Where Date Who Why

① You are given the ground truth labels of the dataset and the predicted values obtained with a clustring algorithm

$y\_true = [\underset{A}{1}, \underset{B}{1}, \underset{C}{2}, \underset{D}{2}, \underset{E}{3}, \underset{F}{3}]$

$y\_predict = [\underset{A}{2}, \underset{B}{2}, \underset{C}{1}, \underset{D}{1}, \underset{E}{1}, \underset{F}{1}]$

Compute the Rand Index [RI] between y_true and y_pred

Remember that the Rand Index is given By:

$$RI = \frac{a+b}{DP} = \frac{TP + TN}{DP}$$

where:

TP (a) is the number of pairs of elements (without ordering) that are in the same set in y_true and in the same set in y_pred.

TN (b) is the number of pairs of elements (without ordering) that are in the difference set in y_true and in different y_pred.

(DP): is the total number of possible pairs in the Dataset (without ordering)

			TP	TN
AB	1 1	2 2	✓	×
AC	1 2	2 1		✓
AD	1 2	2 1		✓
AE	1 3	2 1		✓
AF	1 3	2 1		✓
BC	1 2	2 1		✓
BD	1 2	2 1		✓
BE	1 3	2 1		✓
BF	1 3	2 1		✓
CD	2 2	1 1	✓	×
CE	2 3	1 1		×
CF	2 3	1 1		×
DE	2 3	1 1		×
DF	2 3	1 1		×
EF	3 3	1 1	✓	×
			3	8

$$TP + TN = \frac{3+8}{15} = \frac{11}{15} = .1733$$

A B C D E F

$$\binom{6}{2} = \frac{6!}{2 \cdot 4!} = \frac{6 \times 5 \times 4!}{2 \cdot 4!} = 15$$

Where Date Who Why

② You are given a dataset containing 300,000 records divided into 3 classes with the following schema.

Id | value_0 | ..... | value_100 | Label

You have to build a classification model to avoid overfiting which is the best strategy to turn and test the classifier?

ⓐ `X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=1)`

ⓑ `tscv = timeseriessplit(n_splits=4)`
`for train, test in tscv.split(x)`

✓ⓒ `X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=42)`
`kf = kfold(n_split=4)`
`for train, test. in kf.split(X_train):`

ⓓ `loo = leaveoneout()`
`for train_Index, test_index in loo.Split(x)`

Where Date Who Why

③ You have to compare the daily temperature differences between two cities, sample by sample the trends are showed in the picture below.

Question 1: which kind of preprocessing would you choose?

- ○ consider ① the difference between the series when missing values
- ○ Fill missing values with ①
- ✓ Fill missing values with previous value
- ○ Add zero padding after temperature²

Question 2: Given the strategy that you have selected, calculate the mean of daily differences calculated with the following numpy Function.

```
avg_daily_diff = np.mean(temp1 - temp2)
```

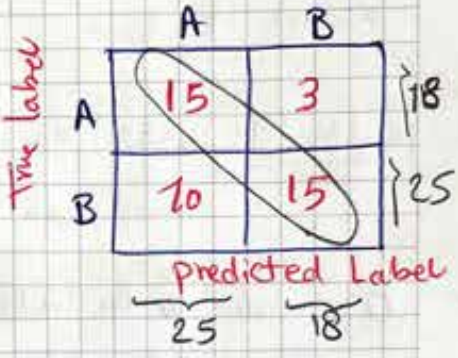
Day	1	2	3	4	5	6	7	8	
temp1	20	23	23	24	25	23	23	23	
temp2	21	22	22	22	24	24	22	22	
diff	-1	+1	+1	+2	+1	-1	+1	+1	⇒5 ⇒ mean = $\frac{5}{8}$ = .625

Where Date Who Why

④ You are given the confusion matrix in the figure below obtained after the validation phase of classifier

Q_1 = Compute the accuracy of the classifier

Accuracy: $\frac{30}{43} = 0.6977$



Q_2 = Compute the precision for class with Label A.

precision (A): $\frac{15}{25}$ → predicted as class "A" ⇒ $\frac{15}{25} = .16$

Recall (A) = $\frac{15}{18}$ → Actually is belong class "A"

$F = \frac{2PR}{P+R}$ ⇒

⑤ You are given a decision tree trained on a dataset with categorical features only.

which Statement is true:

(a) Each internal node of a decision tree can have at most "C" children, where "C" is equal to the number of distinct values of the attribute associated with the internal node.

(?)

Where Date Who Why

⑥ The matrix in the figure below represents the distances between data points, while the table on the right lists the ground truth labels. points a, b, c, d represent training sample while x_1, x_2 are test Samples.

	a	b	c	d	x_1	x_2
a	0	1	2	3	2	2
b	1	0	2	1	2	5
c	2	2	0	3	4	2
d	3	1	3	0	1	4
x_1	2	2	4	1	0	4
x_2	2	5	2	4	4	0

train test

Sample	ground-truth
a	True
b	Tuve
c	False
d	False

Q_1 if the value of K is set to 2, what is the number of distances computed by the k-nearest neighbor algorithm to obtain the label of x_1 ⇒ The number of computed distance is 4

Q_2 ⇒

Where Date Who Why

⑦ Given the transactional dataset shown in the figure below. Apply the Apriori Algoritm to extract all frequent itemsets. The value of minSup is 2.

An itemset is considered to be frequent if it support count is equal to or higer than the misup.

Transactions

ab
cde
bc
ab
acd
abce
acde
de
ac
de

Question-1) list all of frequent Items having lenght 2,

Question2) list all itemset of lenght 3 that have been generated By Apriori after the join and prune steps before counting their support on the database.

Question 3) list all frequent itemsets that are not closed

$Q_1 \Rightarrow$

A	6
B	4
C	6
D	5
E	5

$\Rightarrow$

AB	3
AC	4
AD	2
AE	2 ×
BC	2
BD	Φ
BE	1
CD	3
CE	3
DE	4

$Q_2 \Rightarrow$

ABC	1
ABD	Φ
ABE	1
ACD	2
ACE	2 ×
ADE	1
CDE	2

$Q_3 \Rightarrow$ ad=2

ae=2

Where Date $A \to B$ Who Why

Support

$$sup = \frac{\text{itemset ها تعداد تکرار}}{\text{تعداد کل سطرها}} = \frac{\#\{A,B\}}{|T|}$$

$sup \geq minsup$

Confidence

$$conf = \frac{\text{itemset ها تعداد تکرار آیتم}}{\text{تعداد آیتم مورد نظر}} = \frac{sup(A,B)}{sup(A)}$$

$conf \geq min\ conf$

* Frequent itemset : is a itemset whose support is greater than or equal to minsup threshold

Rand Index

$$Rand\ Index = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{TN + TP}{TN + FP + TP + FN}$$

برای محاسبه RI ابتدا جفت ها را می سازیم (هر دو را در مجموعه خوشه ها می سنجیم نه یا مجموعه دیگر)

$TP = 1 \Rightarrow y_t = 1 \ , y_p = 1$

$TN = 1 \Rightarrow y_t = 0 \ , y_p = 0$

$TP = TN = 0 \Rightarrow y_t \neq y_p$

		دسته بندی واقعی: در دو دسته مجزا	دسته بندی واقعی: در یک دسته
	نتیجه خوشه بندی برای زوج (x,y)	در دو دسته مجزا	در یک دسته
خوشه بندی	در یک خوشه	FP f_{01}	TP f_{11}
خوشه بندی	در دو خوشه مجزا	TN f_{00}	FN f_{10}

در RI تعداد کلاسترها hyper parameter نیست

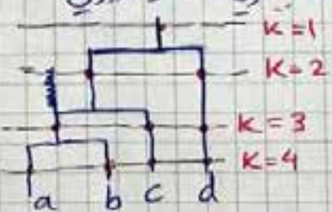
Hierarical clustering:

Single linkage (min)

complete linkage (max)

در ماتریس دنبال کمترین مقدار هستیم ← نزدیک‌ترین فاصله

در ماتریس دنبال بیشترین مقدار هستیم ← بیشترین فاصله یا دورترین



Where Date Who Why

Naive Bayas classifier

$X \in R^n \qquad X = (x_1, x_2, \dots, x_n)$

$P(class|x) = P(x_1|class) * P(x_2|class) * \dots * P(class)$

تعداد آن کلاس در x_1 (تعداد شکل مورد نظر) — مجموع کل شکل مورد نظر (△, *, ○)

△
○
X

Manhatan distance

$$dist(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2|$$

DBSCAN :

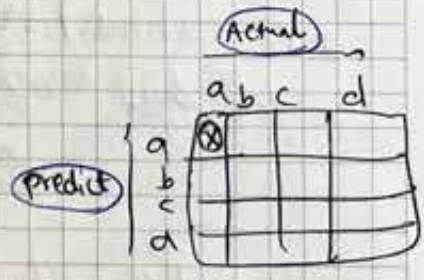
Core : Points with at least minpoints neighbors within a distance

border : at least a core Point in the neigh bor hood

noise : neither a core nor a border

accuracy score in classification

$$Accuracy = \frac{\text{جمع عناصر روی قطر اصلی}}{\text{مجموع کل عناصر ماتریس}} = \frac{\text{تعداد کل درست ها}}{\text{تعداد کل عناصر}}$$



$$Recall(A) = \frac{\text{عددی قطر اصلی سطر آن کلاس}}{\text{مجموع عناصر روی سطر آن کلاس}} = \frac{\text{تعداد درست های کلاس A}}{\text{تعداد کل عناصر متعلق به کلاس A}} \Rightarrow P_{truth}$$

$$Percision(A) = \frac{\text{عددی قطر اصلی آن کلاس}}{\text{مجموع عناصر روی ستون آن کلاس}} = \frac{\text{تعداد درست های کلاس A}}{\text{تعداد کل عناصر انتخاب شده کلاس A}} \rightarrow P_{predict}$$

$$F1\_Score = \frac{2P*R}{P+R}$$

Where Date Who Why

measures of node impurity

classification:

* Decision Tree:
  1) Hunt's Algorithm
  2) CART
  3) ID3, C4.5, C5.0
  SLIQ, SPRINT

⇒
- Gini index
- Entropy
- misclassification error

(X) interpretable

Gini split ⇒

تعداد رکوردها در فرزند i ام

$$gini_{split} = \sum_{i=1}^{K} \frac{n_i}{n} \times Gini(i)$$

تعداد رکوردها در نود P ام

$$Gini(i) = 1 - \sum [p(j|t)]^2$$

تعداد تکرار کلاس j در نود t

* Rule-based classifer

*1) mutually exclusive:
1) دو Rule نمی‌توانند همزمان درست باشند.
2) هر رکورد باید حداکثر تحت پوشش یک قانون (Rule) باشد.

