Classification

dealing with categorical labels in terms of grouping the data

PREDICT → The given features belongs to which group

features : anything
Label : categorical

BINARY CLASSIFICATION

label will have only two unique data points

Yes | No
0 | 1
True | false
spam | ham

- Logistic Regression ⇐
- K-NN
- Naive Bayes
- Decision Tree classifier
- Random forest classifier
- Support vector classifier

MULTI-CLASS CLASSIFICATION

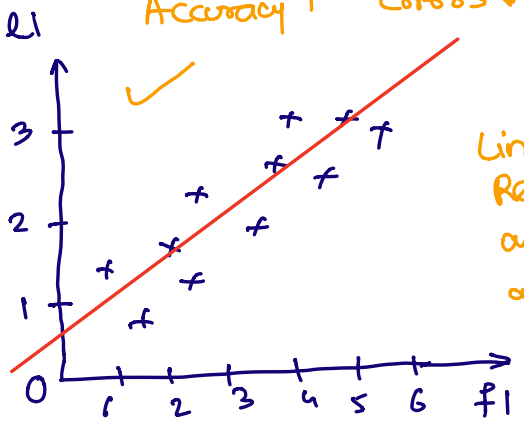
label will have more than two unique points.

setosa | versicolor | Virginia

Primary | Social | Updates | Promotion

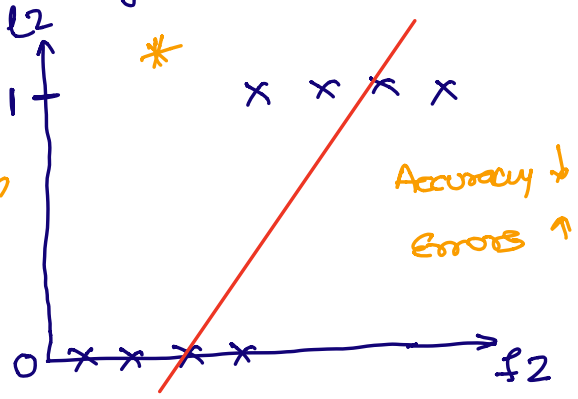
- K-NN
- Naive Bayes
- Decision Tree classifier
- Random forest classifier
- Support vector classifier

Logistic Regression .... Classification algo.
(Binary classification)



Dataset 1
(Scatter Plot)

Linear Regression as base algo.



Dataset 2
(Scatter Plot)

Two values → 0 ✓, 1 ✓

f^n that will return probability

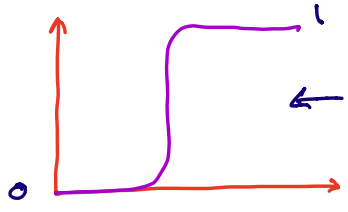
* ① Sigmoid f^n
② Bayes Theorem

The motto of the experiment was to modify the linear Regression formula such that the outcome is a probability (0-1) value and based on probability value we can define the threshold for 0 & 1

Logistic Regression = Linear Regression applied over Sigmoid function.

(Sigmoid f^n)

$$S = \frac{1}{1 + e^{-y}} \quad \text{--- (1)}$$



$$y = b_0 + b_1 f_2$$ —— Linear Regression formula —— ②

Substituting y in eqn ①

$$s = \frac{1}{1 + e^{-(b_0 + b_1 f_2)}}$$ → Logistic Regression equation

APPLIED ML → Logistic Regression

① sklearn.linear_model ... LogisticRegression.

② Applicable only for binary classification

Applied ML → Classification.

① Check whether the given problem statement is a Binary classification or multi-class classification.

eg. data.label.value_counts()

② Check whether the given dataset is a balanced dataset or unbalanced dataset.

Classification dataset

- Balanced: your label data points will have equal no. of data
- Unbalanced: your label data points will have an uneven amt of data.

iris.csv ⟶ 150 observation / data points

Setosa ⟶ 50
Versicolor ⟶ 50
Virginica ⟶ 50
} ⟶ BALANCED DATASET.

