22/02/2022

Machine Learning and Deep Learning

Unit - I

Introduction to Machine Learning

History of ML:

1) 1950 – Introduction of Turing Test by Alan Turing
   Q → "Can Machines Think?"
2) 1952 – Samuel introduced first ML Program
   ie The playcheckers game
3) 1957 – Frank Rosenblatt invented first neural network
4) 1967 – Pattern Recognisation (KNN algorithm introduced)

25/02/2022

5) 1979 – Self Driving Carts → by Stanford university students.
6) 1982 – RNN – Recurrant Neural Network.
7) 1989 – Reinforcement Learning
8) 1995 – Random Forest and SVM (Support Vector Machine)
   (accuracy of model > 95%)
9) 1997 – IBM deep blue (chess game)
10) 2006 – Netflix (Recommendation System)
11) 2010 – Kaggle Website
12) 2011 – IBM Watson (Jeopardy)
13) 2016 – Google Alpha Go

26/02/2022

Human Learning:

In a cognitive science, learning means gaining the information from observation.

There are 3 types of learning. They are.

1) Learning under expert guidance
2) Learning by self
3) Learning guided by knowledge gained from experts

Machine Learning:

It is a computer program which is said to learn from the experience E with respect to some class of task T and performance measure P, if its performance at task T as measured by performance P and improves with experience E.

Well Posed Problem:

1) Play Checkers game
   - E → represents the experience to play the game.
   - T → represents the task of playing checkers.
   - P → Performance measure indicated by percentage of game won by the player.

2) Image Classification.
   - E → taking the labelled data as input.
   - T → To classify the image
   - P → To classify image based on percentage.

   (or)

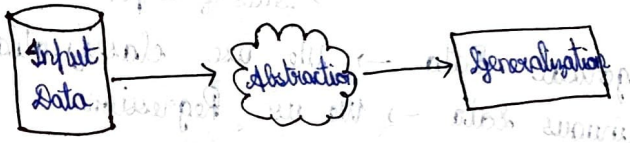
   - E → Past data with images having labels
   - T → Task assigning class to new unlabelled data
   - P → The performance measure indicated by the percentage of image is correctly classified

How do Machines Learn?

To learn a machine, we need to take 3 steps.

1) **Input Data:**
   Past data / information is utilized as a basis for future decision making.
2) **Abstraction:**
   The input data is represented in a broader way through the underlined environment.
3) **Generalization:**
   The Abstracted representation is generalized to form a framework for making decision.

Process of ML:



Machine Learning Types:

1) **Supervised Learning:**
   model is an equation ie mathematical equation to solve a problem.
   It consists of input and output data and a supervisor.
   Input (Independent Variable)

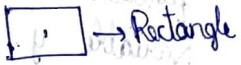
Image

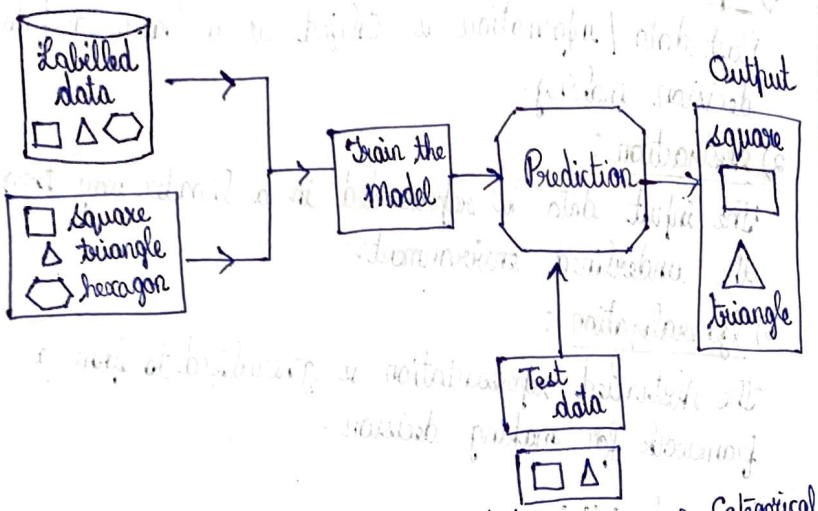
Output

Square, Triangle

Dependent Variable

Testing





Supervised learning → tumor detection (Yes/No) — Categorical Value Data
→ Sales of a product

Categorical Data → We use classification algorithms
Continuous Data → We use Regression

↳ 1) Classification and 2) Regression.