*2) exhaustive rules:
هر ترکیب ممکن از مقادیر ویژگی‌ها باید در نظر گرفته شود
2) هر رکورد باید حداقل تحت پوشش یک قانون باشد.

* Associative classification

* K-nearest neighbor

(Y) incremental

Euclidean distance ⇒ $d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$ Not (X)

$$w = 1/d^2$$

Terms occuring frequently in a Single document but rarely in the Whole collection are prefered.

Where | Date | Who | Why

*Bayesian classification

Fully incremental

not (X)

$\begin{cases} C: \text{any class label} \\ X: \langle x_1, \dots, x_k \rangle \text{ record to be classified} \end{cases}$

Max

① compute $P(C|X)$ for all classes → $(P(C|X))$

② Assign X to the class with maximal $P(C|X)$

$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$

N_C/N

برای هر کلاس ثابت است.

$$P(X_1, X_2, \dots X_k|C) = P(X_1|C) * P(X_2|C) \dots * P(X_k|C)$$

$P(X_k|C)$

$$P(X_k|C) = \frac{|X_{kc}|}{N_C}$$

تعداد instance هایی که مقدار X_k دارند و در آن k متعلق به کلاس c هستند.

برای داده‌های پیوسته ⇒ probability distribution

* Support vector machines (SVM)

Not (Y)

Not (X)

*Artificial neural networks

Not (Y)

Not (X)

tf_idf

tf_idf of term t in document d of collection D (consist of m document)

$$tf\_idf(t) = freq(t,d) * \log(m / freq(t,D))$$

② A collection of heterogenous doc

Suitable for: ① A Single Doc consisting of many sections or subsections

Where | Date | Who | Why

clustering Algorithms:

1) K-means and its variants ⇒ closeness ⇒ Euclidean distance, cosine, correlation, similarity

2) Hierarchical

3) Density-based

tf-df suitable for: ① Single Documents or Parts of document with homogeneous content. ② A collection of documents ranging over the same topic.

Where Date Who Why

Minkowski Distance

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where n is number of dimensions.

if $r=1$ ⇒ Manhatan Distance ⇒ $dis(P_1,P_2) = |x_1 - x_2| + |y_1 - y_2|$

if $r=2$ ⇒ Euclidean distance $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_k - y_k)^2}$
dissimilarity ↵

if $r \to \infty$ ⇒ Supremum ⇒ maximum difference between any component of the vectors.

tf-df

tf-df of term t in document d of collection D

$tf\_df(t) = freq(t,d) * \log(freq(t,D))$

Terms occuring frequently both in a single document and in the whole collection are preferred.

2022/06/13 امیرکبیر

۱) Where Date Who Why

Fp-growth

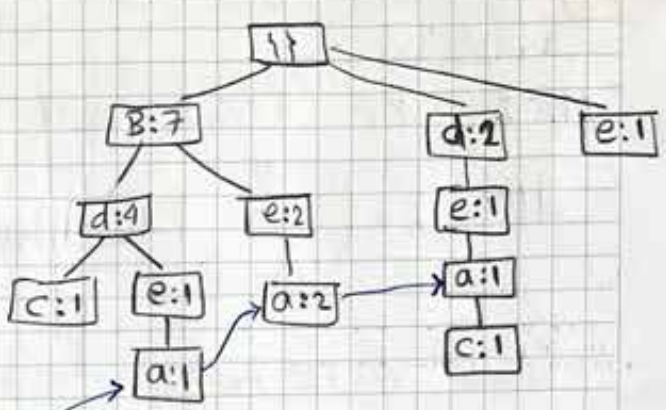
BcD	BDc
ACDE	DEAC
BD	✓
B	✓
ABE	BEA
E	✓
ABE	BEA
BD	✓
D	✓
ABDE	BDEA

① A-CPB
② A-CHT
③ AD-CPB

A	4
B	7
C	2
D	6
E	5

ترتیبی

B	7
D	6
E	5
A	4
C	2



Bde	1	دوبار
Be	2	count بودن
de	1	دارد B,E

B	3
D	2
E	4

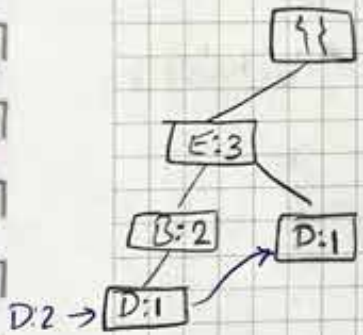
A-CHT

E	4
B	3
D	2

ترتیبی

A-CPB

EBD	1
EB	2
ED	1



EB	1
E	1

E	2
B	1

حذف شده

abd D-CPB=? {ab:1, ae:2, aeb:2}

abde D-CHT=? {a:5, e:4, b=3}

ce DB-CPB=? {a:1, ae:2}

abe

acde

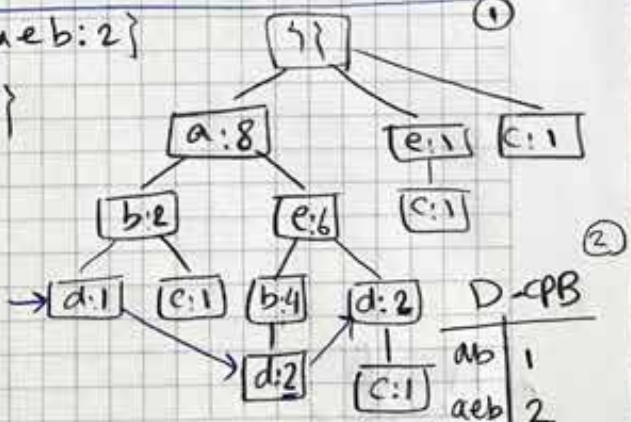
abde

abe

c

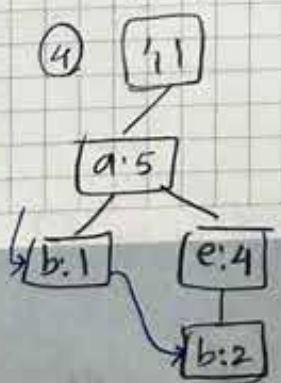
ade

abc



D-CPB

ab	1
aeb	2
ae	2



DB-CPB

a	1
ae	2

D-CHT

a	5
e	4
b	3

(ab(1), aeb(2))
1+2=3

حل تمرین سوال 10

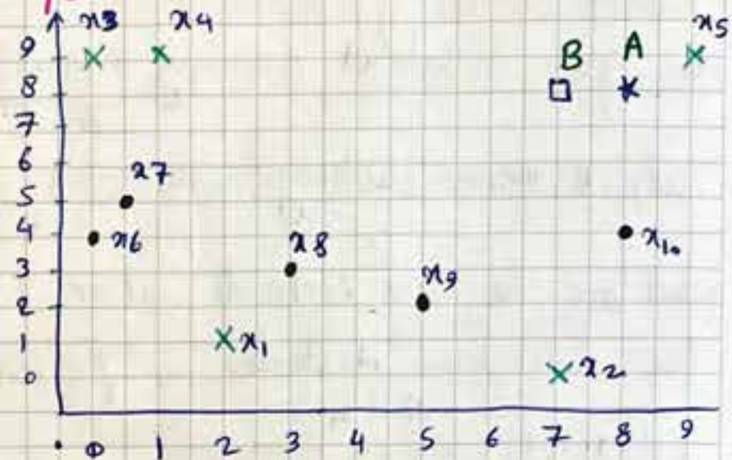
2) امتحان 2021,9,6

$\rightarrow \max_{i=1}^{n} |X_i - Y_i|$ where X,Y is two Points

Where Date Who Why

KNN Algorithm (with chebyshev distance), K=3

	$d(x_i)$	A (8,8)	B (7,8)
X	x_1 (2,1)	7	7
X	x_2 (7,0)	8	8
X	x_3 (0,9)	8	7
X	x_4 (1,9)	7	6
X	x_5 (9,9)	1 ✓	2 ✓
O	x_6 (0,4)	8	7
O	x_7 (1,5)	7	6
O	x_8 (3,3)	5 ✓	5 ✓
O	x_9 (5,2)	6	6
O	x_{10} (8,4)	4 ✓	4 ✓



① هر نقطه روی نمودار یک نام خاص می‌دهیم (X, X2, ..., Xn)

② جدول دو در دو را براساس مختصات‌ها طراحی می‌کنیم

③ محاسبه فاصله (xi , A), (xi , B) براساس $\max_{i=1}^{n} |X_i - Y_i|$

④ برای تعیین نزدیک‌ترین همسایه در هر ستون (K=3)
سه تا از کمترین ها را بررسی می‌کنیم و در آخر اساس همسایه را اختصاص می‌دهیم.

Question 1)

$A \Rightarrow$ 2 O, 1 X $\Rightarrow$ class (A) = circle

$B \Rightarrow$ 2 O, 1 X $\Rightarrow$ class (B) = circle

What is the Label assigned to A,B when neighbors' votes are weighted uniformaly?

چون در صورت سوال گفته weighted uniformly

Question 2) What are the class probabilities for A, B, if the vote cast By each neighbor is $w = \frac{1}{1+distance}$

A		W	class label
x_5	1	$\frac{1}{2}$	X
x_8	5	$\frac{1}{6}$	O
x_{10}	4	$\frac{1}{5}$	O

$\frac{1}{2}$, $\frac{11}{30}$

$P(X|A) \Rightarrow \frac{\frac{1}{2}}{\frac{11}{30}+\frac{1}{2}} = \frac{\frac{1}{2}}{\frac{26}{30}} = \frac{15}{26}$

$P(O|A) \Rightarrow \frac{\frac{11}{30}}{\frac{11}{30}+\frac{1}{2}} = \frac{\frac{11}{30}}{\frac{26}{30}} = \frac{11}{30}$

B		W	class label
x_5	2	$\frac{1}{3}$	X
x_8	5	$\frac{1}{6}$	O
x_{10}	4	$\frac{1}{5}$	O

$\frac{11}{30} \Rightarrow$

$P(X|B) = \frac{\frac{1}{3}}{\frac{1}{3}+\frac{11}{30}} = \frac{\frac{1}{3}}{\frac{21}{30}} = \frac{30}{3 \times 21} = \frac{10}{21}$

$P(O|B) = \frac{\frac{11}{30}}{\frac{1}{3}+\frac{11}{30}} = \frac{\frac{11}{30}}{\frac{21}{30}} = \frac{11}{21}$

روش حل سوال

$\hookrightarrow P(C|X) = \frac{vote_c(X)}{\sum_{i \in C} vote_i(X)} \Rightarrow vote_Y(X) \Rightarrow$ unormalized vote assigned to class Y for sample X

حل سوال 3 امتحان

3) 2021, 2, 10

Where Date Who Why

K-mean clustering - Centroids

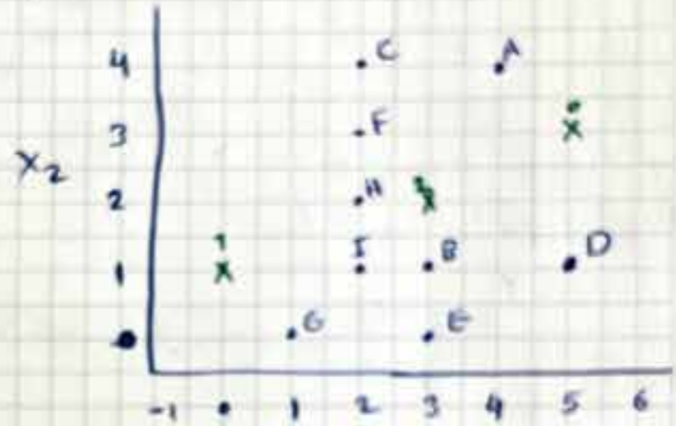
You given a dataset containing 9 points in 2 dimension (X_1, X_2) A - I

Apply K-means clustering with K=3 → (9) Blue points ⇒ (A-I)
(3) red stars ⇒ centroids (0, 1, 2)

What are the new centroids after 1 iteration of K-means algoritm?

