Machine Learning

Aniruddh Arora
IIT2019003

(1) The fourth industrial revolution or Industry 4.0 is referred to as the automation of industrial process using technologies like AI, ML and other modern day techniques. The fourth Industrial revolution started in the early 21st Century [late 2000s].

The relationship among AI, ML, DL can be defined as $DL \subset ML \subset AI$

(is subset of)

In Artificial Intelligence, Turing test is a test designed to check if a machine or software is truly intelligent or not. It comprises of an evaluator, a human participant and a machine participant. The evaluator pitches in a question to both the participants and then

it tries to evaluate whether the machine is intelligent or not on the basis of how perfectly the machine is capable of imitating the human participant.
The turing test was proposed by a British Computer Scientist Alan Turing in 1943.
Sofar, a chatbot named 'Eugene Goostman' have passed the Turing test in 2014.

---

(2) The complex code breaking technique developed by Alan Turing was known as 'Turingery'. It involved the study of logical structures of the encrypted codes and also the periodic rotation patterns that could be seen in the code.
The term Artificial Intelligence was coined by John McCarthy in 1956.

---

(3) GPT-3 is an abbreviation for Third Generation Generative Pre Trained Transformer. It is a machine learning model that could generate any type of text.

Aniruddh Arora
IIT2019003

(5) Structured data: The collection of data which is clearly defined, easy to search and manipulate is known as structured data. It is defined using several parameters and can only exist in the format that is pre defined. The clearly defined structure makes it easy to ~~be~~ searched and analyze the data.

example: The SQL databases stores the data in a structured manner with several set of keys which makes it easy to handle and analyze it.

Unstructured data: The data stored in its raw or native form is defined as unstructured data. The format in which unstructured data is not predefined and it is very easy to handle and analyze this data.

example: The data from internet searches, the chatting services etc is unstructured data.

(6) AI can help in the healthcare Industry:

- By Training a suitable AI model, we can use it to predict whether there is a chance of having a disease or not. for example: Using ML, we can predict whether or not the Tumour shown in the image is cancerous or not.
- Analysing a large number of images/cases might be very time consuming and tiresome for humans. These kind of Big Data Analysis can be easily done in mere seconds using the AI.
- AI can assist healthcare professionals in making decisions that can even save a person from dying.
- AI robots can successfully carry out very complex, sophisticated and long surgeries with ease and minimal chance of making mistakes.
- AI can help synthesizing artificial life support devices by mimicking the natural organ functioning.
- AI can help in analyzing the molecular structure of various life saving drugs and can help in finding their cheaper alternatives.

Aniruddh Arora
IIT2019003

(8) $$h_w(x) = w_0 x_0 + w_1 x_1 \quad \text{---(1)}$$

$$J(w) = \sum_{i=1}^{m} \frac{1}{2m} \left(h_w(x_i) - y_i \right)^2 \quad \text{---(2)}$$

Y = original Values

y^* = predicted Values

$$\bar{Y}^* = \bar{X}\bar{W}$$

$$\begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_m^* \end{bmatrix}_{m\times 1} = \begin{bmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ \vdots & \vdots \\ x_{m0} & x_{m1} \end{bmatrix}_{m\times 2} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}_{2\times 1} \rightarrow W$$

$$\text{Error} = \bar{Y} - \bar{X}\bar{W} \quad [\bar{Y} - \bar{y}^*]$$

$$E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} = \begin{bmatrix} y_1 - (w_0 x_{10} + w_1 x_{11}) \\ y_2 - (w_0 x_{20} + w_1 x_{21}) \\ \vdots \\ y_m - (w_0 x_{m0} + w_1 x_{m1}) \end{bmatrix}$$

$$\text{Squared error} = \begin{bmatrix} e_1^2 \\ e_2^2 \\ \vdots \\ e_m^2 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \underbrace{\begin{bmatrix} e_1 & e_2 & \cdots & e_m \end{bmatrix}}_{1\times m}$$

Squared error $= e_1^2 + e_2^2 + e_3^2 + \_\_\_\_ + e_m^2 = L$ (say)

$$L = \underset{1\times m}{[e_1\ e_2\ \cdots\ e_m]} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}_{m\times 1}$$

$$L = E^T \cdot E = (Y - XW)^T (Y - XW)$$

using properties of transpose $\left[(A+B)^T = A^T + B^T,\ (AB)^T = B^T A^T\right]$

$$L = (Y^T - W^T X^T)(Y - XW)$$

$$L = Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW$$

Now differentiating wrt W_i $i = 0, 1$

$$\nabla_w L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \end{bmatrix}$$

to minimize the squared error put it equal to 0

$$\begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \end{bmatrix} = 0$$

now, $\nabla_w (W^T a) = \nabla_w (a^T w) = a$

so,

$$\nabla_w (L) = \nabla_w (Y^T Y) - \nabla_w (Y^T XW) - \nabla_w (W^T X^T Y) + \nabla_w (W^T X^T XW)$$

$$0 \quad - X^T Y \quad - X^T Y \quad + X^T XW \quad + (W^T X^T X)^T = 0$$

$$\Rightarrow -2X^T Y + X^T XW + X^T XW = 0$$

$$\Rightarrow X^T Y = X^T XW \Rightarrow W = (X^T X)^{-1} X^T Y$$

Aniudh Arora
IIT2019003

using the derived normal eqn.