* Supervised learning is a process of providing input data as well as correct output data to the machine learning model.
* The aim of supervised learning is to find a mapping function to map the input variable x with the output variable y.
* The input variable x is called as Independent variable and output variable y is called as Dependent variable.

Examples:
1) Predicting result of a game.
2) Sales Prediction.
3) Classifying set of emails as spam or not spam.

There are 2 different types of supervised learning
1) Classification and
2) Regression.

Classification:
If we are trying to predict categorical and nominal variables.

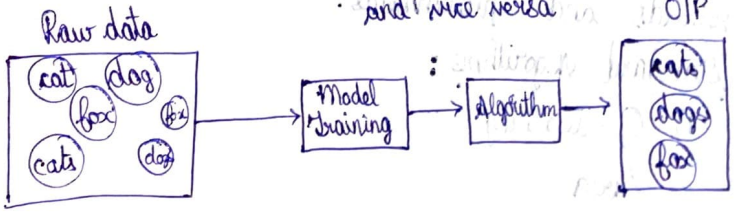
Regression:
When we are trying to predict continuous or real value variables.

Classification: Problems / Examples:
1) Image Classification
2) Prediction of a Disease
3) W/L prediction of a game
4) Handwriting recognition.

Regression is a statistical method to model the relationship between dependent variable (target variable) and independent variables (predictor).
↓
may be multi variable

Regression Problems / Examples.
1) Sales / Price Prediction
2) Weather Forecasting.

2) Unsupervised learning:
Similarity Measures → Distance b/w clusters is more → less similar and vice versa



1) Clustering → clustering similar data into clusters..
2) Association → Market Basket Analysis
Common relation between the different input data.
data

Applications involved in Unsupervised Learning:

* Unsupervised Learning contains 2 types
  1) Clustering
  2) Association

Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into the group and has lesser distance(or) no similarities with objects and categorize them as per the presence and absence of those commonalities.

Association is used to find the relationship between variables in large database.
It determines the set of items that occurs together in the dataset.
Association rule makes marketing strategy more effective.
After finding Association Rule, we can use regression analysis to predict the sales.

Reinforcement Learning is based on rewards and punishments. Based on environment, agent must act.
The model learns and updates itself based on the rewards and punishments.

Standard Algorithms:
1) Q-learning
2) Sarsa

Applications of Unsupervised Learning:

05/03/2022

1) Recommendation Systems
2) Anomaly Detection
3) Association Rule Mining
4) Similarity Detection
5) Product and Customer Segmentation.

Problems not to be solved by using ML:

1) Reasoning Power
2) Contextual Limitations
3) Scalability
4) Internal Working of Deep Learning.

Applications of ML:

1) Banking Sector
2) Health Care
3) Fraud Detection
4) Self driving cars.
5) Image Recognition
6) Traffic Prediction
7) Sales Trading
8) Product Recommendations
9) Speech Recognition.

Tools of ML:

1) R lang
2) Python
3) MATLAB
4) SAS s/w.

Issues of ML:

1) Poor Quality of data
2) Underfitting training data
3) Overfitting training data
4) ML is a complex process
5) Slow Implementation.

09/03/2022

Input → Raw Data

↓

Activities → Step 1.

Preparing a Model:

1) Understanding the type of data in the dataset.
2) Explore the data to understand quality.
3) Explore relationship blw features, attributes, variables etc
4) Find the potential Issues.
5) Remediate the data
6) Apply Preprocessing techniques
   1) Dimensionality Reduction
   2) Feature subset selection → Relevant data.

↓

Output is Refined Data

Refined Data is input to step 2.

Learning: Training the Model.