Use Euclidean distance ⇒ $\sqrt{\Sigma(x_i - y_i)^2} \Rightarrow \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

	dis(x_i)	C_0 (5,3)	C_1 (0,1)	C_2 (3,2)	نزدیک ترین مرکز
A	(4,4)	$\sqrt{2}$	5	$\sqrt{5}$	C_0
B	(3,1)	$\sqrt{8}$	3	1	C_2
C	(2,4)	$\sqrt{10}$	$\sqrt{13}$	$\sqrt{5}$	C_2
D	(5,1)	2	5	$\sqrt{5}$	C_0
E	(3,0)	$\sqrt{13}$	$\sqrt{10}$	2	C_2
F	(2,3)	3	$\sqrt{8}$	$\sqrt{2}$	C_2
G	(1,0)	5	$\sqrt{2}$	$\sqrt{8}$	C_1
H	(2,2)	$\sqrt{10}$	$\sqrt{5}$	1	C_2
I	(2,1)	$\sqrt{13}$	2	$\sqrt{2}$	C_2



⇒

$C_0 \Rightarrow \{A, D\} = \frac{1}{2}\left((4,4) + (5,1)\right) = \left(\frac{9}{2}, \frac{5}{2}\right)$

$C_1 \Rightarrow \{G\} = (1,0)$

$C_2 \Rightarrow \{B, C, E, H, I, F\} \Rightarrow \frac{1}{6}\left((3,1) + (2,4) + (3,0) + (2,2) + (2,1) + (2,3)\right) = \left(\frac{14}{6}, \frac{11}{6}\right)$

① گام اول فاصله هر نقطه را از مرکز کلاستر مربوط به دست می آوریم

② نزدیک ترین مقدار را در بین کلاسترها جهت انتخاب می کنیم.

4) تمرین با سوال ۸ امتحان
تمرین استاد سرمدی

Where Date Who Why

A binary classifier is trained to separate between images of cats and dogs. The test set used to evaluate this model is balanced. with 10,000 images of cats and 10,000 images of dogs.
The classifer only predicts 50 images as being cats.
All those predictions are Correct.
What can said about this classifier?

Cats = 10
dogs = 10
predict (cat) = 5

Dog (10) -5 → تشخیص داده شده dog که در واقع cat است
Cat (10) +5

$\Rightarrow$ Recall (cat) $= \frac{50}{10,000} =$

$\Rightarrow$ Recall (dog) $= \frac{10,000}{10,000} = 1$

$\Rightarrow$ It has high Recall for the class "dog"

There is 3 classes (~~Squares~~, Circle, triangle) After training. (آزمون ۸)
where the label assigned by C is given by the background color of the regin in.

predicted class = ■ ← = ● ← = ▲ ←

supole that the symbols in the image are test points whose shap represents. the respective ground truth label.

$P_{\square} = \frac{3}{4}$
$P_{\triangle} = \frac{5}{5} = 1$
$P_{O} = \frac{8}{11}$

$R_{\square} = \frac{3}{5}$
$R_{\triangle} = \frac{5}{6}$
$R_{O} = \frac{8}{9}$

Accuracy $\Rightarrow \frac{\square\ O\ \triangle}{\frac{3+8+5}{20}} = \frac{16}{20} = 1$

۲۰۲۲/۰۲/۱۱ جلسه ۱

4)

Where | Date | Who | Why

The pearson correlation between two variables: $X = \{x_1, x_2, \dots, x_n\}$ having mean $\bar{x}$ and $y = \{y_1, y_2, \dots, y_n\}$ having mean $\bar{y}$ is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Two Sequences A and B have been sampled from a distribution. The following Seq are included below both in numerical and graphical:

$B = 0.5A^2 + 5$ $\quad A \Rightarrow X = y$

A	B
-5	17,5
-4	13
-3	9.5
-2	7
-1	5,5
0	5
1	5.5
2	7
3	9,5
4	13
5	17,5

what is the pearson correlation between A, B (r_{AB}) = ?

$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 0 \Rightarrow$ صفر خطی $\Rightarrow r_{AB} = 0$

$\bar{x} = 0$

$\bar{y} = 10$ $\Rightarrow$

$(-5 \times 17,5) = -37,5$

$+(-4 \times 13) = -12$

$+(-3 \times 9,5) = +1,5$

$+(-2 \times -3) = +6$ $\Rightarrow$

$+(-1 \times -0,5) = +0,5$

$+(0 \times 5) = 0$

$+(1 \times 5,5) = +5,5$

$+(2 \times -3) = -6$

$+(3 \times -15) = -1,5$

$+(4 \times 3) = 12$

$+(5 \times 7,5) = 37,5$

5)

Where Date Who Why

K-mean algorithm - centroids

$K=5 \leftarrow A, B, C, D, E, F$

$A = \frac{1}{2}(1+0) = \boxed{\frac{1}{2}}$

$B = \boxed{2}$

$C = \boxed{3}$

$D = \boxed{5}$

$E = \frac{1}{5}(\underbrace{7+7+7}_{27} + \underbrace{9+9}_{18}) = \boxed{\frac{39}{5}}$

		(1) A	(2) B	(3) C	(5) D	(8) E	نزدیکترین گروه
0	x_1	(1)	2	3	5	8	A
1	x_2	(0)	1	2	4	7	A
2	x_3	1	(0)	1	3	6	B
3	x_4	2	1	(0)	2	5	C
5	x_5	4	3	2	(0)	3	D
7	x_6	6	5	4	2	(1)	E
7	x_7	6	5	4	2	(1)	E
7	x_8	6	5	4	2	(1)	E
9	x_9	8	7	6	4	(1)	E
9	x_{10}	8	7	6	4	(1)	E

7) Cosine Similarity - one hot encoding - Minkowski

Where | Date | Who | Why

The following definitions refer to any two vectors $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n) \in R^n$

The cosine similarity is defined as:

$$\cos(a,b) = \frac{a \cdot b}{\|a\| \|b\|}$$

The Minkowski distance of order p between is defined as:

$$D_p(a,b) = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{1/p}$$

You are given a categorical dataset with two independent features. X_1 and X_2. The two features can assume one of 13 and 7 distinct values respectively. more specifically.

$X_1 = (n_{00}, n_{01}, n_{02}, \ldots, n_{12})$

$X_2 = (m_{00}, m_{01}, \ldots, m_{06})$

From this Dataset two Records $r_1 = (n_{04}, m_{06})$ and $r_2 = (n_{11}, m_{01})$ are extracted.

① what is the Cosine Similarity between $r_1, r_2 \Rightarrow \cos(r_1, r_2)$.

$n_{04} = (0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$ (positions 0 1 2 3 4 5 6 7 8 9 10 11 12)

$n_{11} = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0)$

$m_{06} = (0\ 0\ 0\ 0\ 0\ 0\ 1)$

$m_{01} = (0\ 1\ 0\ 0\ 0\ 0\ 0)$

$\cos(r_1, r_2) = 0$?

② what is the Minkowski distance of order 1 between r_1, r_2?

$$d_1(r_1, r_2) = \sum_{i=1}^{2} |a_i - b_i| = 4 \quad |1-0| + |0-1| + |0-1| + |1-0|$$

$D_1(r_1, r_2)$

③ $\quad 2 \quad 2 \quad 2 \quad 2$

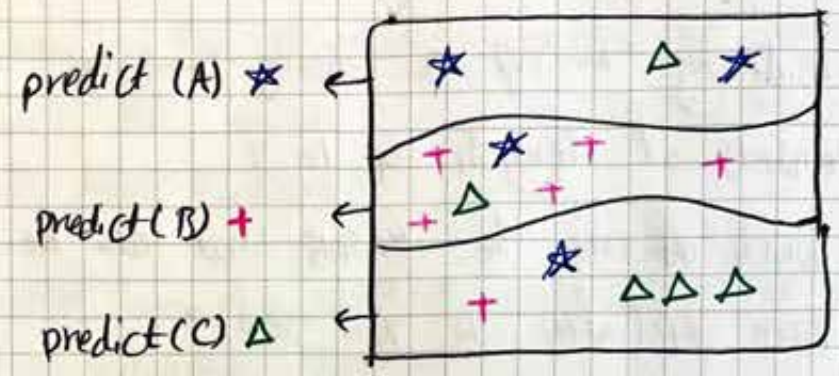
$D_2(r_1, r_2)$

$$d_2(r_1, r_2) = \sqrt{\sum_{i=1}^{2} |a_i - b_i|} = \sqrt{4} = 2$$

8)

Where | Date | Who | Why

Random forest classifier — Percision / Recall



$P(A) = \frac{2}{3}$ $\quad R(A) = \frac{2}{4} = \frac{1}{2}$

$P(B) = \frac{5}{7}$ $\quad R(B) = \frac{5}{6}$

$P(C) = \frac{3}{5}$ $\quad R(C) = \frac{3}{5}$

9)

Where Date Who Why

Apriori Algorithm

Considering minSup = 2 ⇒ only Consider the extraction of itemsets of length up to 3

1) which are the itemsets that can be purned during the purn step (after the join step)?

2) " " " " " " that are purned after computing their support count?