Setosa ⟶ 51
Versicolor ⟶ 49
Virginica ⟶ 50
} ⟶ UNBALANCED DATASET

The reason we need to know whether the dataset is balanced or not is,

⇒ ① To check the quality of the model

* ② To decide the metric for feature optimization

MODEL QUALITY METRICS

↓

× → (statistical inference)
↦ → (domain)

Balanced Dataset
(rely on statistical inferences)
- Accuracy Score of the model

Unbalanced Dataset
- Precision
- Recall
- F1 score
} of the model

Note:- The rule for generalization remains same in Classification. (test accuracy score > train accuracy score)

EVALUATION METRICS FOR CLASSIFICATION

Confusion matrix → ① Accuracy
CM is a table used to ② Precision
describe the performance ③ Recall
of the model. ④ F1-Score

example of confusion matrix (iris.csv) — Balanced → S : 50, Ve : 50, Vi : 50

Dataset used : iris.csv
Features : sepal_length, sepal_width, petal_length, petal_width
label : species

Actual values (y_true)	Predicted: Setosa	Predicted: Versicolor	Predicted: Virginica	Support
Setosa	50	0	0	⇒ 50
Versicolor	0	45	5	⇒ 50
Virginica	0	1	49	⇒ 50

y_pred = predicted by model

$$\text{Accuracy} = \frac{50 + 45 + 49}{50+0+0+0+45+5+0+1+49} = \frac{144}{150} = 0.96$$

(True values / every thing)

(applicable for the entire model)

96% ←

[Go Diagonal]

Precision → Precision is applicable for each label data point

[Go VERTICAL]

	S	Ve	Vi
→ S	50	0	0
→ Ve	0	45	5
→ Vi	0	1	49

$$\text{Precision}_{(\text{setosa})} = \frac{50}{50+0+0} = 1 \simeq 100\%$$

$$\text{Precision}_{(\text{versicolor})} = \frac{45}{45+1+0} = 0.97 \simeq 97\%$$

$$\text{Precision}_{(\text{virginica})} = \frac{49}{0+5+49} = 0.9 = 90\%$$

Recall → is applicable for each data point

(Go HORIZONTAL)

$$\text{Recall}_{(\text{Setosa})} = \frac{50}{50+0+0} = 100\%$$

$$\text{Recall}_{(\text{versicolor})} = \frac{45}{0+45+5} = 90\%$$

$$\text{Recall}_{(\text{virginica})} = \frac{49}{0+1+49} \simeq 98\%$$

	S	Ve	Vi
S	50	0	0
Ve	0	45	5
Vi	0	1	49

example use-case

Email spam classifier.

	ham	spam
ham	10 (1)	100 (2)
spam	5 (3)	30 (4)

① Is the given data set balanced or unbalanced?

Ans: Support of each label

ham = 10 + 100 = 110

spam = 5 + 30 = 35

Total data pts = 145 //

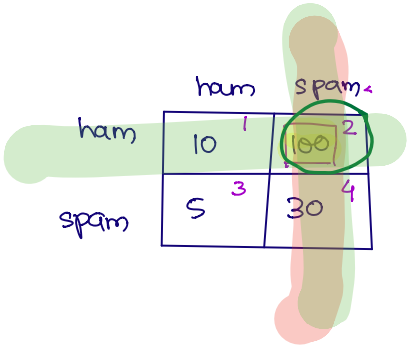
∴ given dataset is unbalanced since n(ham) ≠ n(spam)

② What is the expectation of client in terms of error tolerance in domain perspective?

spam → Inbox [mark as spam]

⇒ ham → Spam [☹]

↳ not tolerable. (domain's perspective)

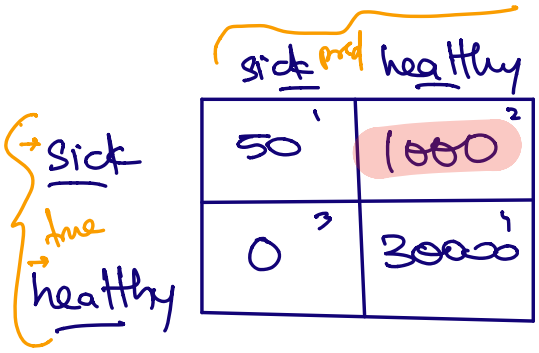


$R_{ham} \rightarrow \frac{10}{10+100} = 0.09 \approx 9\%$

$P_{spam} \rightarrow \frac{30}{100+30} = 0.23 \approx 23\%$

Use-case 2 → Medical diagnostics use-case.

Role of this model is to predict whether the patient is sick or healthy.