1) Partitioning → Dividing the data into
   1) Train Data
   2) Test Data.
2) Model selection → Based on Data, select the correct specified model like SVM etc. to perform specified task.
3) Cross Validation

Performance Evaluation

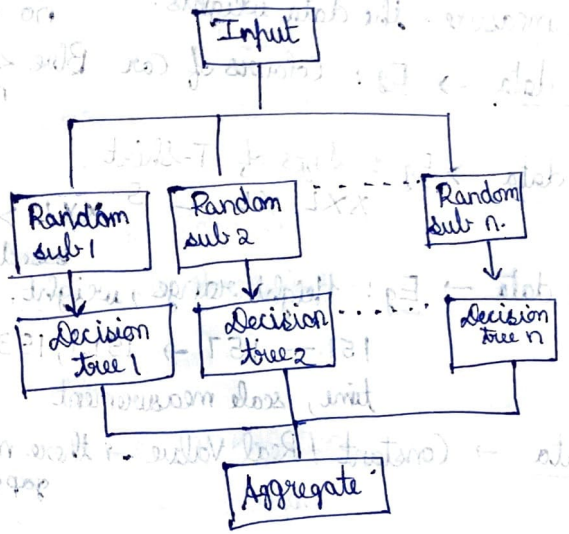
1) Examining Accuracy of Model.
2) Visualization of Accuracy Scores (ROC Curves)
3) ROC Curves plot

Performance Improvement

based on hyperparameters

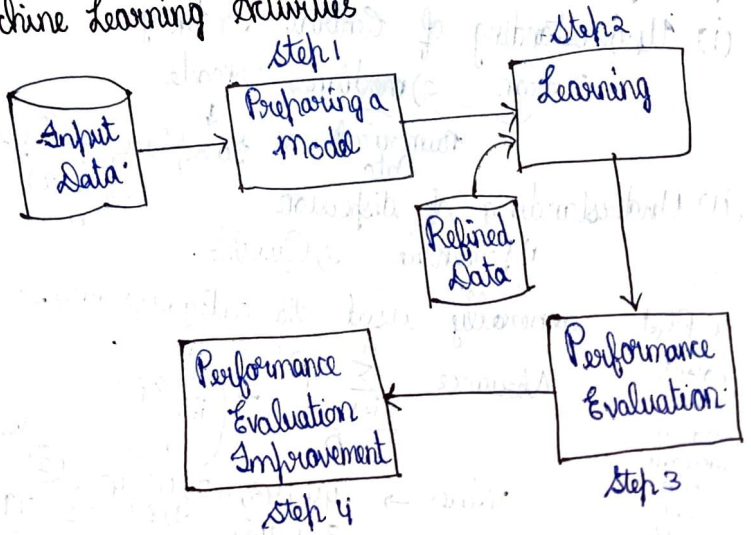
1) Tuning the Model → Trial & Error Method by changing
2) Ensembling → consider n base models → by training & testing find optimal solution
3) BAGGING
4) Boosting

→ Bootstrap and Aggregation



Consider the weak learners/models and train them to obtain better results to become a good learner.

Machine Learning Activities



Types of Data:

1) Categorical Data → Nominal Data, Ordinal Data } Qualitative
2) Numerical Data → Continuous Data, Discrete Data } Quantitative

We cannot measure the data weights.

We can measure the data weights.

Nominal data → Eg: Colours of Car Blue < Black — no order b/w values, not possible

Ordinal data → Eg: Sizes of T-Shirt
XXL XL L S — XXL > XL possible based on order.

Continuous data → Eg: Height range, weight.
151 — 157 → 151.1, 153.6, ----
time, scale measurement

Discrete Data → Constant / Real Value → there may be gaps b/w freqbl.

11/03/2022

Exploring Structure of Data:

1) Exploring Numerical Data

(i) Understanding of Central Tendency
1) mean 2) median 3) mode
Numerical Data
↓
To replace outliers

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

(ii) Understanding of dispersion
1) Percentile 2) Quartile.

→ Box Plot generally used to categorize outliers

count
mean
median
variance
min
25%.
50%.
75%.
max.

Variance $\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$ Variance

Values → 44, 46, 48, 45, 47. ② 79.6
34, 46, 59, 39, 52

1) Central Tendency
- mean } Numerical Data
- median
- mode Categorical data

2) Data Dispersion
- Variance
- Standard Deviation $\sqrt{\text{Variance}}$
- Percentile
- Quartile

25%. 50%. 75%.
Q_1 Q_2 Q_3

25%	25%	25%	25%

IQR - Inter Quartile Relationship
$Q_3 - Q_1$

(ii) Plotting Numerical Data
Use of Boxplot, histogram

Exploring Relation between variables → min Max Q_1, Median Q_3
1) Scatter plot
2) Two-Way-Cross Tabulation x,y.

Data Quality and Remediation:

Best Quality → Standard Datasets.

Issues
1) Incorrect Sample Set Collection
2) Errors in Data Collection.