ADE
AB
BCD
BDE
ABC
AE
CD
ABE
ABCDE
ABDE

A	7
B	7
C	4
D	6
E	6

AB	5
AC	2
AD	3
AE	5
BC	3
BD	4
BE	4
CD	3
CE	1
DE	4

minSup > X

ABC	2
ABD	2
ABE	3
ACD	1
ACE	X
ADE	3
BCD	2
BCE	X
BDE	3

⇒ ACD < minSup ⇒

ABCD	X
ABCE	X
ABDE	✓

① {ACE, BCE, ABCD, ABCE}

② {CE, ACD}

11)

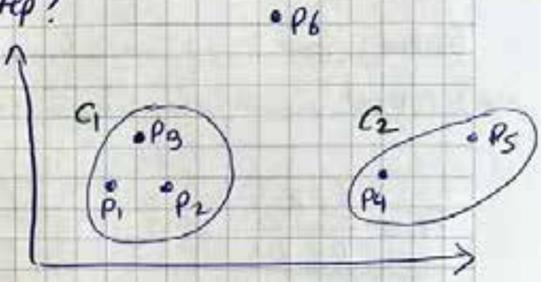
Where Date Who Why

clustering, Eucidean distance - cluster Similarity

		x	y
C_1	P_1	2	2
	P_2	4	2
	P_3	3	5
C_2	P_4	11	2
	P_5	13	24
C_3	P_6	8	7

① What is the proximity matrix at this step write the answer as distances between Pairs of points / clusters?

② What are the clusters obtained after applying an aditional step?



$$C_1 = \frac{1}{3}\begin{pmatrix}(2,2)\\(4,2)\\(3,5)\end{pmatrix} = (9/3, 9/3) = (3,3)$$

$$C_2 = \frac{1}{2}\begin{pmatrix}(11,2)\\(13,4)\end{pmatrix} = (24/2, 6/2) = (12,3)$$

$P_6 = (8,7)$

① هر دو در مرحله بخش اول جواب یکی است

② در نظر می‌گیریم که P6 به کدام مرکز کلاستر نزدیک‌تر است و پس آن را جزو آن کلاستر قرار دادم.

① $\text{distance}(C_1, C_2) = \sqrt{9^2} = 9$

$\text{distance}(C_1, P_6) = \sqrt{41} = 6{,}403$

$\text{distance}(C_2, P_6) = \sqrt{32} = 5{,}657$ ✓

2) $C_1 \Rightarrow (P_1, P_2, P_3)$

$C_2 \Rightarrow (P_4, P_5, P_6)$

Where | Date | Who | Why

2) classification:

Given the following confusion matrix

	a	b	c	d
a	10	0	0	0
b	0	4	0	4
c	0	4	10	0
d	0	2	0	6

(actual / predicted)

Σ 10 10 10 10 $\Rightarrow$ 40

$Q_1 \Rightarrow$ Compute the accuracy score.

$Q_2 \Rightarrow$ compute F-measure (F1) of class b:

$$\text{Accuracy} = \frac{10+4+10+6}{40} = \frac{30}{40} = \frac{3}{4}$$

① جمع عناصر قطر اصلی تقسیم بر جمع کل عناصر ماتریس درهم ریختگی داده شده است (میزان صحت)

② جمع عناصر خارج از قطر اصلی هم برابر جمع نادرست (Incorrect) است.

$$F_1 = \frac{2PR}{P+R} \Rightarrow \frac{2 \times \frac{2}{5} \times \frac{1}{2}}{\frac{2}{5}+\frac{1}{2}} = \frac{\frac{2}{5}}{\frac{9}{10}} = \frac{20}{45} = \boxed{\frac{4}{9}}$$

$$P(B) = \text{percision}(B) = \frac{4}{10} = \boxed{\frac{2}{5}}$$

$$R(B) = \text{Recall}(B) = \frac{4}{8} = \boxed{\frac{1}{2}}$$

3) Regression: multi Regression Pipeline

Given the following dataset with 2 features (x_0, x_1) and 3 data points: $x = [[2,4],[1,2],[2,0]]$

Apply to x the following multi regression pipeline

1) Feature extraction step $\Rightarrow [x_0, x_1, x_0^2, x_1^2, x_0 x_1]$

2) Regression parameters (to be applied on the extracted features)

$B: [0, 2, 0, 1, 1/2]$, Bias = 1

$Q_1 \Rightarrow$ what is the output vector with the predictions? Y_pred = [?]

$Q_2 \Rightarrow$ Given the ground truth predictions $\Rightarrow$ y_truth = [28, 9.5]

Compute the mean Absolute error (MAE) of the obtained predictions (y_pred)?

Where Date Who Why

$$X = \begin{bmatrix} 2 & 4 \\ 1 & 2 \\ 2 & 0 \end{bmatrix} \Rightarrow X_{poly} \Rightarrow \begin{bmatrix} x_0 & x_1 & x_0^2 & x_1^2 & x_0x_1 \\ 2 & 4 & 4 & 16 & 8 \\ 1 & 2 & 1 & 4 & 2 \\ 2 & 0 & 4 & 0 & 0 \\ & * & & * & * \\ B \Rightarrow [& 2 & & 1 & 1,2 \ ] \\ & x & & x & \end{bmatrix} \Rightarrow y^* = \begin{bmatrix} 4\times2 + 16\times1 + \frac{8}{2} \\ 2\times2 + 1\times4 + \frac{2}{2} \\ 2\times0 + 1\times0 + 1,2\times0 \end{bmatrix}$$

Bais

(Q1) $\Rightarrow y^* = \begin{bmatrix} 28 \\ 9 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 29 \\ 10 \\ 1 \end{bmatrix}$

(Q2) $\Rightarrow MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \Rightarrow \frac{1}{3} \begin{bmatrix} 29 \\ 10 \\ 1 \end{bmatrix} - \begin{bmatrix} 28 \\ 9 \\ 5 \end{bmatrix} = \frac{1}{3}(1+1+4) = 2$

$\hat{y}$ → y_predict y → y_truth

4) computation of indices - clustering RI ⇒ Rand Index

Given the labels predicted by a clustering algorithm and ground truth labels:

0 1 2 3

y_truth = [1, 1, 1, 2]

y_pred = [3, 3, 1, 1]

Compute the Rand Index (RI) $\Rightarrow \frac{TP + TN}{\binom{n}{2}} \Rightarrow \frac{n(n-1)}{2}$

	truth	predict	TP	TN
[0,1]	1	1	1	0
[0,2]	1	0	0	0
[0,3]	0	0	0	1
[1,2]	1	0	0	0
[1,3]	0	0	0	1
[2,3]	0	1	0	0
			1	2

$\Rightarrow \frac{1+2}{\frac{4(3)}{2}} = \frac{3}{6} = \frac{1}{2}$

Where Date Who Why

5) clustering - DBSCAN - silh(P_i)

Given the following distance matrix (each cell describes the distance between two points)

Apply DBSCAN clustering - hyperparam ⇒ $\varepsilon = 5$, minpoints = 2

Q_1 ⇒ label each point with B (border), C (core), N (noise)

a	b	c	d	e	f	g
C	B	C	C	N	C	B

Q_2 ⇒ Assign a cluster id to each point:

a	b	c	d	e	f	g
0	1	0	1	-1	0	1

Q_3 ⇒ Compute the silhouette score of point g:

	a	b	c	d	e	f	g
a		6	4	7	8	3	6
b	6		6	3	7	7	6
c	4	6		7	7	3	9
d	7	3	7		6	8	4
e	8	7	7	6		7	8
f	3	7	3	8	7		6
g	6	6	9	4	8	6	

$$Silhouette(g) = \frac{inter(g) - intra(g)}{\max(inter(g), intra(g))}$$

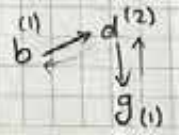
$intra(g) = \frac{1}{2}(6+9) = \frac{10}{2} = 5$ {d, b}

$inter(g) = \min\left(\frac{6+6+9}{3}, 8\right) = \min(7, 8) = 7$ (class(0) {a, c, f}, class(-1) {e})

$\frac{21}{3} = 7$

$Silh(g) = \frac{7-5}{7} = \frac{2}{7}$

⇒ class = 0



class = 1

$$Silh(P_i) = \frac{inter(P_i) - intra(P_i)}{\max(inter(P_i), intra(P_i))}$$

⇒ intra(P_i) ⇒ Avg dissimilarity of P_i with all other Point with in the same cluster

inter(P_i) ⇒ be the lowest Avg dissimilarity of P_i to any other cluster of which P_i is not a member of it

Maximal Frequent itemset

Where | Date | Who | Why

An itemset is frequent maximal if none of its immidate supersets is frequent.

close: A(count) is not greater than it's immidate support.

ABCD
ABCD
ABC
BCD
CD

→ ①

A	3
B	4
C	5
D	4

⇒ ②

AB	3
AC	3
AD	2
BC	4
BD	3
CD	4

AD: 2 < minsup

A(3) ⇒ AB(3), AC(3), AD(2)
- A is not close
- A is not max

B(4) ⇒ AB(3), BC(4), BD(3)
- B is not close
- B → max

max = A's immidiate support, itemset minsup are preset with min

minsup count ≥ 3

C(5) ⇒ AC(3), BC(4), CD(4)
- close: 5 > 3,4,4
- not max: 3,4,4 ≤ 3

D(4) ⇒ AD(2), BD(3), CD(4) ⇒
- not close ⇒ 4 > 2,3,4 ×
- not max ⇒ 2,3,4 ≤ 3

③

ABC	3
BCD	3

AB(3) ⇒ ABC(3) ⇒
- not close → 3 > 3 ×
- not max → 3 ≤ 3

AC(3) ⇒ ABC(3) ⇒
- not close → 3 > 3 ×
- not max

BC(4) ⇒ ABC(3), BCD(3) ⇒
- close: 4 > 3,3 ✓
- not max: 3,3 < 3 ×

BD(3) → BCD(3) ⇒
- not close: 3 > 3 ×
- not max

CD(4) → BCD(3) ⇒
- close ⇒ 4 > 3
- not max ⇒ 3 ≤ 3 ×

④

ABCD	2

< minsup

ABC(3) ⇒ ABCD(2) ⇒
- close ⇒ 3 > 2 ✓
- max ⇒ 2 ≤ 3 ✓

BCD(3) ⇒ ABCD(2) ⇒
- close ⇒ 3 > 2 ✓
- not max ⇒ 2 ≤ 3 ✓

	A	B	C	D	AB	AC	BC	BD	CD	ABC	BCD
close	×	×	✓	×	×	×	✓	×	✓	✓	✓
max	×	×	×	×	×	×	×	×	×	✓	✓

Where Date Who Why

Frequent - Item Set :

is an itemset whose support is greater than or equal to a minsup threshold

Support : $A \rightarrow B$

is the fraction of transactions that contains an itemset exp: {Beer, milk} = 2/5 → تراکنش Transaction

$$Sup = \frac{\#\{A, B\}}{T}$$

confidence : $A \rightarrow B$

is the frequency of B in transactions Containg A

$$conf = \frac{Sup(A, B)}{Sup(A)}$$

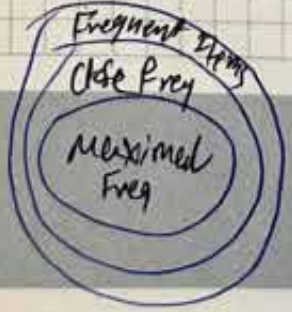
* if $A \subseteq B$ then $Sup(A) \geq Sup(B)$

Maximal Frequent Itemset

An itemset is frequent maximal if none of its immediate supersets is frequent.

