

DAY: _____

DATE: _____

IDS-15

(20/03/24)

KNN \Rightarrow reg + classification.K-mean \Rightarrow clustering.

classification (supervised learning)

 \hookrightarrow label = categorical.KNN \Rightarrow classification + reg. \hookrightarrow no learning
(just store data during training) \hookrightarrow if numbers \Rightarrow % of yes/no \hookrightarrow actual - calculated = error. \hookrightarrow for reg \Rightarrow own exercise. \hookrightarrow for classification. \Rightarrow data is used as it is \Rightarrow lazy learner (learn while predicting). \Rightarrow k \Rightarrow usually odd value selected. \Rightarrow relatively better at binary classification rather than multi class.

- 1- Find distance from all pts
- 2- Sort distance
- 3- select top k values.

 \hookrightarrow usually euclidean distance used.

$$\sqrt{(7-3)^2 + (7-7)^2} = \sqrt{16+0} = 4, A \quad 3$$

$$\sqrt{(7-3)^2 + (4-7)^2} = \sqrt{16+9} = 5, A \quad 4$$

$$\sqrt{(3-3)^2 + (4-7)^2} = \sqrt{0+9} = 3, B. \quad 1$$

$$\sqrt{(1-3)^2 + (4-7)^2} = \sqrt{4+9} = 3.61, B. \quad 2$$

$\left. \begin{array}{l} 3, B \\ 3.61, B \\ 4, A \\ 5, A \end{array} \right\} \rightarrow B$

So point is B.

B

IDS-16

(22/03/23)

Unsupervised Learning:

Clustering \Rightarrow K-mean.

• \uparrow Inter cluster distance: distance b/w pts in diff clusters

\downarrow Intra cluster distance: " " " within cluster

$K \geq 2$

\hookrightarrow particular method used to find value of k .

(x, y) (x, y) $(x-u, y-y) \Rightarrow$ distance

1) $k=3$.

2) $c_1, c_2, c_3 \Rightarrow$ select randomly but have difference.

3) $\left\{ \begin{array}{l} \text{distance of each point from every centroid.} \\ \text{assign point to centroid with min distance.} \\ \text{Avg the distance, update the centroid.} \end{array} \right.$

iterations

\hookrightarrow repeat until no updation in centroid values

Elbow Method (for value of k)

\hookrightarrow generally used to find k .

Expectation-Maximization

$$\sum_{j=1}^c \sum_{i=1}^n \|x_i - c_j\|^2$$

\downarrow \downarrow
 data points no. of clusters

DAY: _____

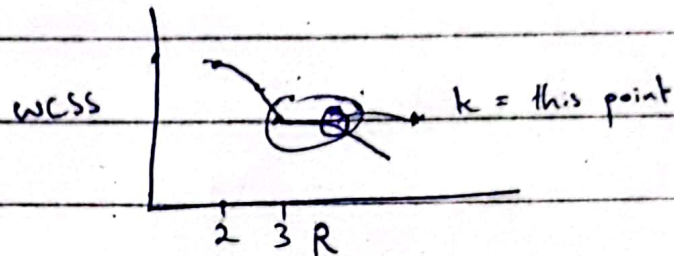
8:35
9:50
10:45
11:40

DATE: _____

1) Elbow Method:

• $\text{sum}((\text{distance of point from cluster-assigned})^2)$

↳ for diff values of k . (e.g 1 to 10)



2) Silhouette Method:

↳ usually used for verification.

a = distance of point from centroid it belongs.

b = distance // // it does not belong.

more +ve value \Rightarrow good no of k

more -ve value \Rightarrow worst no of k .

from -1 to +1.

IDS-17

(27/03/24)

Evaluation Metric

→ Confusion Matrix (for classification)

↳ False Alarm (FP)

overfit \Rightarrow Type 1 error \Rightarrow all FP

DAY: _____

TP FP
FN TN
DATE: _____

Confusion Matrix Example.