Data Remedies:
1) Handling Missing Values
2) Handling Outliers

Remove Outliers
1) Removing Outliers which are unnecessary.
2) Imputation → Assigning similar frequent values
3) Capping ~5% ⟷ 95% ~

2) Handling Missing Values
1) Eliminating records which are having missing values
2) Imputation of missing values
3) Estimating Missing Values.

Dimensionality Reduction:

It is a technique in which the conversion of higher dimension dataset to lower dimension dataset

Prediction:

It is widely used to obtain a better fit predictive model while solving classification & regression values.

Feature Subset Selection:

It is a process to select the subset of relevant features and leaving out irrelevant features

19/03/22

Unit - II

Selecting a Model:

1) Input Data → Raw data
2) Abstraction → Training data
3) Generalization → Testing data

$y = f(x) + c$ → error

- y → Output dependent variable
- f → Target function
- x → input independent variable.

Target functions used:

1) Cost function → error function - tells us how bad our model (Only applicable to training dataset) is going to function.
2) Loss function → applicable on a particular datapoint
3) Objective function → We consider data and model

To select a model, we have 2 parts

1) Predictive
2) Descriptive → unsupervised learning

Predictive:
- Uses classification algorithms for categorical values
- regression for continuous values.
- Supervised learning

Descriptive:
- K means Clustering
- Market Based Analysis
- Recommendation System
- Pattern Recognition

Cost function:

It is also called as an error function which helps to measure the extent to which the model is going wrong in estimating the relationship between x and y. In that sense, cost function can tell how bad the model is performing. R^2 regression function can be used

Loss function:

It is similar to the cost function. The only difference is the loss function is defined on a data point while cost function is for entire training dataset

Objective function:
It is used to find most suitable solution to a point. We need to evaluate the quality and optimality of system. We can also call it goal function.

ML algorithms broadly divided into 2 types

1) Supervised learning — primarily focussing on solving predictive problems.

2) Unsupervised learning — primarily focussed on descriptive problems

Predictive Models — They try to predict certain value using the values in an input dataset. It has a clear focus on what they want to learn and how they want to learn.

Examples: Predicting whether a transaction is fraud.
a customer may move to another product
win loss % of a cricket match.

* The models which are used for prediction of target features of categorical values are called classification models.
* The target feature is known as a class and the categories to which classes are divided into are called as levels.

Some of the popular classification models.

KNN, Naive bayes, decision trees

* The models which are used for prediction of numerical values of the target features of a data instance are called as Regression models

Examples: Prediction of revenue growth
Prediction of rainfall amount
predicting sales of a product in coming years.

Descriptive Models:

(Predictive) Descriptive models which group together similar data instances. Data instances having a similar value of the different features are called clustering models.

Examples: Customer grouping / segmentation based on social, demographic, ethnic factors.

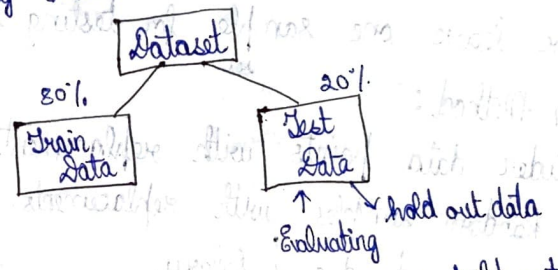
The most popular model for clustering is K-Means.

23/03/2022

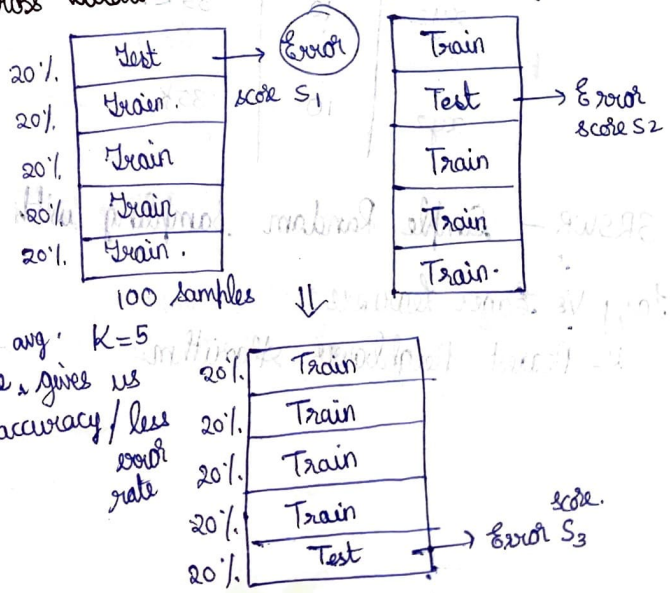
Training of Model (for Supervised Learning):

1) Hold Out Method:

Used to divide training set into training data and testing data. So we hold out some part of training set i.e testing data.



