

Machine Learning (CS60050) Spring, 2020-2021

Instructor: Prof. Aritra Hazra

Scribed by:

Ram Kishor Yadav (20CS60R70)

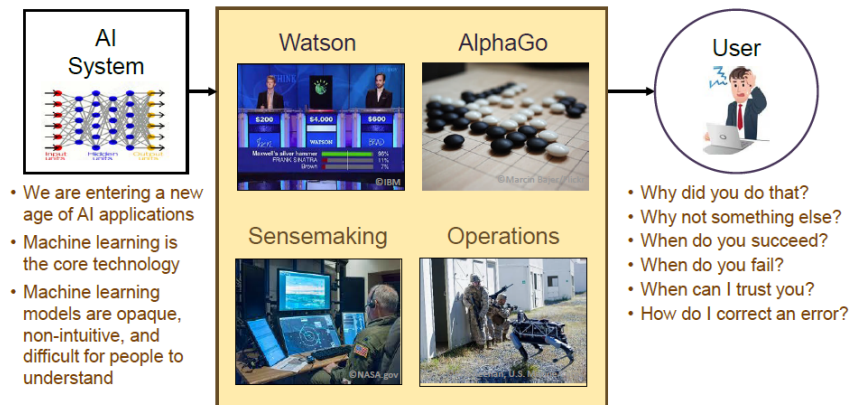
Rohit (20CS60R71)

Wednesday, 14th April 2021

Explainable and Interpretable Machine Learning:

When we gradually form the parameters to learn and keep on improving the number of parameters in a way to make our learning prediction more accurate, the primary point where we miss and gradually get more distant is the explainability or the interpretability factor. Meaning that now days with deep neural networks and all other kinds of techniques from very basic towards the complicated architecture that resulting as a model that is suitable to predict and interpolable wherever we want to use it. The problem of these models are they are very less explainable to the person or to the application where it is being used. Suppose we are using radiology images and trying to detect cancer and suppose our model tells that it's a cancer because your lambda value is 0.5. Patient and doctor nobody will be very much satisfy by that answer and also if a robot tells that I just landed on mars because I can see that ϵ is 0.5 and μ is 0.6 so I landed here. It is not worthy to use it because it cannot explain on this decision. Suppose, we build an autonomous car and it got an accident and if you could not explain why the accident happen it would not be helpful. The present day of research potentially diverted towards how can we make our model explainable or interpretable in that sense.

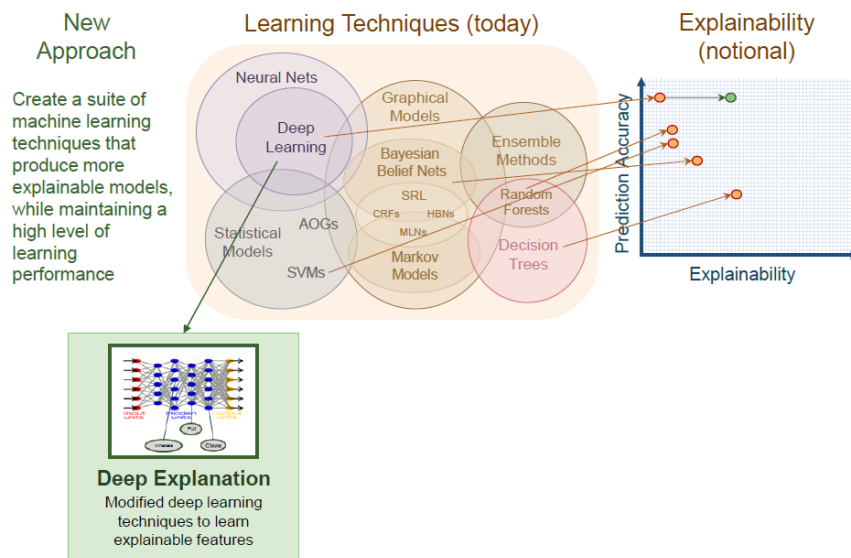
Explainable AI (XAI) – What are we trying to do?



