Creating a ML Program

① Load the data

② Pre-processing tasks

ML algo (sklearn)
- ① Your data must be complete
- ② Your data must be strictly numeric
- ③ Your data must be in the form of numpy array.

(a) Statistical summary. (describe)

(b) Check for missing data (info)

(c) Handle missing data (Imputation) → mean / median / mode / default

(d) Handling categorical data → Dummy variables

(e) Handling ordinal data → Encode the data and convert the ordinal data to discrete numeric.
* Dummy variables (domain & statistics)

③ Separate your data as features and label. Ensure the data is in the form of numpy array.

④ ⇒ Handle Numerical data. (Feature Scaling)

This step is applicable for features.

This step ensure your feature columns follow a common statistical scale. (sklearn)

① Rescale your feature manually by defining a range (0 to 1) (-1 to 1) (MinMaxScaler)

⇒ ② Standardize your data (mean = 0, std = 1) (StandardScaler)

③ Normalize your data. (manual method.... write code using numpy & scipy)

ⓐ L1 normalization → euclidean

ⓑ L2 normalization → manhattan

④ Feature Transformation

ⓐ Log Transformation ⓑ Trignometric Transformation

The above step ensures your scale is reduced, however your magnitude is unaffected.

Rule:- for ML, if your algorithm uses distance formula

len ← ⓐ Euclidean — *

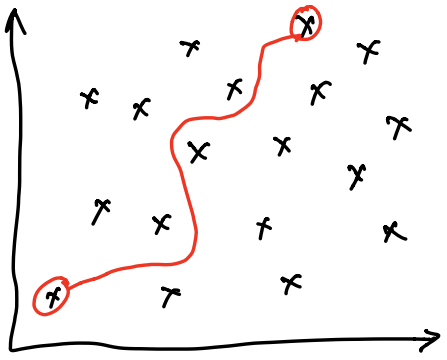
mse ← ⓑ Manhattan

ⓒ Minkowski

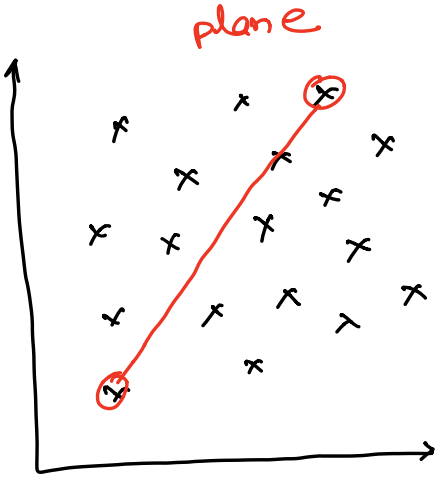
- k-means
- k-NN
- svm (optional)

[Gradient Descent]

for DL, feature scaling is mandatory for ANN.



① manhattan dist.



② Euclidean dist

magnitude remain same

100 → 1
200 → 2
300 → 3
400 → 4

scaling is reduced

$37 \Rightarrow 3.7 \times 10^{1} \Rightarrow 3.7e^{+1}$

exponential

$3.700\ e^{+1}$