2) K-fold cross validation: Extension for hold out method.



10 fold cross validation:

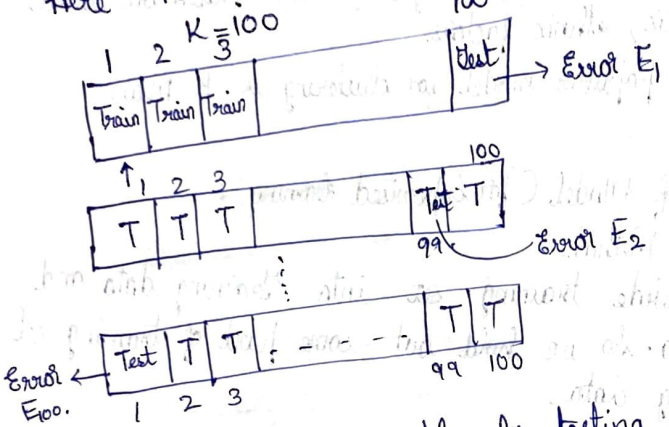
Here K value is 10.

Leave one out cross validation:

Here K = N = Number of inputs

K = 100

Time consuming process.



Always we leave one sample for testing the data set.

Bootstrap Method:

We consider data points with replacements

Random samples with replacements.

Name	Age	Salary
xyz	10	35K.
⋮	⋮	⋮
xyz	10	35K

SRSWR – Simple Random Sampling with Replacement.

Lazy Vs Eager Learners:

K-Nearest Neighbours Algorithm

Hold Out Method:

- The data is partitioned into training and testing data which is by holding back a part of the input data from validating the train model is called Hold Out Method
- There are different hold out methods are used to improve the ML model by avoiding overfitting and underfitting of model.

K-fold Cross Validation:

- Data is divided into testing and training set.
- ML model is developed using the partition of data and then tested on the rest data.
- This process is repeated K-times with different random partitions to generate an average performance.

Bootstrap Method:

- It is a random sample conducted with replacements
- While selecting the sample, the same can appear more than once.
- We can create more than one training data set from original dataset.
- It follows the technique SRSWR.

  S → Simple R-Random S-Sampling

  W-With R-Replacement.

Lazy Vs Eager Learners:

- Lazy Learner:

→ It simply stores training data that means memorising data and wait to get test data.

→ The training time is less and predicted time is less.

Example: KNN

Eager Learners:

→ It takes more time for training data.
→ While new input is arrived it generalizes and gives the quick results.

Example: SVM, Neural Networks, Decision tree.

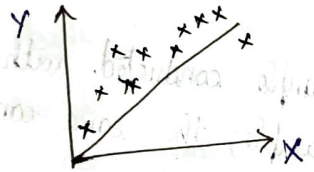
24/03/22

Model Representation:

1) Underfitting: for low amount of data.

Model should be provided with refined input data.

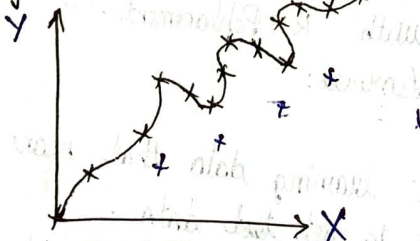
→ Bias - Average value b/w actual value and predicted value.
→ Bias is high i.e more erroneous., less Variance.
→ Low Variance and High Bias → indicate underfitting



To avoid underfitting

1) Consider more amount of data (i.e train data).
2) Removing effective variables.

2) Overfitting: Low Bias, High Variance.



We get low bias while training
We get high variance while testing

To avoid Overfitting

1) Apply K-fold cross Validation
2) Remove unnecessary data from data set

→ Best fit Model - Low bias, low variance.

Underfitting: It occurs when the model is not able to capture the underlying trend of data i.e the model may not learn from the training dataset.

