Machine Learning

Supervised Learning

Regression
Features: any
Label: numeric

Linear Regression
Support Vector R.
Decision Tree R.
Random Forest R.
xgboost - R

Classification
Features: any
Label: categorical

Logistic Regression
K - NN
Decision Tree C
Random Forest C
Support Vector C
Naive Bayes.
xgboost — C

Unsupervised Learning

Clustering
-k-means
-DB scan
-Hierarchical clustering

Association Rule Mining.
- Apriori
- Eclat

Dimensionality Reduction
- PCA
- LDA

Simple Linear Regression.

One feature
One label.

Slope intercept formula
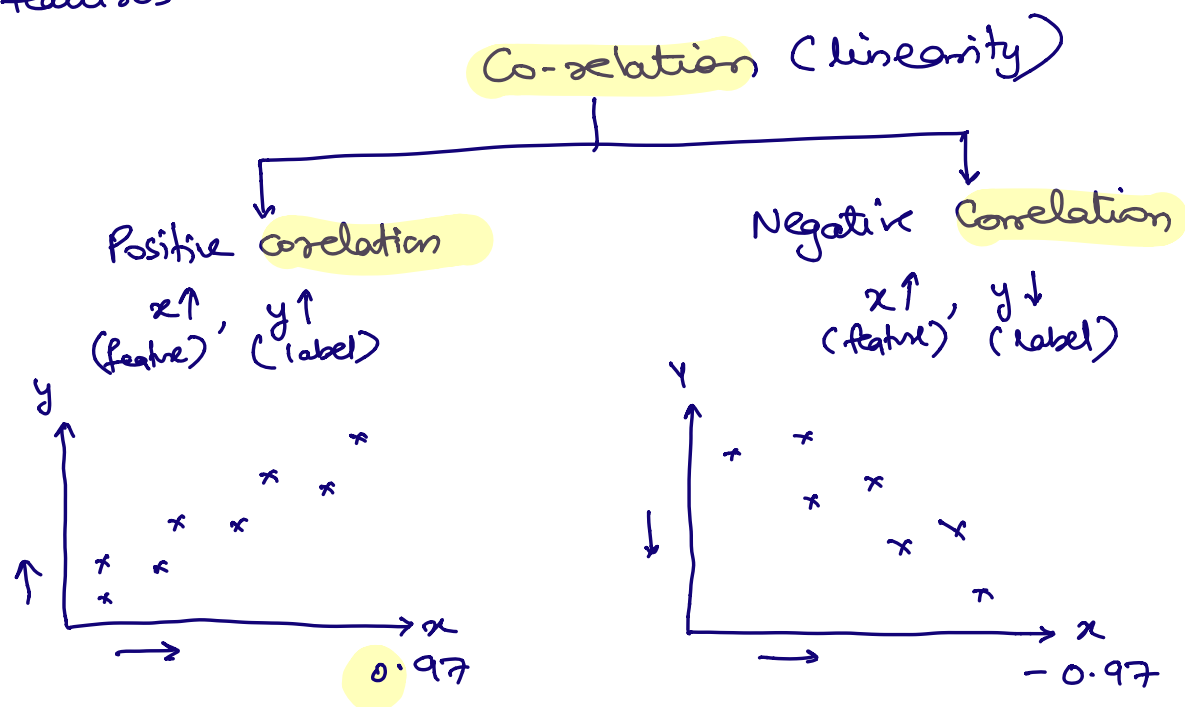
$$y = b_0 + b_1 x_1$$

intercept ↑   coeff of x_1

Use case: HR salary estimator
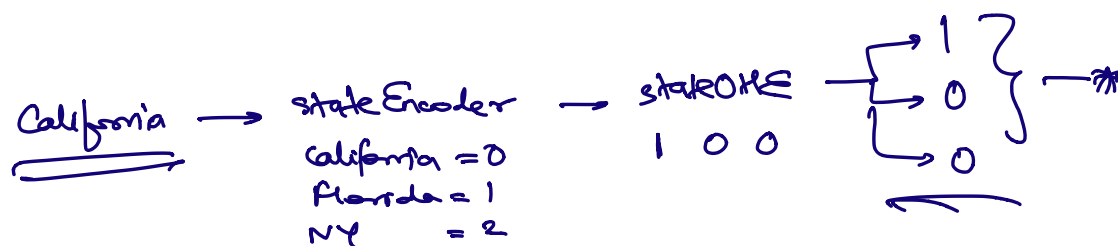
Feature — YearsExp
Label   — Salary

$$Salary = b_0 + b_1 (YearsExp)$$

Data must have some correlation. (linearity)