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \left[X^T X\right]^{-1} X^T Y$$

$$X = \begin{bmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ \vdots & \vdots \\ x_{m0} & x_{m1} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

---

(9) $H_w(x) = W_0 X_0 + W_1 X_1 + W_2 X_1^2$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad X = \begin{bmatrix} x_{10} & x_{11} & x_{12} \\ x_{20} & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ x_{m0} & x_{m1} & x_{m2} \end{bmatrix} \qquad \left[x_{i2} = (x_{i1})^2\right] \forall i = 1, 2, \ldots$$

$$W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

$$E = \begin{bmatrix} y_1 - (w_0 x_{10} + w_1 x_{11} + w_2 x_{12}) \\ \vdots \\ y_m - (w_0 x_{m0} + w_1 x_{m1} + w_2 x_{m2}) \end{bmatrix}$$

Total squared error, $L = E^T E$

$$= (Y - XW)^T (Y - XW)$$

$$L = (Y^T - W^T X^T)(Y - XW)$$

$$L = Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW$$

$$\nabla_w(L) = 0$$

$$\Rightarrow \nabla_w(Y^T Y) - \nabla_w(Y^T XW) - \nabla_w(W^T X^T Y) + \nabla_w(W^T X^T XW) =$$

$$0 - X^T Y - X^T Y + X^T XW + (W^T X^T X)^T = 0$$

$$\Rightarrow -2X^T Y + X^T X W + X^T X W = 0$$

$$\Rightarrow X^T X W = X^T Y$$

$$\Rightarrow W = (X^T X)^{-1} X^T Y$$

$$W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \quad X = \begin{bmatrix} x_{10} & x_{11} & x_{12} \\ \vdots & & \\ x_{m0} & x_{m1} & x_{m2} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

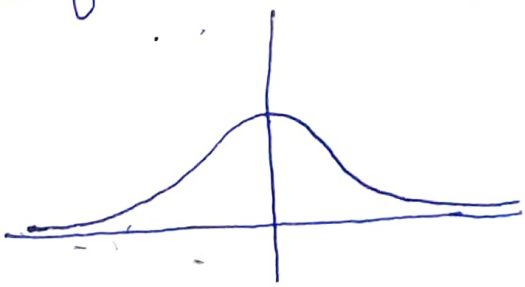
here $x_{i2} = (x_{i1})^2 \quad \forall i = 1, 2, \ldots, m$

---

(11) Locally weighted regression is a special type of Linear regression which focuses more on the near neighbourhood of the point of interest than the far neighbourhood of the point.

We can use an exponential function to define a weighted function.

$$h_i(x) = e^{-k(x - x_i)^2}$$



analyzing the graphical nature, the near neighbourhood will be given more preference than the far neighbourhood known as weights.

here k is known as smoothening parameter.

Bandwidth allows us to define the spread of the neighbourhood and help us limit the weights given to the elements in the neighbourhood of the focus point.

(7) Autonomous System: A system is known to be an autonomous system when it is capable of doing certain activities or tasks as it is designed to do. The designer to pre-code all the possible scenarios that could occur and autonomous machine could handle such predicted scenarios without fail.

Intelligent System: Intelligent systems can be defined as systems that could make a decision while improvising with the actual scenarios. These systems can decide what to do in scenarios that are not predicted already. The systems try to decide the best possible method to handle such unpredicted scenarios.

4 different types of learning are:

- Supervised learning: The classified/structured data is fed to the model in order to set the parameters that are used to predict the outcome on unknown values.
- unsupervised learning: The unclassified/unstructured data is fed to the model for training purposes the machine itself has to classify the fed data.
- Semi supervised: A portion of supervised learning and unsupervised learning is used.
- Reinforcement learning: The observations through environmental interactions are used to reinforce the trained model to more accurately predict the unknown values

Steps to Build a ML model:

- Data Collection: data is needed to be collected for training and checking purposes.
- data arrangemat: data should be arranged in a suitable order in to make it suitable for training.
- Model type: different types of models can be used for different purposes. It is necessary to pick up the best suited model.
- Training: The model is then needed to be trained properly.
- Adjustments: The predicting parameters are then adjusted to increase the accuracy.
- deployment: The model is now ready to be deployed and predict values for new data.

---

(12) In Machine learning, the term bias is referred to as the ~~average~~ difference of average prediction made by our model and the actual value of that point. A model with high bias oversimplifies the training data and leads to wrong predictions this is also known as underfitting problem.

The machine learning model that pays a lot of attention to the training data and tries to fit all the available data is known as a high variance model. Due to perfect values for the training data, the model fails to generalize the trend and provides inaccurate results. This is also known as overfitting problem.

The objective function is structurally similar to the original mean squared error function but with an extra term known as regularizer

$$J(W) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_w(x_i) - y_i\right)^2 + \underbrace{\lambda_j \sum_{i=1}^{m} W_j^2}_{\text{regularizer}}\right]$$

differentiating

$$\frac{\partial J(W)}{\partial W_i} = \frac{1}{2m}\left[\sum_{i=1}^{m} 2\left(h_w(x_i) - y_i\right)(x_i) + \lambda \cdot 2W_j\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(x_i h_w(x_i) - y_i x_i\right) + \frac{\lambda}{m} W_j$$

The regularizer limits the affect of the parameters of the model. Therefore it can be said to have a shrinking affect on the parameter

(4) Top-down strategy

In top down approach, the process of designing the system proceeds from top till bottom in the heirarchy. The identification of main module is followed by division of the main module into several submodules that are placed lower in the heirarchy. The division of modules is carried out till the lowest level untill further division is not possible.

Bottom up strategy

The process of designing a system proceeds in a bottom up manner where most basic and fundamental modules are identified at the lowest level in heirarchy and are combined to form bigger main modules until the last module is formed at the highest level in the heirarchy.