Overfitting: It occurs when the model is fully able to capture the underlying trend of data but not with test data.

Bias: It is an error used to find out the average difference between predicted and actual value.

Variance: It is a measure of how far the set of data points are spreaded out from their mean value. It means to find the difference of deviation from the actual value.

$$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

x_i - i^{th} datapoint
$\bar{x}$ - mean of all DPs
N - No of DPs.

- Underfitting can be avoided by giving more training data and reducing the effective features.
- Overfitting is avoided by K-fold cross validation, Hold back of validation dataset, Removing the nodes with no predictive use.

Evaluating the Performance of a Model:

Classification Metrics

Confusion Matrix:

n = Total no of data points	Actual; No	Actual; YES
Predicted_No	TN	FP
Predicted_YES	FN	TP

Precision

Recall.

Predicted	Actual
Match would be win (T)	Match Won (P) $\Rightarrow$ (TP)
Match win (F)	Match lost (P) $\Rightarrow$ (FP)
Match lost (F)	Match win (N) $\Rightarrow$ (FN)
Match lost (T)	Match lost (N) $\Rightarrow$ (TN)

Accuracy : $\frac{TP + TN}{TP+FP+TN+FN}$

$TN = 9, \quad FP = 4, \quad FN = 2, \quad TP = 85$

$A = \frac{94}{100} = 0.94 \Rightarrow 94\%$ accuracy.

2) Misclassification Rate / Error Rate :

$$\text{Error Rate} = \frac{FP + FN}{TP+TN+FP+FN}$$

→ Confusion Matrix is a performance measurement. Problem
→ It gives the summary on prediction of classification.
→ It is an $n \times n$ matrix used to compare actual and predicted values.

Type 1 Error → FP

Type 2 Error. → FN

→ Accuracy defines how often the model predicts correct output.

→ Misclassification error is an error rate defined how often the model gives wrong prediction.

Precision:
It defines number of correct outputs provided by the model.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall:
It defines out of total +ve classes how our model predicted correctly. It must be as high as possible for best performance of model.

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(Sensitivity)}.$$

Specificity:
It measures the proportion of negative examples which have been correctly classified.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

F-Measure:

$$\text{F-Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

ROC Curves:

ROC stands for Receiver Operating Characteristic Curve.
In this we have True positive Rate (TPR) = $\frac{TP}{TP+FN}$

False positive Rate (FPR) = $\frac{FP}{FP+TN}$

20/03/2022

Performance Measure for Regression:

1) MAE - Mean Absolute Error - Actual, Average differences b/w Predicted and (Average) values.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

y_i → Actual Data Point.
$\hat{y}_i$ → Predicted value.

It is one of the best method/metric for regression.
It is the average of difference between original value and predicted value.

② **Root mean Square Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Eg: $y_i = (10, 5)$, $\hat{y}_i = (8, 6)$.

$$MAE = \frac{1}{2}\left[|10-8| + |5-6|\right] = \frac{3}{2} = 1.5$$

$$RMSE = \sqrt{\frac{1}{2}\left(|10-8|^2 + |5-6|^2\right)} = \sqrt{\frac{1}{2}(4+1)} = \sqrt{\frac{5}{2}}$$

$$= \sqrt{2.5} = 1.582$$

③ **R^2 Error:**

Residual → Error

$$R^2 \text{ Error} = \frac{SST - SSE}{SST}$$

$$= 1 - \frac{SSE}{SST}$$

SSE → Sum of Square error

SST → Sum of Square Total

$$SSE = \sum (y_{actual} - y_{pred})^2$$

$$SST = \sum (y_{actual} - y_{mean})^2$$

$$\therefore R^2 \text{ Error} = 1 - \frac{\sum (y_{actual} - y_{pred})^2}{\sum (y_{actual} - y_{mean})^2}$$

Range of R^2 error is (0 to 1)

④ **Adjusted R^2**

0% to 100%

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1-R^2)(N-1)}{(N-P-1)}\right]$$

R^2 value; Total no of value; Predicted value.

Unsupervised learning metrics:

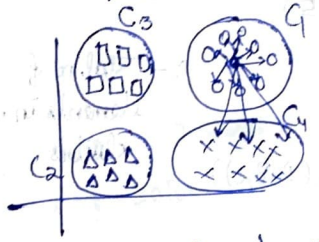
1) **Clustering:**

Internal Evaluation → Euclidean, Manhattan distances

i) Silhouette Co-efficient $= \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

Internal evaluation methods generally measures the cluster quality. Based on homogenity of data, belongs to same cluster and heterogenity of data, belong to different clusters.