Close Item Set

if none of it's immidiate suport has the same support as the item set.



maximal frequent itemset:

Where
ABC
ABCD
BCE
ACDE
DE

→

Date	
A	3
B	3
C	4
D	3
E	3

Who	
AB	2
AC	3
AD	2
AE	1
BC	3
BD	1
BE	1
CD	2
CE	2
DE	2

Why	
ABC	2
ABD	1
ACD	2
CDE	1

ABCD	1

minSup=2

closed

① C(4) ⇒ AC(3), BC(3), CD(2), CE(2) ⇒ 4>3,3,2,2 ✓

② D(3) ⇒ AD(2), BD(1), CD(2), DE(2) ⇒ 3>2,1,2,2 ✓

③ E(3) ⇒ AE(1), BE(1), CE(2), DE(2) ⇒ 3>1,1,2,2 ✓

④ AC(3) ⑤ BC ⑥ CE ⑦ DE ⑧ ABC ⑨ ACD

maximal

① CE(2) ⇒ CDE(1) ⇒ 1<2 ✓

② DE(2) ⇒ CDE(1) ⇒ 1<2 ✓

③ ABC(2) ⇒ ABCD(1) ⇒ 1<2 ✓

④ ACD(2) ⇒ = = = ✓

Robustness to the noise

Where Date Who Why

Classifications:

1) Decision Tree (Hunt's Algorithm) ✗

2) classification Rules ✓

3) Association Rules ✓

4) Neural Network

5) Bayesian Networks ⇒ Fully incremental ✗

$P(X|C)$

$P(x_1, x_2, \dots x_n | \Delta)$

↳ assign x to class with maximal $P(C|X)$

6) k-nearest neighbours (kNN) incremental

majority vote ① → based on distance

weighted vote ② → according to distance

7) support vector Machines (SVM) ✓

8) Random forest ✓

Classification:

Given the following confusion matrix

	a	b	c	d
a	10	0	0	0
b	0	4	0	4
c	0	4	10	0
d	0	2	0	6

actual ↓ / predicted →

Q1: Compute the accuracy score

Q2: Compute F-measure (F1) of class b

Q1: $\Sigma_{correct} = 10+4+10+6 = 30$ (اول عناصر قطر اصلی را جمع میکنیم و جمع عناصر روی قطر اصلی را حساب میکنیم)

$$\text{Accuracy} = \frac{\text{total correct prediction}}{\text{total prediction}} = \frac{30}{30+10} = \frac{3}{4}$$

(مجموع عناصر ماتریس غلط) Incorrect

Q2: $F1 = \frac{2P \cdot R}{P+R}$

Precision: P
Recall: R

For class (b) →
- $P = \frac{4}{10}$ → Correct (عناصر قطر اصلی) / تعداد عناصر ستون b
- $R = \frac{4}{8}$ → تعداد قطر اصلی آن سطر / تعداد عناصر سطر

$$F1 = \frac{2 \times \frac{4}{10} \times \frac{4}{8}}{\frac{4}{10}+\frac{4}{8}} = \frac{4}{9}$$

Regression: (3 point)

- Given the following dataset, with 2 features (X0, X1) and 3 data points:

$$X = [[2,4],[1,2],[2,0]]$$

- Apply to X the following multinomial regression pipeline

1 - Feature extraction step

$$[X_0, X_1, X_0^2, X_1^2, X_0X_1]$$

2 - Regression parameters (to be applied on the extracted features)

$B = [0, 2, 0, 4, 1/2]$, $Bias = 1$

$$\begin{bmatrix} 28 \\ 9 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 29 \\ 10 \\ 1 \end{bmatrix} = \hat{y}_{pred}$$

- Q1 = what is the output vector with the predictions?

y_pred = [?]

$$X = \begin{bmatrix} 2 & 4 \\ 1 & 2 \\ 2 & 0 \end{bmatrix} \Rightarrow ① \; X_{poly} = \begin{bmatrix} x_0 & x_1 & x_0^2 & x_1^2 & x_0x_1 \\ 2 & 4 & 4 & 16 & 8 \\ 1 & 2 & 1 & 4 & 2 \\ 2 & 0 & 4 & 0 & 0 \end{bmatrix}$$

② $B = [0 \; 2 \; 0 \; 4 \; 1/2]$

$$* \; y = \begin{bmatrix} 4\times2+16\times4+\frac{8}{2} \\ 2\times2+4\times1+\frac{2}{2} \\ \cdots \end{bmatrix} = \begin{bmatrix} 28 \\ 9 \\ 0 \end{bmatrix}$$

Q2: Given the ground truth predictions

- y_truth = [28, 9, 5]

- compute the Mean Absolute Error (MAE) of the obtained predictions (y_pred)

$$y = \begin{bmatrix} 28 \\ 9 \\ 5 \end{bmatrix}$$

ground truth: y_i ↓ ; Prediction: $\hat{y}_i$ ↑

$$MAE = \frac{1}{n}\sum_{i}^{n} |\hat{y}_i - y_i| \Rightarrow \frac{1}{3}\left|\begin{bmatrix} 29 \\ 10 \\ 1 \end{bmatrix} - \begin{bmatrix} 28 \\ 9 \\ 5 \end{bmatrix}\right| = \frac{1}{3}|1+1+4| = 2$$

4. Computation of indices

- Given the labels predicted by a clustering algorithm and ground truth labels:

- y_true = [1, 1, 1, 2] (indices 0, 1, 2, 3)

y_pred = [3, 3, 1, 1] (indices 0, 1, 2, 3)

- Compute the Rand Index score (RI)

$$RI = \frac{TP + TN}{\binom{n}{2}} = \frac{TP + TN}{\frac{n(n-1)}{2}}$$

- where TP = number of pairs of elements that are in the same set in y_true and in the same set in y_pred
- TN = number of pairs of elements that are in different sets in y_true and different sets in y_pred
- n = number of data points

$n = 4$

	y_true	y_pred	TP	TN
[0,1]	1	1	1	0
[0,2]	1	0	0	0
[0,3]	0	0	0	1
[1,2]	1	0	0	0
[1,3]	0	0	0	1
[2,3]	0	1	0	0
			1	2

$$RI = \frac{1+2}{\binom{4}{2}} = \frac{3}{\frac{4!}{2!\,2!}} = \frac{3}{\frac{4\times3\times2!}{2\times1\times3!}}$$

$$RI = \frac{3}{6} = \frac{1}{2}$$

5_ Python-related questions

$\sqrt{\sum (x-y)^2}$ axis = 1

برداشته شود   متناظر ستون از اولی

متناظر سطر از دومی برداشته می‌شود

- Given two Numpy vectors
  - X = with shape (100, 50)
  - y = with shape (50,) = (1, 50)

a) np.sqrt((X - y)**2).sum(axis = 1))
is the euclidean distance between rows of X and y and the result has shape (100, 1)

b) np.sqrt(((X - y)**2).sum(axis = 1))
has shape (100,)

c) np.sqrt(((X - y)**2).sum(axis = 0))
has shape (100,)

d) np.sqrt(((X - y).sum(axis = 1))**2)
has shape (100, 1)

X = [ ] 100 — 50 (100x50)   y = [ ] —50— (1x50)

we can broadcasting

euclidean = $\sqrt{\sum (y_i - x_i)^2}$ → دقیقاً تک‌ها پیاده‌سازی این فرمول می‌کنند.

one dimensional & row factor   ← axis = 1

7- Python-related questions

- Given a Dataframe with four columns (category, year, month, # subscriptions)

a) df[['category', 'year']].pivot_table(#subscriptions, index='category', columns='year', aggfunc='mean')

return the info about average no of subscription for each combination of category/year

b) df.groupby(by=['category']).sum().unstack()

returns information about the total number of subscriptions for each combination of category and year.

c) df.pivot_table('#subscriptions', index='category', columns='year', aggfunc='sum')

returns information about the maximum number of subscriptions for each combination of category and year.

d) df.drop(columns='month').groupby(by=['category', 'year']).sum().unstack()

returns information about the total numbers of subscriptions for each combination of category and year.

e) None of the previous answers is correct

1- Preprocessing

(2024-12-03)

- The following list represents training set values of a specific attribute:

[10, 0, 5, 3, 3, 0, 3, 4, 4, 7, 5, 7, 8, 4, 9] = تعداد 15

- Use these values to train an equal-frequency based discretization with three bins (low, medium, high). which statement is correct?

a) The test vector [1, 7, 9] is discretized to [low, medium, high]

b) The test vector [10, 7, 4] is discretized to [high, medium, medium]

c) The test vector [3, 4, 7] is discretized to [low, medium, high]

d) The test vector [5, 4, 2] is discretized to [high, medium, low]

Width-Based

MIN ——— MAX

15 عدد داریم که باید به 3 بازه تقسیم شود.

$\frac{15}{3} = 5$ هر بازه 5 تا عدد دارد.

Frequency

		اعداد
low	2	0
	3	3
medium	3	4
	2	5
high	2	7
	1	8
	1	9
	1	10

بازه‌ها (low, Medium, high) 3bins

2- Time Series

- Which is the most significative pair of features for distinguishing between the two periodic time series depicted in the figures below?

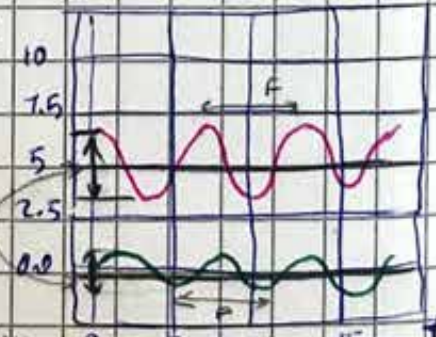
a) Mean, First deprivative

b) Mean, percentiles

c) First derivative, percentiles

d) Percentiles, Frequency

e) All of the pairs above are equivalent for distinguishing between the two series



3- Classification

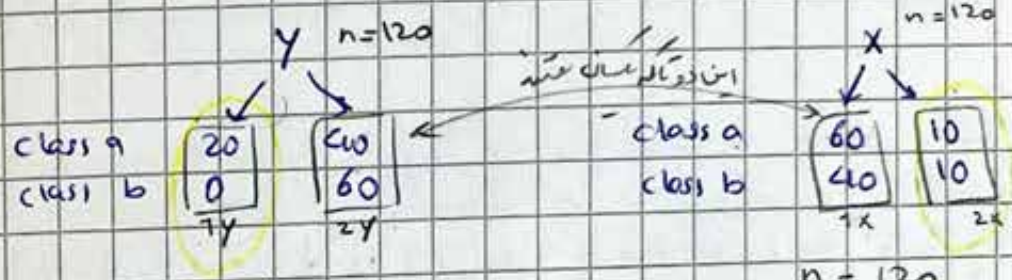
The two dataset splits depicted in the figure represent an intermediate step of Hunt's algorithm.

