Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Machine Learning (CS60050)
Instructor : Aritra Hazra

Deepanshi Pandey : 20CS60R46 || Ashish Kumar Singh : 20CS60R48

26-March-2021

Core learning principles

The fundamental principles of learning which may be essential for designing learning algorithms for various ML problems. Ignoring them sometimes can lead to fitting hypothesis in a bad manner.
These include :

1. Simple Hypothesis :

A lower order polynomial(simple hypothesis) is used instead of a higher order polynomial if they can fit the data well. Out of sample performance would be good. Also it is better in terms of generalisation.

Occum Razor : From the data given, simplify the hypothesis till it fits the data well and becomes a good representative of the data.

What is **SIMPLE**?

Complexity of hypothesis(h) :

- Order of polynomial.
- Minimum description length(sometimes it is easier to represent large data using it's description).

Complexity of hypothesis space(H) : Mathematical notion of complexity of a model can be known using:

- VC dimension (VC dimension high : complicated in learning).

- Entropy based measures.

What is the link between chosen hypothesis and a hypothesis set?

The notion of simple also comes from how easy is it is to get a hypothesis out of a hypothesis set.

Suppose **h** takes **l** bits to represent.Therefore **h** is one among 2^l elements of **H**.

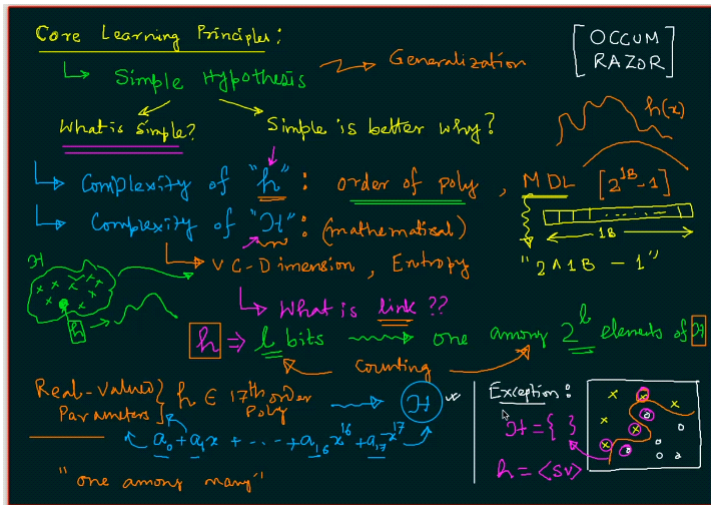
Counting mechanism to link: Simpler is the representation power of a chosen hypothesis, lesser is the amount of search space in the hypothesis set.

Example : Suppose h is a 17^{th} order polynomial,

$$a_0 + a_1x + a_2x^2 + ... + a_{16}x^{16} + a_{17}x^{17}$$

where $a_0, a_1, ...$ are real valued parameters.

The 17^{th} order polynomial is one among many hypothesis present in the hypothesis set.
An exception is a support vector machine, though it looks complex, the hypothesis is actually simple and is defined by a few support vectors.



Why is simple better ?

Puzzle : An astrologer everyday predicts about how your day is going to be. Suppose he gives correct prediction for 5 days consecutively free of cost. Now you are eager to know about how he does that. The 6th day the astrologer

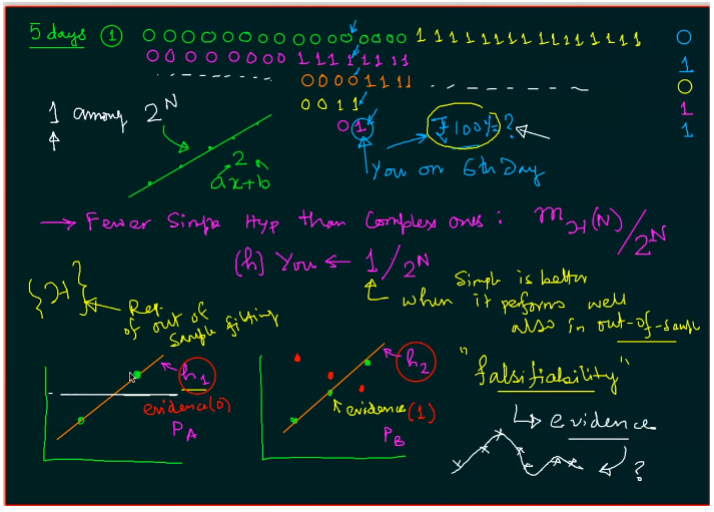
comes and asks you if you wanted to know the prediction of 6th day. But now he charges you Rs.100/- for the prediction.

What would you do?

The astrologer basically follows a deterministic approach. On the first day, he goes to 32 houses in the locality (including yours). He predicts a good day randomly for 16 houses and a bad day for others. He continues this trend for 5 days but with only those people for whom the predictions were correct for all of the previous days. You got all the predictions correct for the 5 days so he now asks you to pay him. So basically he doesn't predict, he only samples. You were one among many people. Hence you are very much interested in paying him for knowing the prediction.
Even though the hypothesis set is huge, you are tempted to go by a simple hypothesis without even considering the whole space. Therefore better or simpler is not elegant until the cost is being considered. This is the reason why the whole hypothesis set should be considered.

