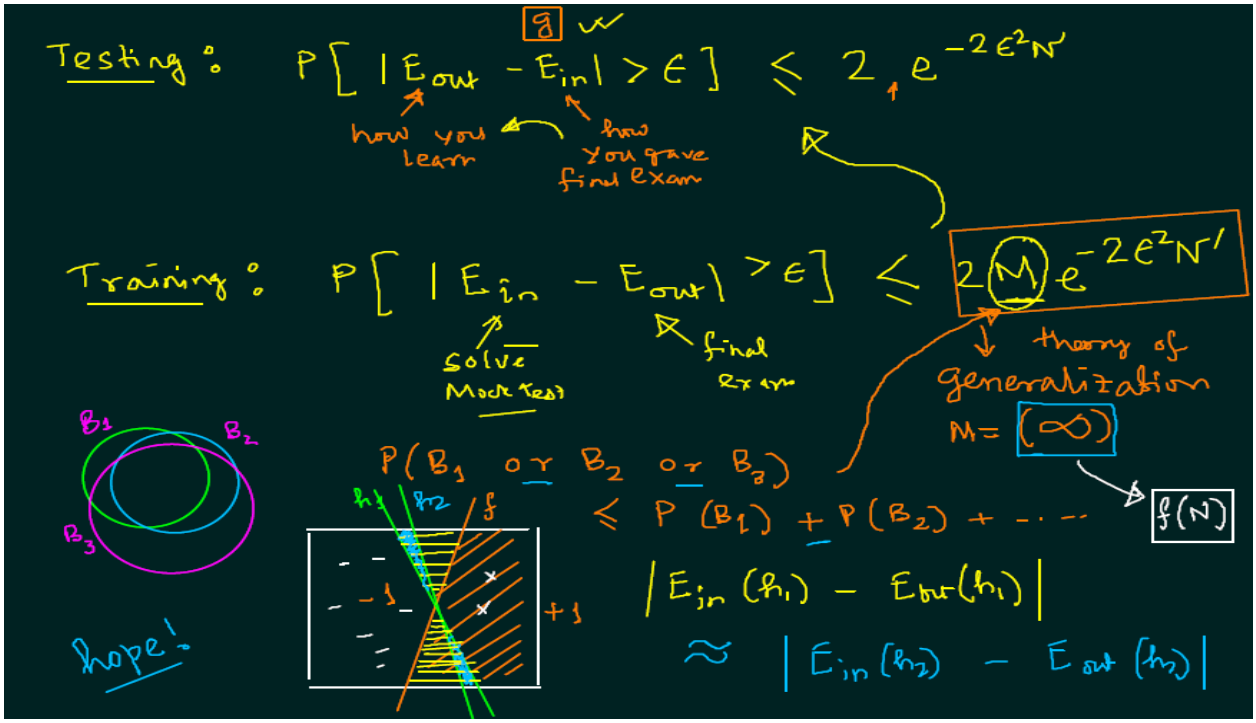

Figure 7: Intuitive estimation of significance of number of hypothesis

- Testing i.e the final exam E_{out} – *how you actually learned* and E_{in} – *how you perform in final exam*
- After getting training by learning and solving the solution it's time to appear for the exam. So we need to minimize the generalization of learning M. We get this M due to non overlapping bad events $P(B_1 \, or \, B_2 \, or \, B_3)$ with one pessimistic bound.
- In case of perceptron learning hypothesis $h_1 and \, h_2$ are overlapping with each other except the part mark in blue in Figure 7 plot. The interfacial area is so small that is $E_{in}(h_1) - E_{out}(h_1) \approx E_{in}(h_2) - E_{out}(h_2)$ is almost equal. Thus M does not have much effect upon perceptron concept.

[1]

[1] Reference:
1. This scribe is based on a lecture taught by Prof. Aritra Hazra on 18-feb-2021 in Machine Learning(CS60050) course.
2. All the figures in this document are taken from either handout-10a or Slide-10a uploaded on Machine Learning(CS60050) course website.