ⓐ. Compute the Gini index of the two splits
- Gini(X), Gini(Y)?

ⓑ. which of the two attribute splits will be selected by the algorithm?

a- X
b- Y

Partition : Pa



$n = 120$
$n_{1X} = 100$
$n_{2X} = 20$
$n_{1Y} = 20$
$n_{2Y} = 100$

X)
$$Gini_{split_j} = \sum_i \frac{n_{ij}}{n} Gini_{pa_i}$$

$$Gini_{pa_i} = 1 - \sum_k \left(\frac{m_k}{n_i}\right)^2 \Rightarrow 1 - \left(\frac{x_a}{X}\right)^2 - \left(\frac{x_b}{X}\right)^2$$

1X) $1 - \left(\frac{60}{100}\right)^2 - \left(\frac{40}{100}\right)^2 = \frac{12}{25}$ ← $Gini_{pa(1X)}$

2X) $1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 = \frac{1}{2}$ ← $Gini_{pa(2X)}$

$$Gini_X = \frac{1X}{n} \times Gini_{pa(1X)} + \frac{2X}{n} \times Gini_{pa(2X)} = \frac{100}{120} \times \frac{12}{25} + \frac{20}{120} \times \frac{1}{2} = \frac{58}{120}$$

1Y: $1 - \left(\frac{20}{20}\right)^2 - \left(\frac{0}{20}\right)^2 = 0$

2Y: $1 - \left(\frac{40}{100}\right)^2 - \left(\frac{60}{100}\right)^2 = \frac{100 - 16 - 36}{100} = \frac{48}{100} = \frac{12}{25}$

$$Gini_Y = \frac{20}{120} \times 0 + \frac{100}{120} \times \frac{48}{100} = \frac{4}{10}$$

ⓑ در جواب باید بیشتر آن را انتخاب کنیم که Gini_index کوچکتری داشته باشد

$Gini_Y = \frac{40}{100}$ ✓

$Gini_X = \frac{58}{120}$

$Gini_{Pa\,1Y} = 0$ <

$Gini_{Pa\,2X} = \frac{1}{2}$ | چون $Gini_{Pa\,1Y}$ کوچکتر است

4- Hierarchical Clustering

- Given the following distance matrix, apply agglomerative hierarchical clustering with single-linkage (min).

- Which statement is correct?

a) with $k=3$ clusters, a and b are in the same cluster

b) with $k=2$ clusters, c and d are in different clusters

c) with $k=3$ clusters, b and c are in different clusters

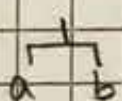
d) with $k=2$ clusters b and c are in the same cluster

e) All of the previous answers are correct.

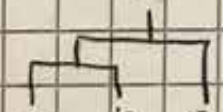
$d(x,y) = d(y,x)$

distance matrix

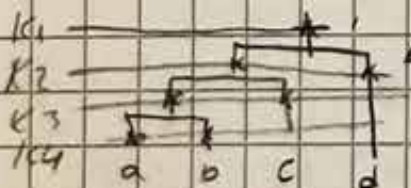
	a	b	c	d
a	0	1	4	3
b	1	0	2	4
c	4	2	0	3
d	3	4	3	0



	AB	c	D
AB	0	2	3
C	2	0	3
D	3	3	0



	ABC	D
ABC	0	3
D	3	0



تعداد کلاستر k

$k=1$ ABCD

$k=2$ ABC D

$k=3$ AB C D

$k=4$ A B C D

$a,b=1$ ← انتخاب کمترین فاصله ← گام ①

$a,c=4$
$b,c=2$ ✓ انتخاب کمترین فاصله ② گام

$a,d=3$ ✓ انتخاب کمترین فاصله
$b,d>4$

$c,d=3$ انتخاب کمترین فاصله

$AB,C=2$ ← انتخاب کمترین فاصله در جدول X1 ③ گام

$AB,d=3$
$C,d=3$ ← انتخاب کمترین فاصله در جدول X1

④ گام

1-(3points)

2020/6/16 امتحان

Let X be a feature vector, $X \in R^N$, i.e., $X = (x_1, x_2, \dots x_n)$

Naive Bayes classifiers compute the output class based on the following definition:

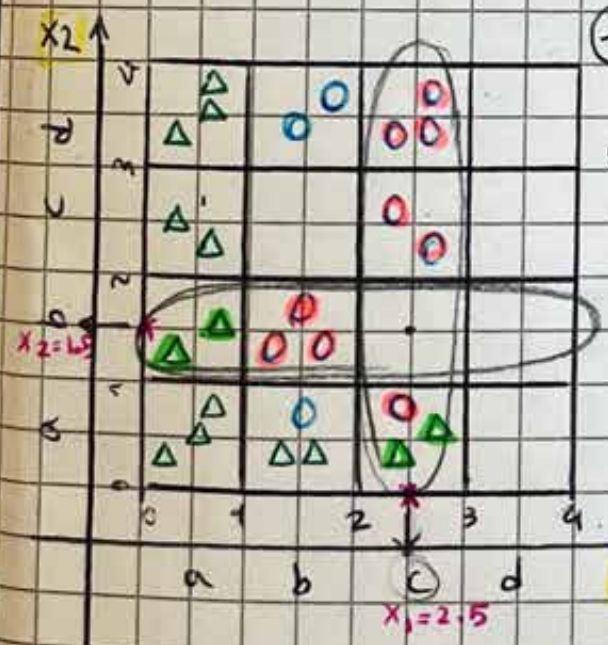
$P(class|X) = P(x_1|class) * P(x_2|class) * P(x_3|class) \dots * P(x_n|class) * P(class)$

If is given a dataset represented in Figure, where x1 and x2 are two continuous features, while triangle and circle are two class labels. The two features are discretized in four bins [a, b, c, d] depicted in the figure with a grid.

Suppose that a Naive Bayes classifier has been trained on the provided dataset, which contain 14 triangles and 12 circles.

classify the test data sample X with feature $x_1 = 2.5$, $x_2 = 1.5$

write in the box the following values one answer per row:



1- P(triangle|X)

2- P(circle|X)

3- The class assigned to X (circle or triangle)

① : $P(\Delta|X) = P(x_1|\Delta) * P(x_2|\Delta) * P(\Delta)$

$= P(c|\Delta) * P(b|\Delta) * P(\Delta)$

$= \frac{2}{14} \times \frac{2}{14} \times \frac{14}{26}$

تعداد مثلث ها در ستون c ← تعداد کل مثلث ها ← تعداد کل مثلث ها در سطر b ← تعداد کل دایره و مثلث

$\rightarrow P(\Delta|X) = \frac{1}{91}$

② : $P(O|X) = P(x_1|O) * P(x_2|O) * P(O)$

$= P(c|O) * P(b|O) * P(O)$

$= \frac{6}{12} \times \frac{3}{12} \times \frac{12}{26}$

$= \frac{3}{52}$

$\frac{3}{52} > \frac{1}{91}$ → O → اشکال دارد نمیشه

3. The class assigned to X (circle or triangle)

2-(3 points)

The ~~Manhattan distance~~ is defined as follows:

$dist(P1, P2) = |X1 - X2| + |y1 - y2|$

in the DBSCAN algorithm the datasets points can be labeled as:

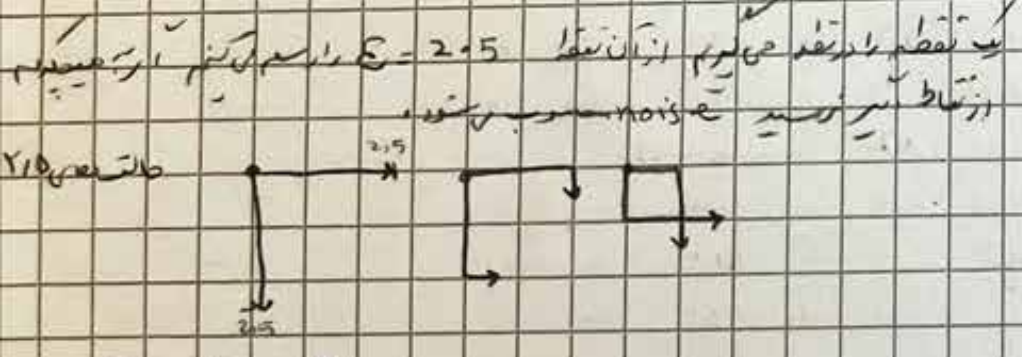
- core (points with at least minpoints neighbors within a distance ϵ)
- ~~border~~ (at least a core points in the neighborhood)
- noise (neither a core point, nor a border point)

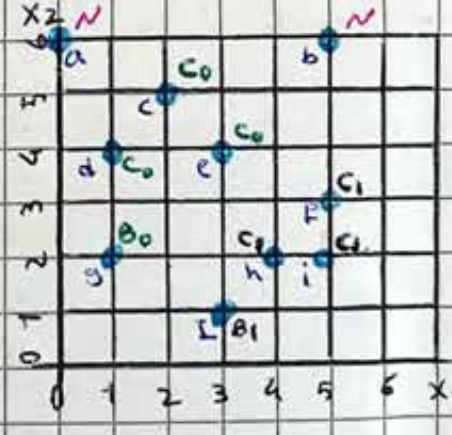
Using the Manhatan distance, apply the DBSCAN algorithm to the following points in the bidimensional space.

use the following hyperparameters: $\epsilon = 2.5$, minpoints = 2 (at least 2 points as neighbors) for each point write:

The assigned label (N=noise, B= border, C = core)

The assigned cluster id (order of cluster ids is not important, use -1 for noise points





a	N	-1
b	N	-1
c	C	0
d	C	0
e	C	0
f	C	1
g	B	0
h	C	1
i	C	1
l	B	1

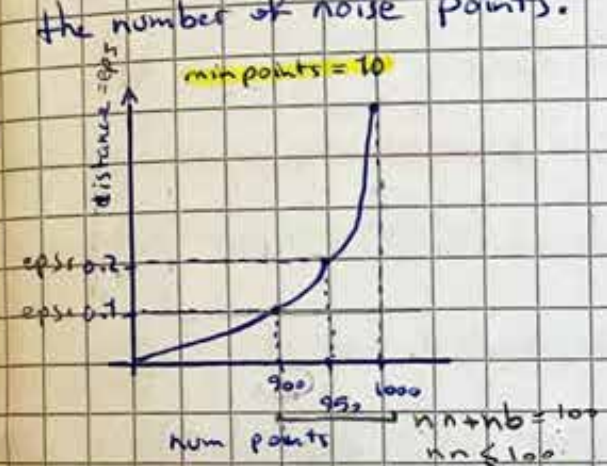
3- (4 points)

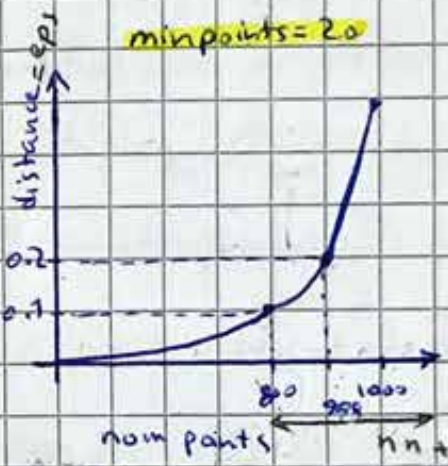
The chart shown in the figure below is typically used to set up epsilon (eps) and minpoints values in DBSCAN.