Fewer simple hypothesis than complex ones: $m_H(N)/2^N$.
The chosen hypothesis: $1/2^N$.
Therefore, simple hypothesis is better if it performs better for out of sample also. To know the out of-sample, the hypothesis set should also be known as it is the representation of out of sample fitting as well.

We also need to consider while fitting the data, how much evidence do we have to satisfy and falsify the data.
Therefore if we cannot falsify the given claim by the dataset we have, the learning is not justified.



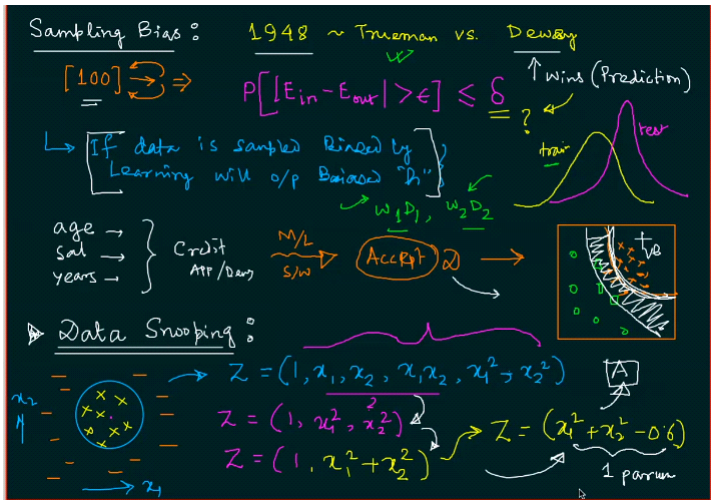
2. Sampling Bias :

Puzzle : In the 1948 US presidential elections, after the voting ended, a newspaper agency called(phone call) a section of people from all over US to predict who among Harry S. Truman and Thomas E. Dewey would win. They finally predicted that Dewey would win and printed it on their newspapers of the next day. But the final results announced Truman to be the winner.
What went wrong? Would it be better if the agency conducted the poll 100 times and take the average winner?
Mathematically, the in sample error should track the out of sample error within a confidence bound (δ)

$$\mathrm{P}(|E_{in} - E_{out}| > \epsilon) \leq \delta$$

We may think that δ is the problem. But actually there exists a bias here. In 1948, phone calls were only accessible to rich people and they preferred Dewey over Truman. Therefore the repetition over 100 people again would not change the prediction because of the sample bias that the 100 people only represent rich people and do not represent the entire population. Hence, if the data is sampled in a biased manner then the learning is also biased.

This can be removed by giving weights to individual data points. This way the bias is nullified.



3. Data Snooping :

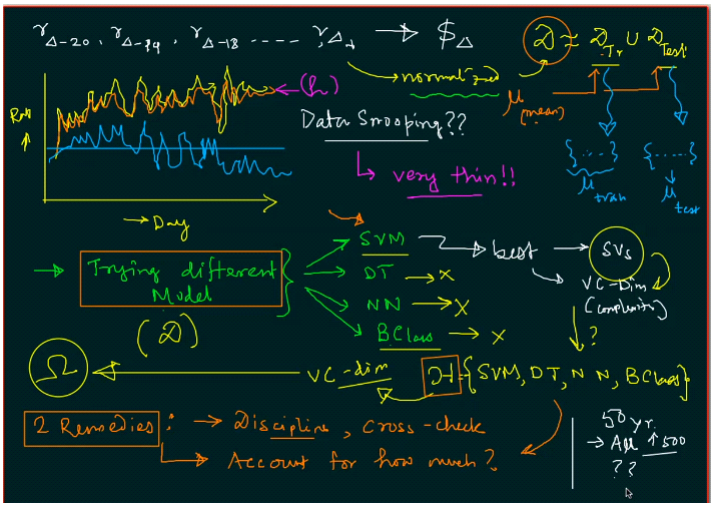
If data is being compromised during learning, then the outcome will also be compromised. This bias is because you have also looked into the data and found best fit i.e, you have done the work of the machine learning model. Therefore, *the outcome won't generalise well to the data that is not given.*
Example: Predict the rate of dollar vs rupee. While looking at the rupee value for last 20 days, we try to predict the dollar value for the next 4 days.

For this, we first normalise data (using the mean of the entire data i.e, test and train) and divide into test and train. After training a ML model, the results are very good for the test set. But when performed on the actual data, the performance was not good. Where did data snooping happen?

Data snooping happened when the data was normalised with one mean on both test and train, rather it should be normalised on two different means (one for test and train each). It is important to note that we need to check how much we compromise the data at each step.

Trying different models: Suppose we try on four models: SVM, DT, NN, Bayesian classifier and find out one best among them, suppose SVM. Now it should be considered that the hypothesis set is not only SVM but also DT, NN and Bayesian classifier.
Therefore, when you reuse a dataset, charge accordingly the penalty of the complexity of the dataset during generalisation as well.

PUZZLE: Suppose there is a stock market broking company which has 50 years data. It is a buy and hold share. You need to predict which company's share would be more profitable if we keep kit for 20 years. They collected data for top 500 companies over past 20 years and strictly followed buy and hold policy. But they did not make profit. Why?



Contribution:

Deepanshi Pandey: 20CS60R46 −− > Upto point 2:Sampling Bias.
Ashish Kumar Singh: 20CS60R48 −− > After point 2:Sampling Bias including point 2.