Recall → Sick → $\frac{50}{1050} = 0.05 = 5\%$

Precision → healthy → $\frac{30000}{31000} = 0.96 \simeq 96\%$

① Is the data balanced or not? → Unbalanced

② Binary / multiclass

③ To check generalization what need to be performed?
→ Raise a query of generalizn check to DS.
The model is generalized!

④ Please check the quality of the model and send the same for approval.
Ask for error tolerance

⑤ Hospital is saying following points:

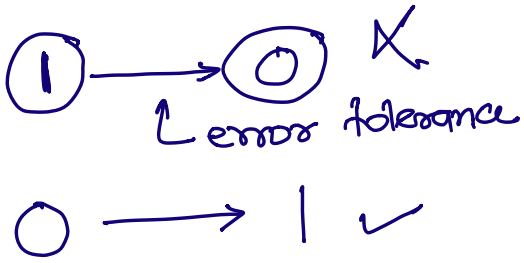
Healthy → Sick ☺ Business

Sick → Healthy ☹

↑ not tolerable

Shopping mall dataset

	0	1
0	238 (1)	19 (2)
1	43 (3)	100 (4)



$$\text{Precision}_{(0)} = 0.84$$

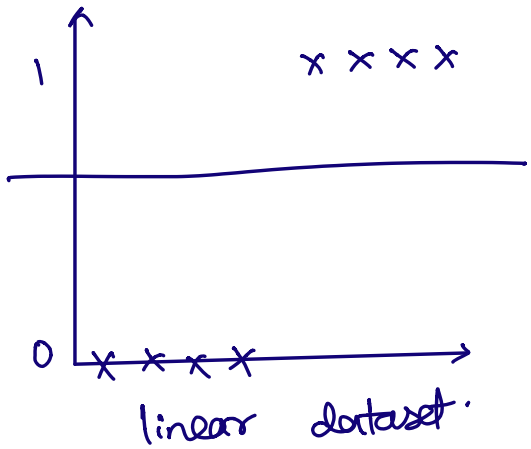
$$\text{Recall}_{(1)} = \frac{100}{143} = 0.69$$

K- Nearest Neighborhood algorithm

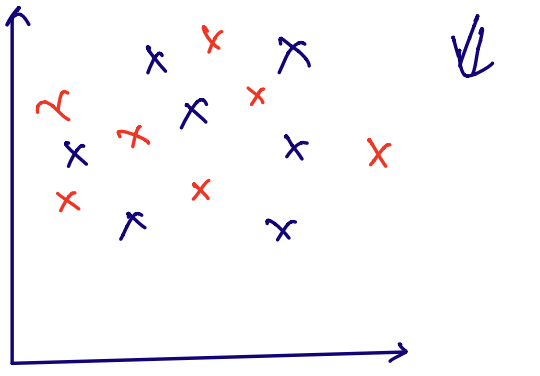
Expectations of most of the classifn algo:

① Your data must be linearly seperable

② Your use-case must be binary classifn (Logistic Regression)



linear dataset.



non-linear dataset

K-NN is BEST when it comes to dealing non-linear seperable datasets!

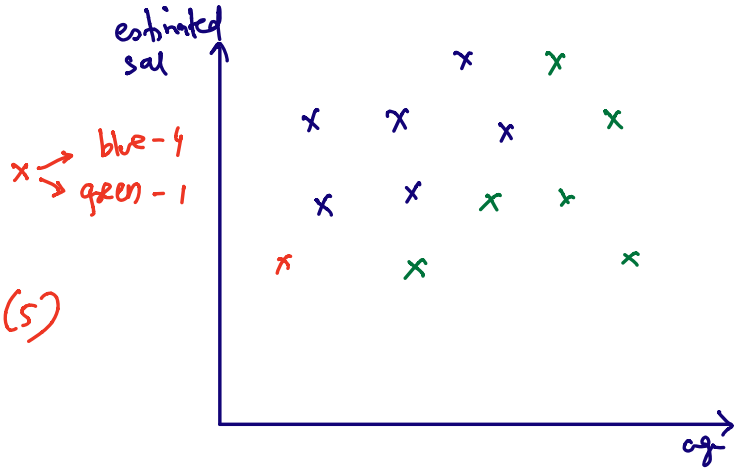
Algo:-

① Define the value of k. (k → no of neighbour)

② Get k nearest data points from unknown data pts.

③ Perform voting and select the majority class.

④ Reassign majority class as final class o/p for given feature.



k = 5

12 calc → sort in asc → head(5)