It shows on the y-axis, for each point in the dataset, the distance from the K-th nearest point, where k is equal to minpoints.

The x-axis represents all the dataset points, ordered by increasing distance from the k-th nearest point.

Let nc the number of core points, nb the number of border points, nn the number of noise points.





which of the following statements is true?

(a) with minpoints = 10, eps = 0.1, nc ≤ 900, nn = 100 ✗ nc = 900

(b) = 20, eps = 0.2, nc < 900, nb ≤ 100 ✗ nc = 900

(c) = 10, eps = 0.2, nc+nb = 950, nn = 50 ✗ nb = 0 ... 950 ...

(d) = 20, eps < 0.1, nc = 800, nn ≤ 200

(e) = 10, eps < 0.1, nc = 900, nb < 50 ✗

(f) = 20, eps < 0.2, nc = 900, nn = 100 ✗ nn = 100

nc = 900

Core ... Core

nn + nb = Core

nn ≤

nb ≤

4_(1 point)

Broadcasting rules for executing an operation between two Numpy vectors are defined as follows:

- The shape of the array with fewer dimensions is padded with leading ones
- if the shap along the dimension is 1 for one of the arrays and >1 for the other, the array with shape=1 in that dimension is stretched to match the other array
- if there is a dimension where both arrays have shape>1 then broadcasting cannot be performed

Apply the broadcasting rules to the following numpy vectors.

```
X = np.array([[1,2,3],
              [0,0,0],
              [1,1,1]])
y = np.array([1,2,3])
Z = np.array([[1],
              [0],
              [-1]])
```

row vector

$$X + y = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}$$

shape: (3,3) (1,3)

$$y + Z = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & 3 \\ 0 & 1 & 2 \end{bmatrix}$$

row vector

column vector

(b) $X+y = [[2,4,6],[1,2,3],[2,3,4]]$ and $y+Z = [[2,3,4],[1,2,3],[0,1,2]]$

---

5 (1 point, -15% penalty for a wrong answer

An itemset is defined to be frequent maximal if none of its immediate supersets (i.e. including one more item) is frequent.

After the analysis of a transactional dataset, the complete set of maximal frequent itemsets with minsup = 900 are. (abc = 930 , def = 980)

(a) sup(ab) can be in range [900, 929], sup(abcd) cannot be in range [0, 930]

(b) sup(ab) can be a value in range [900, 929], sup(abcd) can be a value in range [900, 930]

(c) sup(ab) can be a value in range [930, 1000], sup(cd) cannot be in range [900, 930]

(d) sup(ab) can be a value in range [930, 1000], sup(abcd) can be a value in range [900, 930]

(e) None of the other answer is correct ✗

(f) sup(ab) can be a value in range [900, 929], sup(cd) cannot be in range [900, 930]

$\Longrightarrow$ sup(ab) $\geq$ 930

b-(1 point / -15% penalty)

A neural network is trained on a training set with 2 features x_1, x_2
- x_1 is a continuous value
- x_2 is a nominal value among the following: "high", "med", "low"

Before training, the features are preprocessed in the following way:
- x_1 is normalized with z-score: $x_1 norm = (x_1 - np.mean(x_1)) / np.std(x_1)$ → training
- x_2 is encoded with a one hot vector ('high': 001, 'med': 010, 'low': 100)

when applying the trained classifier on unlabeled data (denoted as test set) which of the following statements are true (multiple answers maybe true)

a) the x_1 feature of the test record must be normalized with the mean and standard deviation previously computed for the training set.
b) the x_1 feature of the test record can be provided to the classifier without normalization
c) the x_1 feature of the test record must be normalized with the mean and standard deviation computed for the test set.
d) it is suggested to process a record with a value of x_2 not included in 'high', 'med', 'low' by choosing a random value for the one hot features.
e) it is suggested to process a record with a value of x_2 not included in 'high' 'med', 'low' by adding an input neuron to the neural network
f) a record with a value of x_2 not included in 'high', 'med', 'low' cannot be processed.

x_1
x_{Hot}
x_{med}
x_{low}

7- (1 point/-15% penalty)

A model's capacity defines its ability of fitting more complex data.

Regularization is defined as a set of techniques to avoid overfitting.

Reffering to a polynomial regression problem, which of the following statements is true?

(a) All the statements are correct.

(b) Adding higher power degrees (e.g. $x^4 x^5$) is always useful to obtain better predictions, since it increases model capacity.

(c) Regularization decreases model capacity and it is only suitable for output variables that are linearly correlated with features

(d) Regularization is always necessary when the output variable presents complex non linear relationships with the features

$MSE + \lambda \sum |w_i|$

(e) Regularization helps to decrease model capacity and fit more complex data.

(f) Adding higher power degrees (e.g. $x^4 x^5$) to the input features increases model capacity and the likelihood of having overfitting

8- (1point/ -15% penalty)

To access the quality of a classification result the following measures were defined.

- accuracy: the percentage of correctly classified items.
- precision (c): the percentage of correctly classified items among those predicted with class c
  ▲ تعداد correct class c / جمع تعداد آبجکت‌هایی که به کلاس c اختصاص داده شده
- recall (c) the percentage of correctly classified items among those with actual class c
  ▲ correct class c / تعداد واقعی آن کلاس

Let C be a supervised classifier and a two-dimensional input space. The points drawn from belong to one of the three classes: circle, square, Triangle. After the training stage, C learns the following decision boundaries:

where the label assigned by C is given by the background color of the region in I.

suppose that the symbols in the image are test points whose shape represents the respective ground truth label (either a circle, a square, a triangle)

which of them is True?

$P_{\square} = \frac{3}{4}$ $R_{\square} = \frac{3}{5}$

$P_{\triangle} = \frac{5}{5} = 1$ $R_{\triangle} = \frac{5}{6}$

$P_{\circ} = \frac{8}{11}$ $R_{\circ} = \frac{8}{9}$

(a) None of the answers is correct

(b) The recall for square is $3/4 \neq \frac{3}{5}$

(c) The precision for Triangle is 1

(d) The precision for square is $3/5 \neq \frac{3}{4}$

(e) The accuracy is 1 → we have some points that are not predicted correctly. $\frac{3+8+5}{20} = \frac{16}{20} \neq 1$

1) (2 points) 2022-01-13 ✓ (امتحان 2021-01-25)

A KNN classifier is trained on the training set in the table below which only contains categorical features.

with categorical features, the distance between two samples **a** and **b** can be computed as the number of features with different values

$$dist(a,b) = \sum_i \ell(a_i, b_i)$$

where:

$$\ell(m,n) = \begin{cases} 0 & \text{if } m=n \\ 1 & \text{otherwise} \end{cases}$$

for each neighbor X_i of x the vote is weighted:

$$weight(x, X_i) = \frac{1}{dist(x, X_i) + .5}$$

	width	weight	speed	class	dist(t_1)
X_1	big	light	fast	A	1
X_2	big	heavy	fast	B	0
X_3	small	heavy	slow	B	2
X_4	big	heavy	slow	A	1
X_5	small	light	slow	A	3

Given the test point below, write in the answer box

هر ستون که با ستون t1 نابرابر باشد

1- the list of the 3 neighbors of t1 → تعداد تفاوت ها را شمارش می‌کنیم و در ستون dist(t1) می‌نویسیم

2- the class assigned to t1 with k=3

	width	weight	speed	class
t1	big	heavy	fast	B

use the following notation

neighbors = {list of neighbors for t1}

class = class assigned

انتخاب ۳ تا با کمترین فاصله

	dist(t1)	weight(x, X_i)	class
X_1	1	$\frac{1}{1.5} = \frac{2}{3}$	(A)
X_2	0	$\frac{1}{1/2} = 2$	B
X_4	1	$\frac{1}{1.5} = \frac{2}{3}$	(A)

جواب سوال ①

مقدار آرا $A = \frac{2}{3} + \frac{2}{3} = \frac{4}{3}$

مقدار آرا $B = 2$

درست که تعداد آرای A بیشتر از B است اما وزن کلاس B بیشتر از A است و ما براساس وزن انتخاب می‌کنیم نه تعداد

• k-means ⟵ نقاطی که در مرحله اول انتخاب شدند cluster را میانگین می‌گیریم که centroid جدید را دربیاریم و دوباره هر نقاط را از نو با نقطه جدید می‌سنجیم و دسته‌بندی می‌کنیم

22 (1 point)

The z-score normalization is applied to a feature x by removing the mean μ and dividing by its standard deviation σ the resulting feature x' is $= \frac{x - \mu}{\sigma}$

$t_1 - \text{mean}$ $\quad \mu = 0$ $\quad t_1 - \text{Mean}_{train}$ $\quad t_1 \ldots$

$\sigma = 1$ $\quad \sigma_{train}$

Consider the regression task described by the training set and the test set below. in which the features have not been normalized the mean and the standard deviation (std) for each column are included at the bottom of the tables For your convenience. The dataset contains three input features (Freq, rpm, power) and an output variable (Y).

The input features have been normalized with z-score (not shown in the tables before training a regression model to predict the value of Y.