$a(i)$ – Average distance between i^{th} datapoint and all other data instances belonging to same clusters.



$$a(i) = \frac{a_{i1} + a_{i2} + \cdots + a_{in}}{n_1}$$

find out similarities

$b(i)$ – The lowest average distance between i^{th} data instance and data instances of all other clusters.

b_{14}(average) $= \frac{b_{14}(1) + b_{14}(2) + \cdots + b_{14}(n)}{n_4}$

(b_{14}: cluster 1, cluster 4)

b_{12}(average) $= \frac{b_{12}(1) + b_{12}(2) + \cdots + b_{12}(n)}{n_2}$

b_{13}(average) $= \frac{b_{13}(1) + b_{13}(2) + \cdots + b_{13}(n)}{n_3}$

$$\therefore b(i) = \min\{b_{12}(\text{average}), b_{13}(\text{average}), b_{14}(\text{average})\}$$

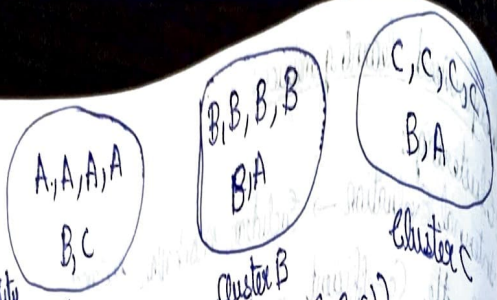
find out dissimilarities.

External Evaluation:

(1) Purity $= \frac{1}{n}\sum_{K} \max(K \cap C)$

K → Total No of classes present in the cluster

C → Target class label name.



majority elements Cluster A

Cluster B

Cluster C

For Cluster A = (K=3) $((A,B,C) \cap (A,B,C))$

$= (A,B,C)$.

$\max(K \cap C) = \max(4,1,1) = 4$

Cluster A = 4

For Cluster B = (K=2) $((A,B) \cap (A,B))$

$= (A,B)$

$\max(K \cap C) = \max(1,5)$

$= 5$

For Cluster B = 5

n → Total no of elements in all clusters

For Cluster C = (K=3) $((A,B,C) \cap (A,B,C))$

$= (A,B,C)$

$\max(K \cap C) = \max(1,1,4)$

$= 4$.

Cluster C = 4

$$\text{Purity} = \frac{1}{18}(4+5+4) = \frac{13}{18} = 0.72$$

We assign a label to each cluster based on most frequent classes in it then the purity becomes the number of correctly matched class and cluster labels divided by number of total data points.

$$\text{Purity} = \frac{\text{Cluster}(A+B+C)}{\text{Total}}$$

The Purity value increases when the number of clusters increases.

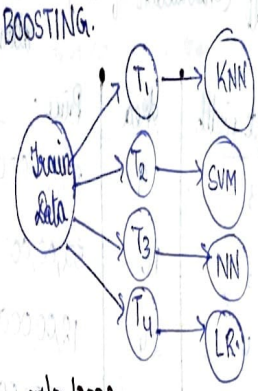
n → Total number of data instances

K → Number of distinct classes in a cluster

C → Class / Cluster label.

Improving the Performance of a Model:

1) Model Parameter Tuning
2) Ensembling
   (i) BAGGING
   (ii) BOOSTING.



06/04/2022, 07/04/2022

Basics of Feature Engineering:

→ To find relevant attributes in a dataset.

→ It is a preprocessing step

1) Feature transformation → Construction, Extraction
2) Feature subset selection.

Feature Engineering is a process of translating the data into features such that these features are able to represent the dataset more efficiently.

It is an important preprocessing step which improves the performance of a model.

It consists of 2 variants

1) Feature Transformation
2) Feature subset selection

Feature Transformation:

Transforming the data (structured / unstructured) into new set of features which represents the underlined problem which ML is trying to solve.

