Chapter-18

① Supervised Learning:

Supervised learning is a type of machine learning where the model is train using labeled data - that means each input has a corresponding correct output (label).

labeled data example

customer_id	Age	gender	Income (Lakh)	Years-employed (years)	total Wealth Lakh
1021210	27	Female	1 lakh	1	10
20306212	28	male	2 lakh	10	20

Types of algorithm used in supervised learning:

Classification

(i) Logistic Regression
(ii) K-Nearest Neighbour (KNN)
(iii) Decision Tree Classifier
(iv) Random forest classifier
(v) Naive Bayes classifier
(vi) SVM (support vector machine
(vii) Gradient Boosting (XGBoost, LightGBM, CatBoost)
(viii) NN (ANN, CNN, RNN)

Regression

(i) Linear regression
(ii) Ridge / Lasso Regression
(iii) Polynomial regression
(iv) Decision tree Regression
(v) Random forest "
(vi) SVR (vii) XGBoost, LightGBM

Unsupervised learning → Unsupervised learning is a type of machine learning ~~to~~ where model is trained using unlabeled data - meaning there is no pre-define output.

Examples:

(i) Grouping customer behavior. (Clustering)

(ii) Reducing dimentions for visualization (PCA)

some examples of unsupervised learning algorithm

(i) K-mean clustering

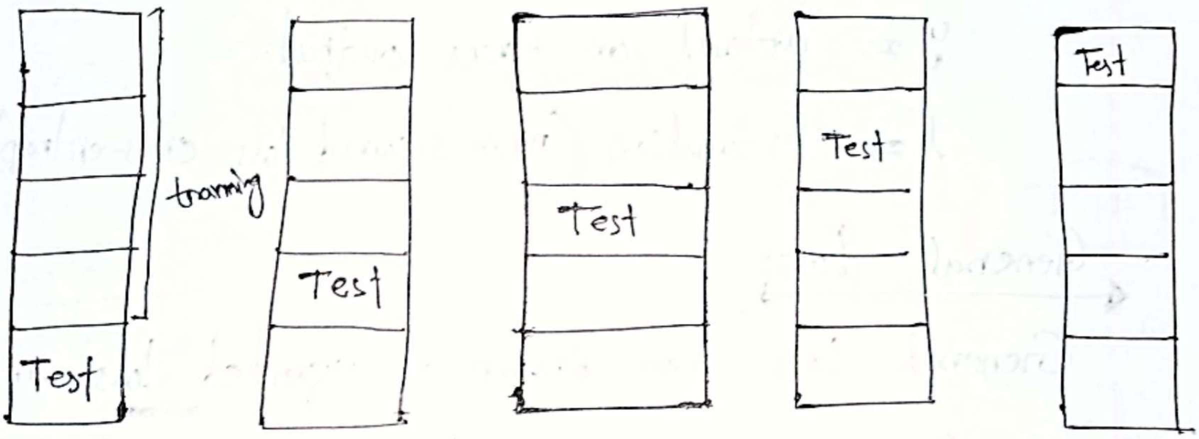
(ii) Hierarchical clustering

(iii) DBSCAN (Density-Based Spatial clustering)

(iv) GMM (Gaussian Mixture Models)

K-Fold cross validation: K-fold Cross Validation is a model evalution technique used to how well a machine learning model perform on unsean data.

It works dividing the dataset into K equal folds (parts) then tranning and testing the model K-times, each time using a different fold as the test set and the remaining folds as the tranning set. Let's assume, K=5 then



Dataset

	↓	↓	↓	↓	↓
Accuracy	a_1	a_2	a_3	a_4	a_5

$$\text{K-fold - cross validation score} = \frac{a_1 + a_2 + a_3 + a_4 + a_5}{5}$$

Empirical loss also called tranning loss:

Empirical loss is the average loss of a model over the ~~#~~ tranning dataset. It measure how well the model fits the tranning data.

$$L_{emperical} = \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i)$$

n = number of tranning sample / batch size

$f(x_i)$ = model prediction for input x_i

y = actual or true output.

l = loss function (mean-squared loss, cross-entrophy loss).

General Loss:

General loss also known as expected loss or generalization Loss. General loss is the expected value of the loss over the entire data distribution, including unseen and future data. It measure how well a model generalizes or perform to new, unseen dataset.

