13/8/2024

Regulations are inversely proportional to innovation → False Myth

→ Not much policy on IT/Digital sector

Q/ Does any IT Company come under MSME? → No IT comes under it

↳ Come under shopkeeper act.

In the last 1 decade, rate of people being hired in digital sector has declined.

Some much power to very few Companies → Paranoia

↑ only to replace the fired people

⇒ "Do No Harm" in the google people. Google ethics Council set up to pass for any product having AI and produced in google.

⇒ Google ethics council was fired (They wrote the paper). Largest shifts in tech policies occured because of the paper.

→ Companies are incapable of self Governance. because aim of Company is to increase share holder value. (As much money as possible)

⇒ Either by Innovation or Exploitation ⇒ Physical need of the Company.

↳ has an upper limit

→ After google many companies have ethics Council. Microsoft has ethics council but they came in case of selling fire for Border Security Force.

→ Greenwashing is the use of environmental language of exploiting by the power Companies

Pinkwashing → "Ethics" washing"

"Techno solutionism and reification"

story: 17th Century Victoria Britain, ⇒ women are not considered people, they are considered property.

⇓

Elon Musk comes from South Africa

↳ Pressure cooker given by him

Why is problem solutions still not be solved?

→ Societal problems are not solved, when rushed through in the technology.

Any societal problem can't be solved by replaced by structural things by an accelerator but it does not change the fundamentals of problem.

⇒ Power is not changed by technology, but velocity increases only.

Bangalore app for traffic control, does not solve the real problem.

⇒ App to track malnutrition, but nurses are limited only.

~~Bias~~, ~~leaderboard~~ Reification

Assuming Cooking is a ~~math~~ bunch of distinct categories, we had incentive to agree ~~the~~ on the categories. By creating categories, false reality was made.

⇒ Benchmark in the AI space ⇒ Make categories → More and more people take part in it → SOTA on the paper.

[Creating reality by creating Categories]

We all agree that money has value ∴ Money has value ⇒ [Reification]

AI as a scapegoat → Entire field of AI is technology

→ To use ~~or~~ or not tchnogy comes under policy.

⇒ ~~often~~ people are wrongly copystriked on youtube due to wrong M.L. algorithm.

→ The decision to use the code is with the companies or political ~~companies~~ decisions.

→ ~~Tech~~ Every tech has some internal qualities, like AI has ability to learn from the past.

→ Problems arising from it can't be just put on AI. Humans also need to be held accountable.

What is Machine Learning?

Trademark → Science

Algorithms that improve automatically through experience by using data.

Training data has annotations in the supervised M.L. models.

⇒ Image → Features → output

⇓

useful qualities extracted from the image.

Can be represented as vector. (n Dimensionals as well)

For every feature we add 1 Dimension.

→ Feature vector gives the location in the N-Dimensional space.

↳ Couldn't throw raw data, had to find features.

→ Having too many features led to curse of dimensionality

Computationally M.L. becomes very costly.

Dimensionality Reduction using PCA, SVD

↳ Singular Value Decomposition

We can use raw data as well.

⇒ Jargons

① Cross Validation → Validation set is smaller as compared to Train data.

↓ How to make it more Robust

80,000 Images of (Mango + orange)

(8000) Images (10 part) → 9 part (72,000) → Train

→ 1 part (8,000) → Valid

Repeat process 10 times with different combinations.

10 times Validation results and take the Average of these.

(Repeatedly validate and take Average)

A new species of Mango. (All our data does not come at once)

Re-training with all data ~~We don~~ at every time is not feasible

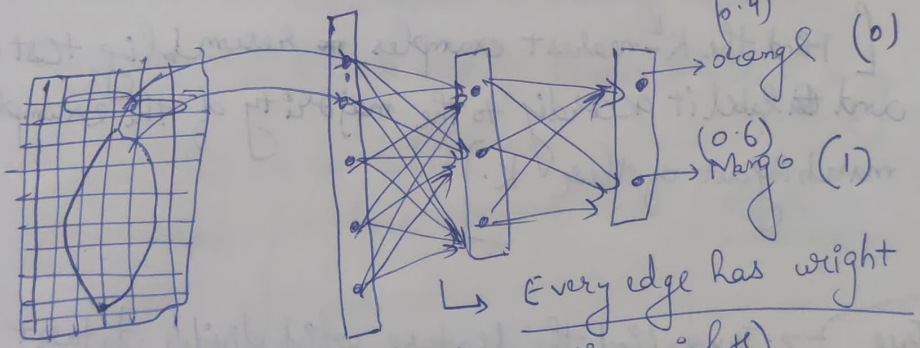
We should only be adding images which are telling something new.
[Use the old Machine Learning to see which images are giving the Error] → Size is increasing only on the datapoint which are actually adding value to the model.
⇒ useful are the ones which are confusing the old model.