(i) Feature Construction:

Ex

apt_length	apt_width	Price
3	3	23,00,000
2	5	52,00,000
6	4	18,00,000
3	1	30,00,000

⇓ Construction of length × width = Area attribute

apt_length	apt_width	Area	Price
3	3	9	23,00,000
2	5	10	52,00,000
6	4	24	18,00,000
3	1	3	30,00,000

Encoding Categorical Variables: (Ordinal)

Eg: Predicting win/loss for an athelete (nominal)

Age	City of Origin	Parent Athelete	Chance of Win
18	City A	yes	yes
20	City B	no	yes
23	City C	yes	no
19	City B	yes	yes

Eliminate them

⇓

Age	City A	City B	City C	Parent Athelete		Chance of win	
				yes	No	yes	No
18	1	0	0	1	0	1	0
20	0	1	0	0	1	1	0
23	0	0	1	1	0	0	1
19	0	1	0	1	0	1	0

Since dimensions increases, space usage increases so we drop unrelevant features.

Encoding Categorical (Ordinal) files: low value → high priority

A=1, B=2, C=3, D=4

Maths	Science	Grade
78	75	B
56	62	C
89	90	A
91	95	A
66	78	D

⇒

Maths	Science	Grade
78	75	2
56	62	3
89	90	1
91	95	1
66	78	4

Converting Numerical (Continuous) features to Categorical features

Fg Eg: Prediction of House Pricing Category

apt_area	apt_price	apt_price	Ordinal
4720	23,60,000	Medium	2
2430	12,15,000	Low	1
4368	21,84,000	Medium	2
3969	19,84,500	Low	1
6142	30,71,000	Medium	2
7912	39,56,000	High	3

Range: $x < 20L \rightarrow$ Low

$20L < x < 35L \rightarrow$ Medium

$x > 35L \rightarrow$ High.

If we assume Low = 1, Medium = 2, High = 3. then.

Text → Tokenization → Variable Count → Normalization

file with text

12/04/2022

Feature Extraction:

Principal Component Analysis:

1) Standardization
2) Co Variance Matrix
3) Eigen Vectors & eigen values
4) Principal components
5) Reducing dimensions

(i) **Standardization:**

$$Z = \frac{\text{value} - \text{mean}}{SD} = \frac{x - \mu}{\sigma}$$

Covariance Matrix:

$$Cov(a,b) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{(n-1)}$$

$$|A - \lambda I| = 0$$

$$\text{Covariance Matrix} = \begin{bmatrix} Cov(a,a) & Covariance(a,b) \\ Cov(b,a) & Cov(b,b) \end{bmatrix}$$

Eigen Values and Eigen Vectors:

$$\begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix} \begin{bmatrix} 0.6778 \\ 0.7351 \end{bmatrix} \quad Cx = \lambda x$$

Principal Components:

Singular Value Decomposition (SVD):

$$A = \begin{bmatrix} \quad \end{bmatrix}_{m \times n} = U \Sigma V^T$$

$UU^T = U^T U = I$

$VV^T = V^T V = I$

Orthogonal matrices
Unitary matrices

U, V are orthogonal & unitary
Σ is a singular matrix.

U - columns → left singular vectors
V columns → right singular vectors
Σ — values of A.

U → Patterns of attributes
V → Patterns among instances.

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

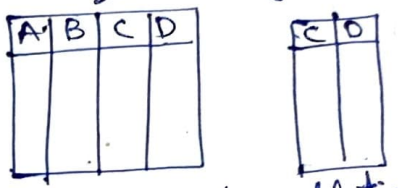
U → 1) Eigen vectors of AA^T
2) Orthogonalize (divide with distance)

V → 2) $A^T A$.

13/04/22

Feature Subset Selection:

Issues in huge data
1) Performance of model decreases.



Advantages:
→ faster
→ efficient

Key Drivers of feature selection:

feature relevance feature redundancy.

Supervised learning
Predictor Variables → irrelevant
→ weakly redundant.

In Unsupervised learning → no labelled data
Test data → grouping into clusters ⇒ based on irrelevant feature elimination
ie relevant features.
→ potentially redundant

Measures to be taken for feature relevance

1) Supervised learning.
(i) **Mutual information:**

$$MI(c,f) = H(c) + H(f) - H(c,f)$$

$$-\sum_{i=1}^{t} P(i)\log_2(P(i))$$

$$-\sum P(f=x)\log_2(P(f=x))$$

(i) Overall feature selection Process:

Steps involved:

* Feature generation $n \rightarrow 2^n$

1) Forward Selection

Take an empty set and insert the attributes one by one in forward direction

2) Backward selection

3) Bidirectional