Dramatic success in machine learning has led to an explosion of AI applications. Researchers have developed new AI capabilities for a wide variety of tasks. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems will be limited by the machine's inability to explain its thoughts and actions to human users. Explainable AI will be essential, if users are to understand, trust, and effectively manage this emerging generation of artificially intelligent partners.

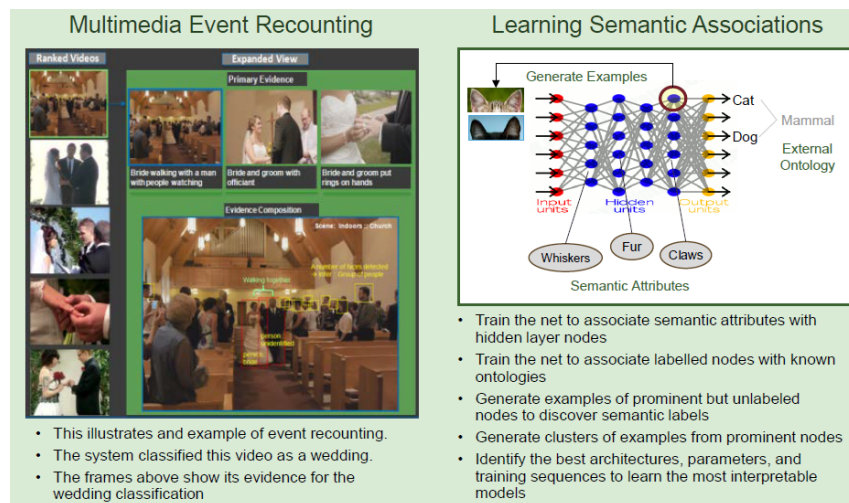
When we gradually move from AI to ML to DL, Our machine becomes much more accurate in prediction Like in Watson's quiz model, that can perform equally or even better than very famous quiz master. So, it won the quiz competition. The story of Alpha go, which is a reinforcement learning based solution to beat Go players in the go games and it basically a learning agent which plays between and gradually learns the moves and ultimately beats the best players. Also these kind of technologies are used in workforce and other sensor operations where the robot guides a particular troop of the police or the border forces to navigate there, collect information and give it. But on the flip side of it Whatever it does something goes wrong and something goes drastically very well. We should know why It made a correct prediction in case of failure. We must know that where it went wrong. In this concept we are lacking in the neural networks and other ML techniques, We only know the parametric value. Research is emerging to find that what is interim that a model is looking for to make decision. In the present day research, your learning process not just give a value or a parametric estimation of something to be of some object or some outcome we wish but also give why the outcome we want has been made by the model. So, it brings two things in the flow earlier we have a training data, a learning algorithm that produces an outcome model which can predict. Now we try to incorporate something as an explainable model and explainable interface because not only we should have a model, but we should also query that model so that we could get what is happening inside. And then we are happy with all the proceeding that we are getting is actually perfectly being done.

Explainable AI (XAI) - Performance versus Explainability



We have many methods like Bayesian belief nets, neural nets, ensemble method, decision trees, some statistical methods like SVM. There are three parts primarily, one is neural networks side, one is statistical side or extended version of it that is graphical models and other is decision tree which are mostly interpretable. Decision tree is much more explainable than neural network because decision tree has an if else structure so it is more explainable whereas neural network is just a parametric learning so it has higher accuracy because we can go deeper. So, prediction that is what we can achieve, is very much good, but the explainability, why it is doing so, is a question mark. So quest of recent time is that can we push this barrier? We have getting prediction accuracy higher and also explainability to certain acceptable margin. Researchers are trying to find the explanation from the hidden nodes of a neural network whether the hidden nodes can us some information about what they are looking at while they are composing the previous layer data and forwarding it to the next layer.

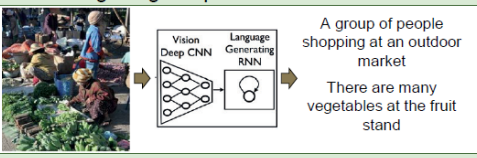
Learning deep Explanations



One of the first paper come out in Multimedia event detection and recounting paper (2014) in which they train with an image with a little bit of pretraining what is in the image. Suppose in the above image if you are exchanging a ring then it must be a wedding event. If there is bride, groom and someone who is officiating the event then that event must be a wedding event. This kind of some training happened and finally they try to learn semantic association in between. For example, for a cat if it try to find fur, the claws of a cat then how this network forwards the probability of claws into the cat as an output. So this is one of the previous approach, they train it with some supervise set of what is happening if we try to match that it try to figure out from intermediate node some associations out of it.