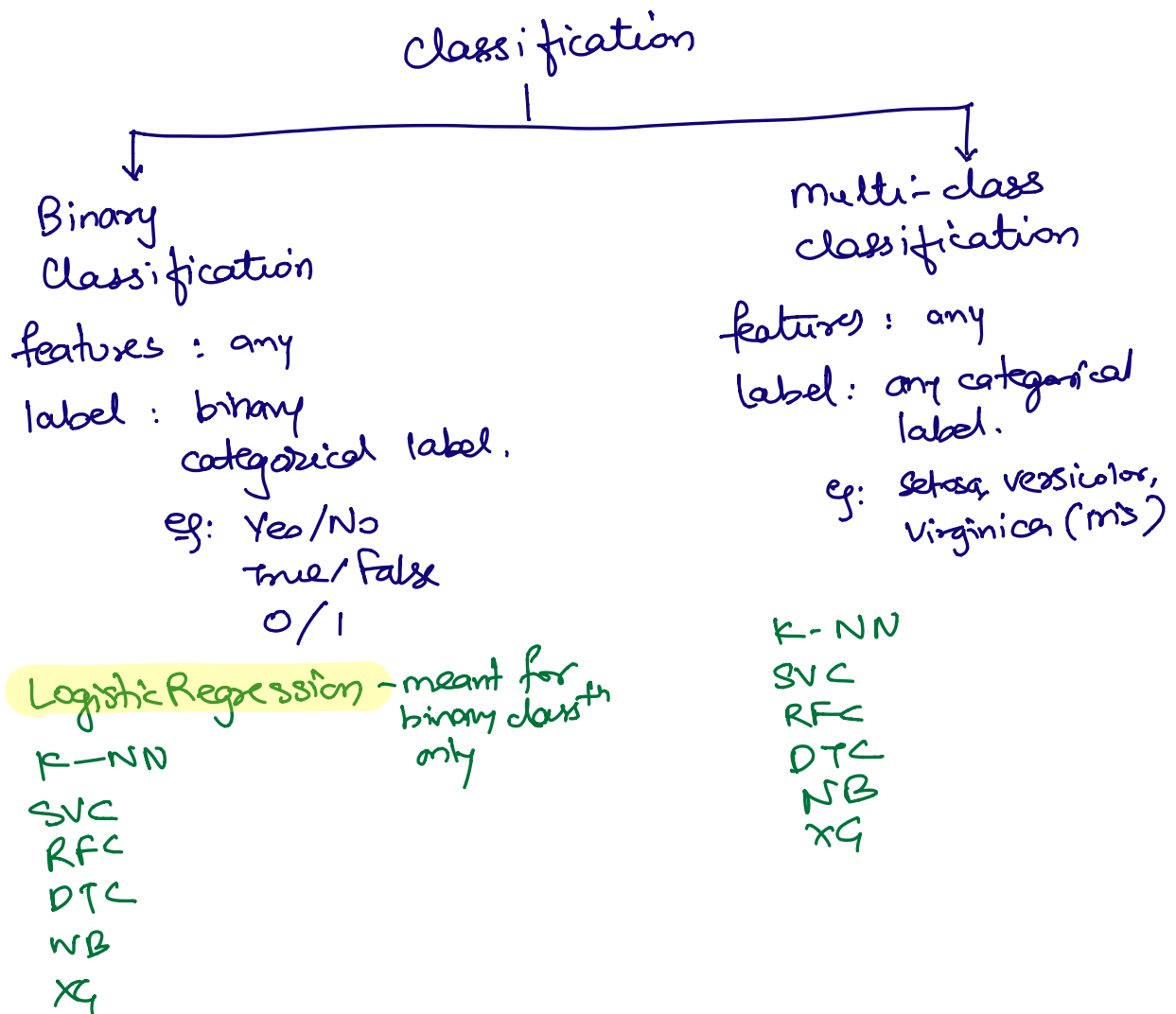There must exists correlation between label and each features.