Ensemble Learning → Take a bunch of Models and join them.

oracle is an algorithm that creates synthetic testing use case.
(Synthetic Apple and orange images to test)

overfitting → ~~Training~~ Datapoints in training set lack variety, the model is very blinkered and has narrow scope of vission
→ (Brittle Model) (New Testing Data can break it easily)

Explainability → Models that can give indication of the results it gives
→ Decision Trees are explainable while Neural Networks are not.



↳ Every edge has weight
(Backpropagation changes the weights)
(Difference in intensity gets the edges of images inside the Neural Network)

[Perceptron and Deep Neural Network]

16/8/2024

Supervised M.L.

Classification

Regression
(Any kind of problem where output is not discrete)

[Inter Annotator] Score and Annotator Agreeing Matrix

→ Multiple annotators sometimes used to hide certain datasets that do not exist.

→ Annotator agreement is not a guarantee for Dataset quality, but a bare minimum for Dataset validity.

Best Algorithm you take is not the most acurate for the task. It can also be due to limitations in hardware as well.

High variance is not good as well as too much bias is not good for predict

→ K-Nearest Neighbour → High Performance for low tech design

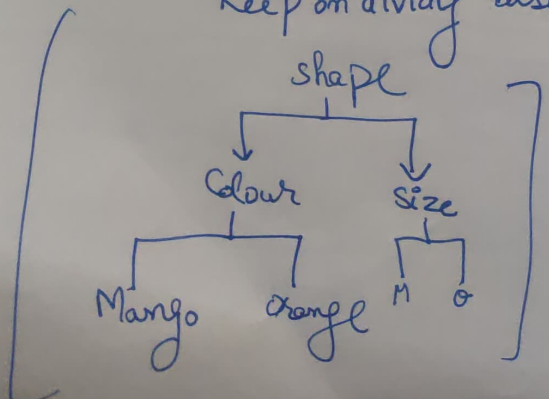
↳ Take training data, take the test set example

[Find the K-nearest examples resembling test example and label it accordig to the majority of the examples matching out of these K.]

→ Decision Tree → Try to find the feature which divides dataset into 2 parts.

Find the feature which best divides the classes

Keep on dividig based on features usefulness.

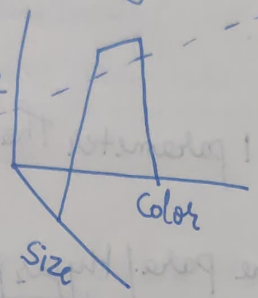


Bayesian Classifier → Use of probability theory
↳ Naive Bayes Algorithm

Simplest classifier assumes that features do not affect each other (Independent)

(Makes Maths simple) ⟹ Naive Bayes Classifier

⟹ We are assuming linearity pre-conditioning



⇒ $f(f_1, f_2 \cdots f_n) = \text{Label}$

Assumes Linear

If not Linear → then we use Neural Network (A sufficient Complex N.N. Can model any relation ⩾ 3 Layers)

Another model that can model (SVM) [Support Vector Machine]

Kernels used to split data points into classes only 2

→ Advant → Not as complex as Neural Networks

Disadvant → It can be used for binary problems only

But we can do like

Cars | others ⟹ Bike | Remaining

⇒ Use chain of classifiers → N-ary SVM classifier

N different SVM's can be used.

Unsupervised M.L.

(No Annotation is provided)

↓

No indication of data is.

→ we try to find useful internal patterns like Natural groups

⇓

Clustering

We can give the no of clusters and model will break them into those clusters.

Parametric Algo -: Clustering has 1 parameter that are given manually.

⇓

which require some para./Hyper para.

Bayesian Non-parametric clustering Algo -:

Chinese Restaurant and Indian Buffet Algorithm

Data Point ⇓

Stay or go to another cluster.

Topic Modelling -: we are given 100K digitised q.p. of IIT's.

⇓ → It finds out statistical distribution of words in the corpus.

It finds overlapping themes and figure out words in each of these themes.

Anomaly Detection -: Video Camera footage

Find out features → cluster them

Some photos won't fit in any of the clusters.

GAN (Generative Adversial Network)

↳ one part figures data patterns and other tries to differentiate and classify real and fake

(Realistic Data points)

Q) What is Coding ?

↳ steps of instructions give to Computer to perform tasks

⇒ Human comprehensible Instructions came (Assembly languages)

↳ very long code lines

⇒ Python ⟶ Assembly (HLL) or Directly to machine Language

↳ Ease of understanding

Python do not have pointers → Unlike C++ which has pointers having direct interaction with hardware

(Inefficient)

Compilers take Code and translate to Machine Language

We cannot use reserved words in Python.

Feature overloading (Same '+' can be used for different purpose)

We can use split function and ':' join (x)

x.split(' ')