Learning to Generate Explanations

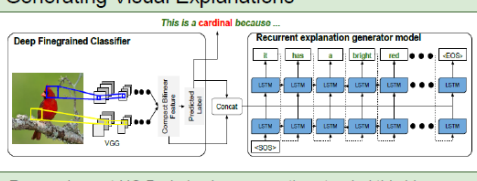
Generating Image Captions



A group of people shopping at an outdoor market
There are many vegetables at the fruit stand

- A CNN is trained to recognize objects in images
- A language generating RNN is trained to translate features of the CNN into words and captions.

Generating Visual Explanations




This is a cardinal because ...

Researchers at UC Berkeley have recently extended this idea to generate explanations of bird classifications. The system learns to:

- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)

Example Explanations



This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

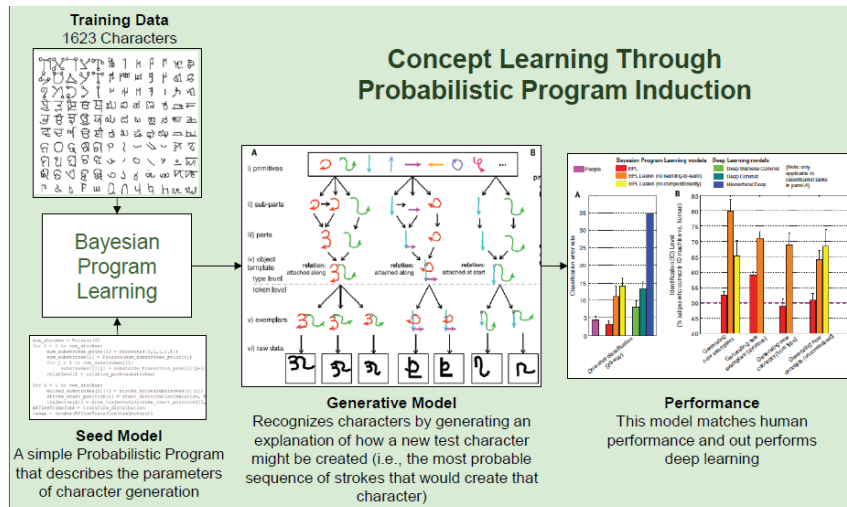
Limitations

- Limited (indirect at best) explanation of internal logic
- Limited utility for understanding classification errors

In era of deep learning, generating captions from images is a kind of explainability measures that we can provide. Suppose we look into the above image and predict something. We can also give some explanations, there are vegetable fruits in the market. A group of people selling, a group of people shopping in the market, that give us a better description why we categorize this image as a market image. This is not difficult because we know that CNN can identify objects and RNN can text the description from the object. This approach has been done in Generating visual explanations paper (2016) that generated visual explanations just by having the concatenation of the CNN structure with the RNN structure. Where RNN in particular uses LSTM because LSTM is sometime needed to remember some information from the path and to forget some information in order to generate a meaningful caption. So, this is one of the information people try to generate from visual data and they are quite successful in generating what is happening inside it. This is through the deep explanation generation.