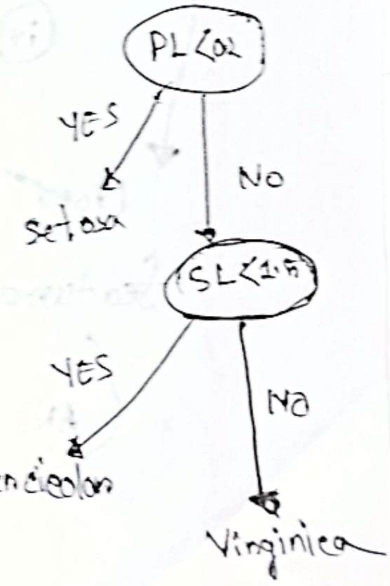
$$L_{general} = \mathbb{E}_{(x,y) \sim P_{data}} [\, l(f(x), y) \,]$$

Difference

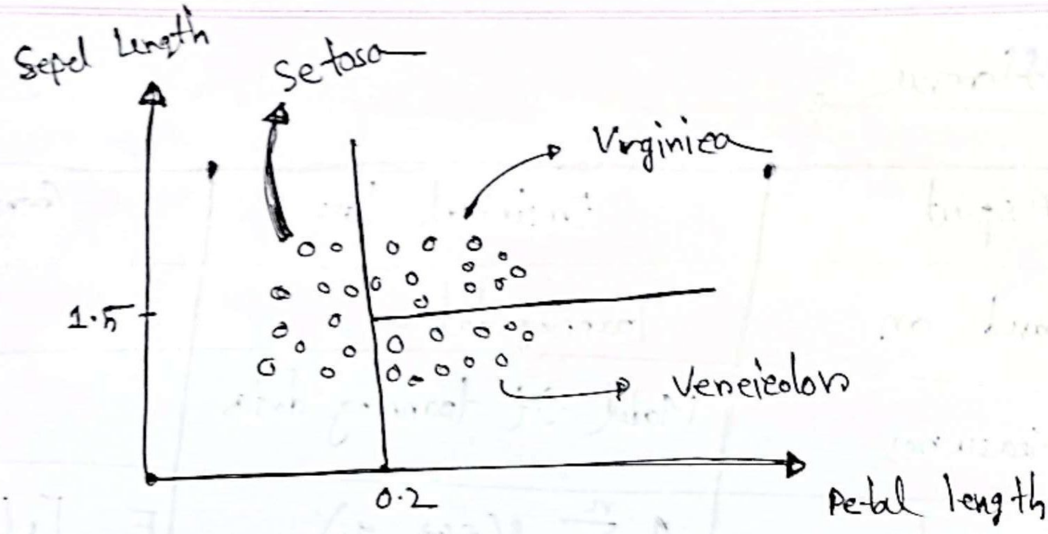
Aspect	Empirical loss	General (Expected) Loss
Based on	Training data	
Measures	Model fit training data	
Formulas	$\frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$	$E_{(x,y)}[l(f(x), y)]$
Depends on	The specific training dat	Real-world data distribution
Risk	Low emperical loss may cause overfitting	Low general loss means good generalization

Decision Tree (classification Problem)

Petal Length (PL)	Sepal Length (SL)	Type
1.34	0.34	setosa
3.45	1.45	Vencicolor
1.67	0.98	setosa
2.56	1.79	Virginica
3.00	1.13	vencicolor
1.3	0.88	setosa



Decision tree model is nothing more than a collection of nested if-else statement.



Decision trees steps

(i) Dataset

(ii) Best features (→ Previous problem 1st PL than SL)

(iii) Split data base on best features

(iv) Repeat (1, 2, 3)

Now our question is how to find the best features?

ANS → [with entrophy and information gain.]

Entrophy ⟷ Entrophy is nothing but the measure of disorderness or the measure of impurity. The mathematical formula of entrophy is

$$E(s) = \sum_{i=1}^{c} -P_i \log_2 P_i ,$$

P_i is the frequency probability of an element/class in our data.