Co-relation (linearity)

Positive correlation

$x\uparrow$, $y\uparrow$
(feature) (label)

0.97

Negative Correlation

$x\uparrow$, $y\downarrow$
(feature) (label)

-0.97

R&D, Adm, Marke, State

State encoder
State ONE

California → State Encoder → stateONE
California = 0   1 0 0
Florida = 1
NY = 2

Classification :
  features : any
  label : categorical

classification

Binary Classification

features : any

label : binary categorical label.

  eg: Yes/No
     True/False
     0/1

Logistic Regression — meant for binary class'th only

K—NN
SVC
RFC
DTC
NB
XG

multi-class classification

features : any

label : any categorical label.

  eg: setosa, vessicolor, virginica (ms)

K-NN
SVC
RFC
DTC
NB
XG

Logistic Regression

Linear Regression $+$ Sigmoid fn $\longrightarrow$ Logistic Regression

$y = b_0 + b_1 x_1$

$$\frac{1}{1 + e^{-y}}$$

y
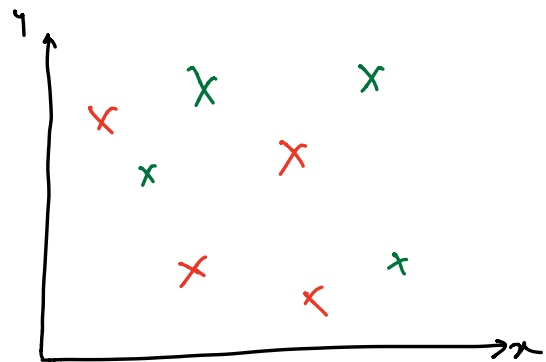
1 — ① $\quad$ 1 True $\quad$ x x x x

0.5 $\quad$ 0 false

① x x x x $\longrightarrow x_1$

$$\frac{1}{1 + e^{(b_0 + b_1 x_1)}}$$

K-NN (K-nearest neighbourhood).

y

1 $\longrightarrow$ x x x x

0 x x x x $\longrightarrow x$

<u>linear data</u>

y

x $\quad$ x $\quad$ x
x $\quad$ x $\quad$ x
x
x $\quad$ x $\quad$ +

$\longrightarrow x$

$\uparrow$ KNN
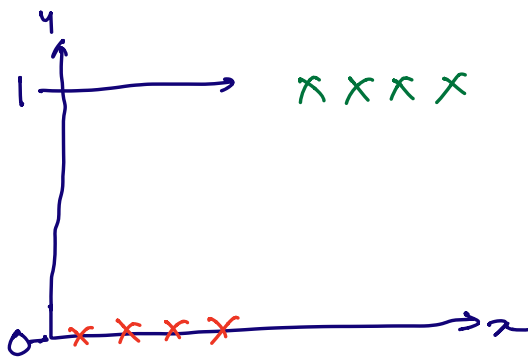
<u>non linear data.</u>

w.r.t Classification

k-NN
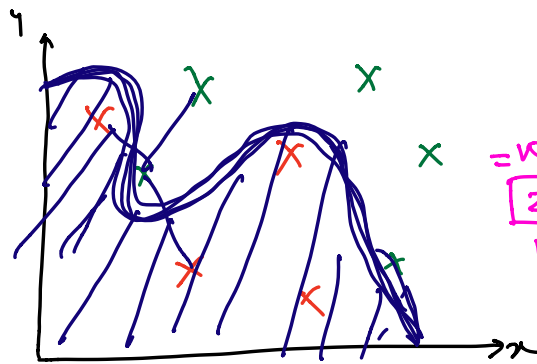
= voting   no. of neighbours. = 3

2 - green
1 - red

no. of unique labels
is 2   red ✓
          green ✓

Euclidean   distance formula   region