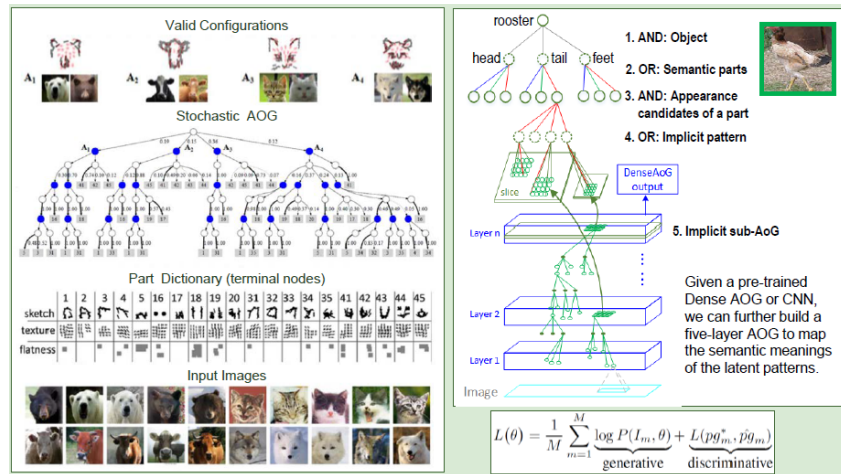
Another thing is that people try to introduce another model on top of it. Suppose we learn a model, I will also give another model as an auxiliary model whatever we learned with it. So, that you could question the auxiliary model and get interpretable answers. This auxiliary model often sometimes be an AND-OR graph. It may be hybrid Bayesian net where the probability value as well as the path can dictate you about the decision you are making.

Learning more interpretable models



Suppose we have a set of characters as a training data. We are trying to train something. Here we generate from the subparts and keep on generating different orientations, different variations of the data. Whatever the way we write such generative model will aid you and dictate why it categorize as a particular letter in your native language. So, we have the seed model which is the oriented alphabet. From there we generate certain different variations of it. These kind of things are associated with whatever we learn can help us why we see that this character is me. Somebody could write it straight forward, somebody could write it in curly way. So, this also can give you a little bit of more explanation and they called it as concept learning through probabilistic program induction. There is a very famous paper “Human level concept learning through probabilistic program induction”, in which they try to imbibe human understanding of concepts using this kind of model.

Stochastic AND-OR Graphs



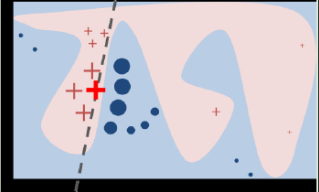

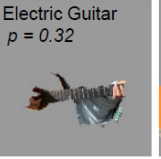

One of the generalizations is the AND OR graphs (AOG). There not only particular structure we do from probability model but we take an AND-OR graph to make a decision objective presentation. Suppose we want to identify a rooster. Our AND node will be whether the head of rooster looks like something which resembles a rooster AND its tail resemble a rooster AND its feet resembles a rooster. Our OR node will be its head could be 90° tilted OR its head could be same orientation OR its head could be 180° tilted. When we go deep down to the AND OR graph we gradually focusing a part of it and try to take down what are the saline characteristics in terms of variations as in composition of output and in terms of possibilities which I composite and thereby if I supplement this auxiliary model along with our model of prediction then when we predict something we traverse some path and we will say this head is 90° rotated but it has a red mark on it and finally we get decision tree. So, At the bottommost layer we will have some textures, boundaries then a little above that we compose that boundary to make a fish like structure, ear like structure and eye like structure then we come up again then we find the face, the tails and all that and gradually this models help us by going through a path along with the decision making we make to explain why that decision has been made.

We can also have supplementary interpretable models along with it not only directly putting inside our original prediction model. Sometime our model also come as a black box and then it is called model induction. It is a technique to infer an explainable model from our black box. So, you do not know what the parametric values has been set. You can only see it as a black of your model. You can give input, it predicts some output, but during experimentation can you

infer certain explainability out of what it predicts. There are many approaches going after it. One of the first famous approach is LIME.

Model induction

Local Interpretable Model-agnostic Explanations (LIME)

Black-box Induction	Example Explanation		
 <p style="font-size: small;">The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.</p>	 <p style="font-size: x-small; text-align: center;">(a) Original Image</p>	<p style="font-size: small;">Electric Guitar $p = 0.32$</p> 	<p style="font-size: small;">Acoustic Guitar $p = 0.24$</p>  <p style="font-size: x-small; text-align: center;">(c) Explaining Acoustic guitar</p>
<ul style="list-style-type: none"> • LIME is an algorithm that can explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model. • SP-LIME is a method that selects a set of representative instances with explanations as a way to characterize the entire model. 			