Salary	Age	Purchase
2 0000	21	Yes
1 0000	45	No
6 0000	27	Yes
15 000	31	No
12000	18	No

→ Purchase → Yes / No (Value of c = 2)

$$E(d) = \sum_{i=1}^{2} -P_i \log_2 P_i$$

$$= -P_{Yes} \log_2 (P_{Yes}) - P_{No} \log_2 (P_{No})$$

$$= -\left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right) - \left(\frac{3}{5}\right) \log_2 \left(\frac{3}{5}\right)$$

$$= 0.97$$

Information Gain: Information gain is a metric used to train Decision Trees. Information gain is used in decision trees to find the best attributes/column to split the data at each node. Formula

Information Gain = Entrophy − Weighted Entrophy

(Parent or target column entrophy)
কমছে, ওর তার

D.T.

Outlook	Humidity	Wind	Play Tennis
sunny	High	weak	No
sunny	High	strong	No
Rain	High	strong	No
Rain	Normal	weak	Yes
Rain	Normal	strong	No
sunny	Normal	strong	Yes

Step: 1

entropy of playtennis $= -P_{No}\log_2 P_{No} - P_{yes}\log_2 P_{yes}$

$= -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.92$

1. outlook column

$E_{sunny} = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.92$

$E_{Rain} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92$

$\text{Weighted } E = \left(\frac{3}{6}\times 0.92\right) + \left(\frac{3}{6}\times 0.92\right) = 0.92$

$IG_{outlook} = 0.92 - 0.92 = 0$

2. Humidity

$E_{High} = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0$

$E_{Normal} = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.92$

~~IG Humidity~~

$\text{Weighted } E = \left(\frac{3}{6}\times 0\right) + \left(\frac{3}{6}\times 0.92\right) = 0.459$

$IG_{Humidity} = 0.92 - 0.459 = 0.461$

3. wind

$E_{weak} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$E_{No} = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$

$\text{Weighted } E = \frac{2}{6}\times 1 + \frac{4}{6}\times 0.811 = 0.874$

$IG_{wind} = 0.92 - 0.874 = 0.046.$

CamScanner

$IG_{Humidity} > IG_{wind} > IG_{outlook}$

$\therefore$ Root Node (Humidity)

Step: 02

Humidity

High

Normal

High	No
High	No
High	No

[pure don't need spliting]

outlook	Humidity	wind	play tennis
Rain	Normal	weak	yes
Rain	Normal	strong	No
~~Rain~~ Sunny	Normal	strong	Yes

$E_{Normal} = 0.92$ (from previous)

Wind:

$E_{strong} = 0 - \frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

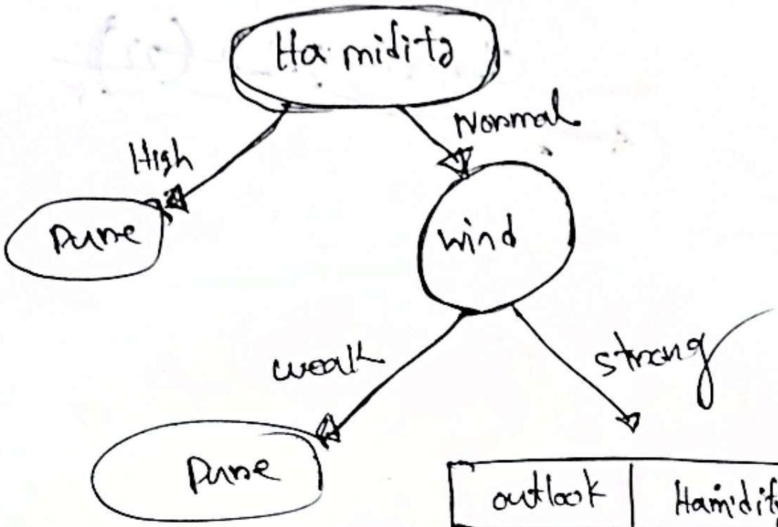
$E_{weak} = 0$

Weighted entropy $= \left(\frac{1}{3}\times 0\right) + \left(\frac{2}{3}\times 0\right) = 0.33$

value may be same take anyone

$IG_{wind} = 0.92 - 0.33 = \cancel{0.92}\ 0.59$

(highest IG Gain) [split base on this.]



outlook	Hamidity	wind	play
Rain	Normal	strong	No
sunny	Normal	strong	Yes