TP: 3 TN: 4 FP: 1 FN: 2

Positive class: Dog

3 1

2 4

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} = \frac{3 + 4}{3 + 4 + 1 + 2} = \frac{7}{10} = 70\%$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{3}{3 + 1} = 75\%$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{3}{3 + 2} = 60\%$$

IDS-18 (29/03/24)

Naive Bayes \Rightarrow collection of algos.

Conditional Probability:

\rightarrow 1) Assume features are independent.

Naive \Rightarrow Basic

\downarrow
equal contribution
on ~~by~~

$$P(y|x) = \frac{P(x|y) * P(y)}{P(x)}$$

$$P(y|x_1, x_2) = \frac{P(x_1, x_2 | y) * P(y)}{P(x_1, x_2)}$$

DAY: _____

DATE: _____

$$P(\text{write}|\text{infected}) = \frac{1}{4}$$

$$P(\text{write}|\text{clean}) = \frac{2}{3}$$

$$P(\text{exec}|\text{infected}) = \frac{4}{4}$$

$$P(\text{exec}|\text{clean}) = \frac{0}{3}$$

$$P(\text{large}|\text{infected}) = \frac{0}{4}$$

$$P(\text{large}|\text{clean}) = \frac{2}{3}$$

Apply Laplacian Smoothing

$$P(\text{Write}|\text{Infected}) = \frac{2}{5}$$

$$P(\text{Write}|\text{clean}) = \frac{3}{4}$$

$$P(\text{Exec}|\text{Infected}) = \frac{5}{5}$$

$$P(\text{Exec}|\text{clean}) = \frac{1}{4}$$

$$P(\text{Large}|\text{Infected}) = \frac{1}{5}$$

$$P(\text{Large}|\text{clean}) = \frac{3}{4}$$

$$P(\text{Read}) = \frac{5}{8}$$

$$P(\text{Write}) = \frac{4}{8}$$

$$P(\text{Exec}) = \frac{5}{8}$$

$$P(\text{Large}) = \frac{3}{8}$$

P

IDS-19

(03/04/24)

Spam-filter

$$P = p_1 * p_2 * p_n$$

multiply probability
of n words

$$1 - P \Rightarrow$$

Example:

$$P(S|W) = \frac{P(W|S) * P(S)}{P(W)}$$

$$P(s | \text{"Review vs Now"}) = ?$$

$$P(s) =$$

→ If probability is 0 in prior probabilities
then apply Laplace smoothing.

↓ Add 2 in denominator (total)

total records + = no of classes

$$P(\{1, 0, 1, 0, 0, 0\})$$

↓ Each index indicating presence/absence of word in table (for multiplication).

⇒ Unknown word ⇒ ignore.

⇒ If prob of spam > threshold ⇒ spam
spam 12.3% or spam > ham ⇒ spam
ham 87.67%.

$$0.0625 \times \frac{2}{6}$$

$$0.0044 \times \frac{4}{6} + 0.0625 \times \frac{2}{6}$$

IDS-20

(05/04/24)

Data Preprocessing:

↳ Most imp part of DS.

Data collection + preprocessing ⇒ 80% time.

DAY: _____

DATE: _____

- Remove Irrelevant Data:

$\frac{100\% \text{ same}}{(\text{variance} < 0)}$ or $\frac{100\% \text{ d/f}}{(\text{emails})} \Rightarrow \text{no info}$

Standardize Capitalization

↳ when working with text.

↳ so generally all text converted to lower case.

Convert Data Type:

Handle Missing Values

↳ some models don't work on missing values.

• MCAR

↳ no relationship of missing data

↳ why missing? X

• MAR

↳ you can identify

why data is missing?

↳ missing within range

• MNAR

↳ ~~identification~~ ✓

↳ investigate

↳ ~~can~~ not related with collected data

Transformation:

- Label encoding \Rightarrow biasness to large values.

- One hot encoding \Rightarrow vector to each category.

↳ No one hot encode for label

↳ only for feature

DAY: _____

DATE: _____

Feature Scaling:

- Standardization

↓

mean = 0

↓

variance = 1

- Normalization

(0-1)

(prefer) as we assume data is normalized.