LIME tries to give new new images and suppose in the image red region is the plus region and green region is the minus region. I do not know how do I categorize in the model because model came to me as a black box I can only experiment with it with an image. Now the near boundary that it find with the higher probability it tries to see which is the plus one and which is the minus one. Let's say it get a image of this and I want to find out what kind of guitar it is. There are 2-3 approaches that immediately follow from this . One is saliency guided , that I can see that saliency map of from the decision being made to the black box model. I need to know the saliency how it is looking into and you can see the way it look into the probability value. This LIME is one of the first approach that it starts with the black box model with arbitrarily try to predict the boundary and try to give confidences of positive and negative sample points and the far away the points, it give less confidence about whether the positive thing properly classified or not. But this has been extended to multiple facet often sacrificing the black box model to a white box model and also often extending into SPI-LIME. But “why should I trust you? Explaining the prediction of any classifier” paper is one of the pioneer paper on black box model.

Bayesian rule list (BRL)



Model Induction

Bayesian Rule Lists (BRL)

- **if** hemiplegia and age > 60
 - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
 - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
 - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
 - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
 - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
 - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)

Clock Drawing Test

Best Testing AUC vs. Model Category for Screening Tasks

Model Category	Testing AUC
All All Features	~0.92
All Classifier Features	~0.88
BRL All Features	~0.85
All Synthetic Features	~0.82
SVM All Features	~0.81
RF Classifier Features	~0.80
RF Synthetic Features	~0.78
SVM Classifier Features	~0.75
SVM Classifier Synthetic	~0.72
Operationalization	~0.70

- BRLs are decision lists—a series of if-then statements
- BRLs discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements.
- Experiments show that BRLs have predictive accuracy on par with the current top ML algorithms (approx. 85-90% as effective) but with models that are much more interpretable

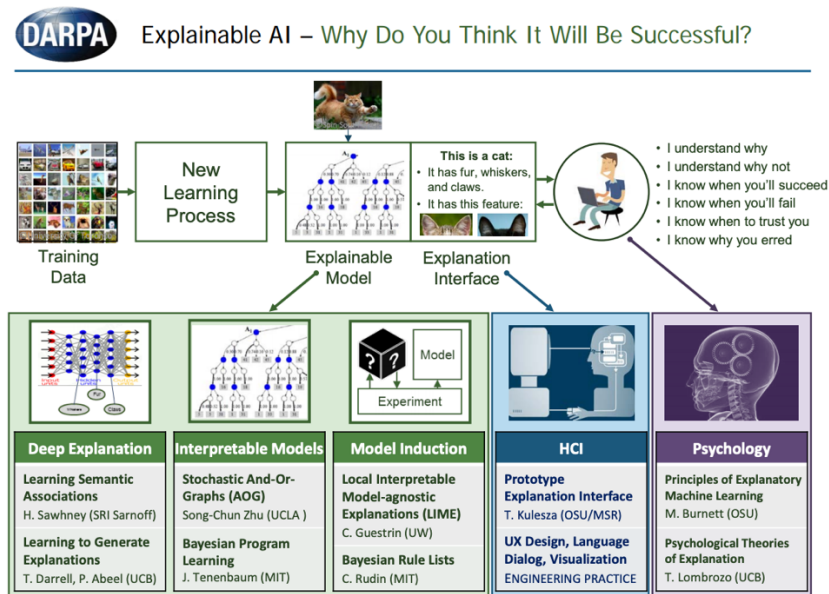
Letham, B., Rudin, C., McCormick, T., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* 2015, Vol. 9, No. 3, 1350-137

There has been ramification of LIME model in terms of different types of psychological and cognitive science areas as well because sometimes how a person is drawing or behaving certain things, it dictate some cognitive pairment as well. So, if I have a model for detecting such kinds of cognitive impaired or some decenties then I can use which portion of it or which is the boundary line I am highly confident that we have a cognitive impairment and some person do not has cognitive impairment. “Interpretable classifier using rules and Bayesian analysis” paper again just extends the line and apply it in mefical and cognitive psychological domain. So, that the explainability whether a person has cognitive impairment I can also get it by looking at how can see whatever you are doing. Sometimes not representing or just hiding a part of model also can dictate which part of model is more significant. So, sometime people just try to hide a portion of an image and try to see whether the outcome of image dictated similarly if we hide. Suppose I am trying to identify guitar if I hide the face the outcome will not matter but if I hide the string the outcome will matter. Without this string you would not figure out this thing is guitar or not. So, that dictates that which part of the neural network is primarily sensing which part to make a decision. So, people have experimented all these and there are papers which often exploits saliency guided maps, blurring of the image seeing blurr part does it matter?, changing the foreground background ,does it matter?

Explainable AI – Why Do You Think It Will Be Successful?

In an explainable model part as we have seen that we have incorporated some of the explanation in model. we provide an additional model to provide explanation. We query Blackbox and sometime we can infer certain decision making. And little bit advanced techniques that as discussed that align the guided techniques that we can use

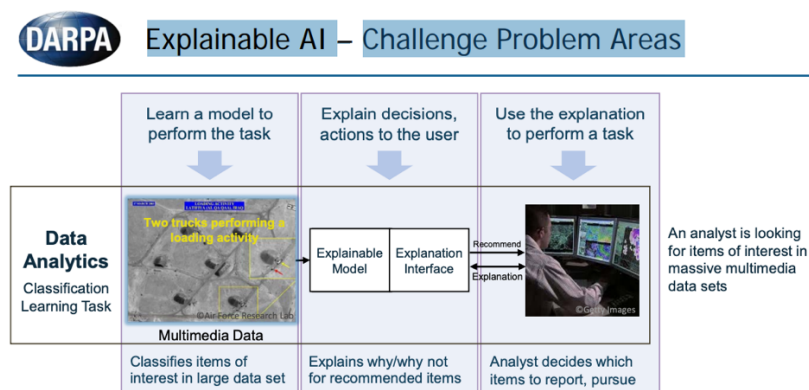
But of similar importance will be of explainable interface as well because not only you supply a model or you equip the model to make some analysis for you but how you represent analysis to user who is using it. So that he could refine it or he can refine his decision-making and all is very important. This is primarily more important when you go into safety critical domain in devices or in case of medical domain or in psychological domain because some learning agent is trying to dictate a doctor about or help doctor about a operation. Obviously, that agent should mention in such a way that is interpretable to the doctor and doctor should communicate with the agent in order to make the decision. Therefore equally important is also the explanation interface, which is often we, call as human computer interface.



This particular thing has also guided certain new research arena now since machine deployed widely in many practical scenarios they are predicting the low outcome in a port they are predicating the medical outcomes and all. So they should fair and unbiased.

Explainable AI – Challenge Problem Areas

Primarily it is deployed in many data analytics platform where people in multimedia in medical image analysis try to recommend and try to for an explanation of it. And some time it is also applied autonomy environment forces where there are drones can say that which region is more attack prone and which region is less attack prone and accordingly guide your troops of soldiers to that region. So it need to explain if you guide your soldiers to certain region and where there is bombing then agent need to explain that why it is taking decision to misguide some soldiers in a place. This should be explainable.



This should me explainable that we still yet not at the position where we should be in terms of machine learning explainability.

The path we travelled: ROADMAP

Theory:

- VC- dimension: it gives us opportunity to think that learning is possible. it is question of possibility verses probability we can learn and DC-dimension give rise that probably, approximately correct ways of learning
- Bias- Variance: How good we can search through and find our hypothesis.
- Complexity
- Bayesian theory

Techniques:

1. Models:
 - linear

- Artificial Neural Network
- Support Vector Machine
- K- nearest neighbours
- Radial basis function
- Gaussian process
- Graphical models(Bayesian net)
- Decision tree
- concept learning: it is just a Boolean formula

2. Methods:

- Aggregation (ensemble technique)
 - Bagging
 - boosting (adaboost)
- Regularization
- Validation
- Input processing(Transform)

Paradigms:

- Supervised: You learned from experience which is labelled
- Unsupervised: Often used to discover something new when you don't have labels
- Reinforcement: This paradigm has laid us to find out that we don't have immediate labels like supervised and it is better than unsupervised because some trainer is there he just tells you that you shouldn't go there as it is not rewarding something like that
- Neural: It is also a paradigm nowadays because it just invented or reinvented the *DEEP LEARNING* aspects.