You are required to perform the normalization on the test point t1 (see test set table)

write in the answer box the normalized row corresponding to the test point t1, after applying the proper z-score normalization on the features.

training set →

	Freq	rpm	power	y
x1	10	20	0	0
x2	5	0	2	10
x3	5	0	0	10
x4	10	20	2	40
mean	7.5	10	1	15
std	2.5	10	1	15

t1 = [freq, rpm, power, y]

✓ $freq = \frac{5.5 - 7.5}{2.5} = \frac{-4}{5}$

$rpm = \frac{1 - 10}{10} = \frac{9}{10}$

$power = \frac{4 - 1}{1} = 3$

$y = 3$

$freq = \frac{t_{1(freq)} - Mean_{(freq)\,train}}{Std_{(train)}}$

	freq	rpm	power	y	test
t1	5.5	1	4	3	
t2	0.5	6	8	6	
mean	3	3.5	6	4.5	
std	2.5	2.5	2	1.5	

normalize کردن برای input هست، y ما همان output واقعی می‌باشد

$Y_{pred} = Y_{true}$ (class(A))

TP ← تشخیص داده A هست و در واقعیت هم A بوده

TN ← تشخیص داده A نیست ولی در واقعیت A بوده

FP ← تشخیص داده A هست ولی در واقعیت A نبوده

FN ← تشخیص داده A نیست و در واقعیت هم A نبوده

Y_{true} = [A A B C]

Y_{pred} = [A B A B]

TP TN FP FN

---

3. (2 points)

The Gini index of the node is computed as follows:

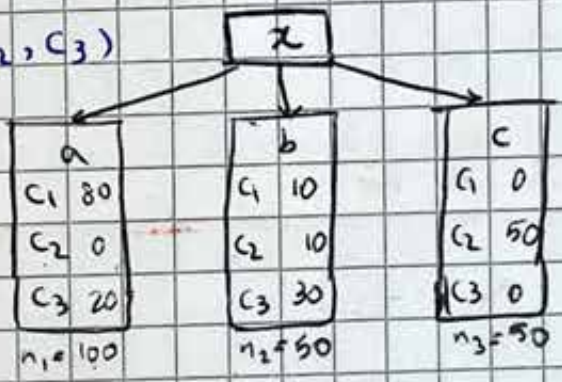
$gini(node) = 1 - \sum_j P(j|t)^2$

where $P(j|t)$ is the relative frequency of class j at node t

The Gini index of a spilit with parent p and children c_i is computed as follows: $gini(split) = \sum_i \frac{n_i}{n} gini(c_i)$

where n_i is the number of records at child c_i and n is the number of records in p.

In the figure it is shown a split x, with three children (a, b, c) for each child you are given the number of elements belonging to each of three classes. (c_1, c_2, c_3)



gini(a) = value

gini(x) = value

1- the gini index of child a

2- the gini index of the split x

$Gini\ a = 1 - \left(\frac{80}{100}\right)^2 - \left(\frac{20}{100}\right)^2 = 1 - \frac{16}{25} - \frac{1}{25} = \frac{8}{25}$

$Gini\ b = 1 - \left(\frac{10}{50}\right)^2 - \left(\frac{10}{50}\right)^2 - \left(\frac{30}{50}\right)^2 = 1 - \frac{1}{25} - \frac{1}{25} - \frac{9}{25} = \frac{14}{25}$

$Gini\ C = 1 - \left(\frac{50}{50}\right)^2 = 0$

$Gini\ X = \frac{100}{200} \times \frac{8}{25} + \frac{50}{200} \times \frac{14}{25} = \frac{4}{25} + \frac{7}{50} = \frac{15}{50} = \frac{3}{10} = 1/3$

4) (2 points)

Given the similarity matrix in the figure, apply agglomerative hierarchical clustering with single (MIN) linkage. The similarity metric ranges from 0 (more distant) to 1 (less distant)

	a	b	c	d	e
a	1	0.4	0.4	0.8	0.5
b	0.4	1	0.3	0.3	0.4
c	0.4	0.3	1	0.4	0.9
d	0.8	0.3	0.4	1	0.5
e	0.5	0.4	0.9	0.5	1

let K be the number of obtained clusters. write in the answer box:

1. The clusters obtained with K=3

2. The clusters obtained with K=2

Use the following notation:

K=3 { points of the first cluster }
{ points of the second cluster }
{ points of the third cluster }

K=2 { points of the first cluster }
{ points of the second cluster }

	A	B	CE	D
A	1	0.4	0.5	0.8
B	0.4	1	0.4	0.3
CE	0.5	0.4	1	0.5
D	0.8	0.3	0.5	1

D,C → 0.4
D,E → 0.5 ✓
A,C → 0.4
A,E → 0.5 ✓
B,C → 0.3
B,E → 0.4 ✓

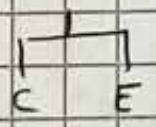
	AD	B	CE
AD	1	0.4	0.5
B	0.4	1	0.4
CE	0.5	0.4	1

B,A → 0.4 ✓
B,D → 0.3
A,CE → 0.5 ✓
D,CE → 0.5

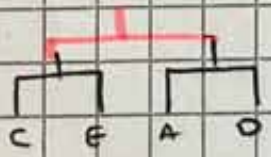
	ADCE	B
ADCE	1	0.4
B	0.4	1

AD,B → 0.4
CE,B → 0.4

ظاهر 1) در گام اول انتخاب با بیشترین عدد است. (0.9)
(ترکیب C,E)



ظاهر 2) انتخاب بیشترین مقدار دوم (0.8)
ترکیب (A,D)



ظاهر 3) انتخاب بیشترین مقدار دوم (0.5)
ترکیب (AD, CE)

q1) K=3 clusters : (B, CE, AD)

q2) K=2 clusters : (B, ACDE)

5) (1 point - 15% penalty)

Consider the following numpy array:

```
X = np.array([[2, 4, 6, 8],
              [3, 6, 9, 12],
              [4, 8, 12, 16]])
```

The following indexing techniques are applied to obtain a and b

```
a = X[1:, ::-1]
b = X[[False, True, True], :0:-1]
```

a = [[12, 9, 6, 3], [16, 12, 8, 4]]

b = [[12, 9, 6], [16, 12, 8]]

Which of the following assertion is correct?

a) The script will provide an error because b cannot be computed.

b) The value of a is [[3,6,9,12],[4,8,12,16]]. the variables a & b contain the same values.

c) The value of a is [[12,9,6,3],[16,12,8,4]]. The variables a & b contain different values.

d) The values of a is [[3,6,9,12],[4,8,12,16]]. The variables a & b contain different values.

e) The value of a is [[12,9,6,3],[16,12,8,4]]. The variables a & b contain the same values.

f) None of the other answers is correct.

6) (2.5 points) Silhouette { range between -1 and +1
* typically between 0 and 1
* the closer to 1 the better

for each data point P_i

- let intra(P_i) be the average dissimilarity of P_i with all other points within the same cluster
- let inter(P_i) be the lowest avarage dissimilarity of P_i to any other cluster of which P_i is not a member.

The Silhouette score of point P_i is defined as:

$$silh(P_i) = \frac{inter(P_i) - intra(P_i)}{max(inter(P_i), intra(P_i))}$$

for two n-dimensional points $a=(a_1, a_2, \dots, a_n)$ and $b=(b_1, b_2, \dots, b_n)$

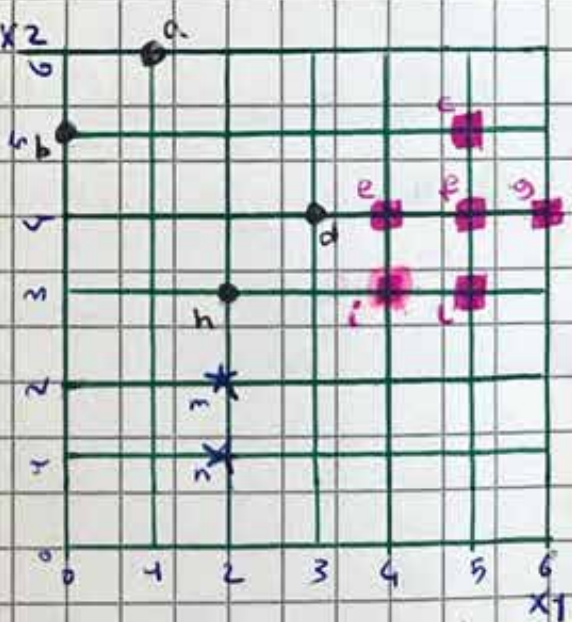
the Manhatan distance is defined as follows.

$$dis(a,b) = \sum_i^n |a_i - b_i|$$

in the figure below, three different clusters are represented with the following shapes: circle, star, square. Use the Manhatan distance as the dissimilarity metric.

consider point (i) belonging to the squer cluster

- intra(i) $= \frac{1}{5}(1+1+2+3+3) = 2$
- inter(i)
- silh(i)



notation:

- intra(i) = value
- inter(i) = value
- silh(i) = value

$inter(i) = \min(d(i,\bullet), d(i,*)) = \frac{1}{n}\sum_{j \in \bullet} d(i,j)$

$d(i,\bullet) = \frac{1}{5}(1+1+2+3+3) = 2$ (1 = d, 1 = h, 2 = a, 3 = b)

$d(i,\bullet) = \frac{1}{4}(2+2+6+6) = 4$

$d(i,*) = \frac{1}{2}(3+4) = 3.5$ (3 = m, 4 = n) → min

→ $inter(i) = 3.5$

$silh(i) = \frac{3.5-2}{3.5} = \frac{3}{7}$

#7 (1.5 points - 15% penalty)

An itemset I is closed if none of its immediate supersets has the same support as I.

An itemset 'I' is represented by a collection of literals (e.g. abc) and a number representing its support count (e.g. abc: 12).

After the analysis of a transactional dataset, the list of frequent itemsets found (with support counts) is the following:

length=1, a:150 b:160 c:180 d:150 e:150

length=2, ab:140 ac:120 ad:130 ae:140

length=3 abc:120 acd:110 abe:130

length=4, abce:120

which of the following statement is true?

a) The only non-closed itemsets are ac, ad, abc, abce

b) The only non-closed itemsets are ac, ad

c) The only non-closed itemsets are ac, ad

d) None of the other answers is correct

e) The only closed itemsets are ac, abc

D) The only non-closed itemsets are ac, abc

	closed
4 120	ABCE
3 110	ACD
3 130	ABE
2 140	AB
2 130	AD
2 140	AE
1 150	A
1 160	B
1 180	C
1 150	D
1 140	E

	non-closed
3 120	ABC
2 120	AC

sup count

8) (1 point - 15% Penalty)

- Percision (c) is the fraction of correct predictions among the samples predicted with class c
- Recall (c) is the fraction of correct predictions among the samples with actual class c

after training two models (M1, M2) we obtain the confusion matrices shown below.

consider class b only. Which of the following statements is true?

M1	a	b	c	d
a	15	10	1	0
b	0	50	5	5
c	5	20	70	0
d	5	10	0	60

true class

Predicted class

M2	a	b	c	d
a	14	10	0	2
b	5	40	5	10
c	5	5	80	5
d	10	5	10	50

(a) M1 has higher recall than M2 and lower Percision

(b) M1 has the same recall as M2 and lower Percision

(c) None of the other answers is correct

(d) M1 has lower recall than M2 and higher Percision

(e) M1 has higher recall than M2 and higher Perecision

(f) M1 has lower recall than M2 and lower Precision

	P	R
M1	$\frac{50}{90} = \frac{5}{9}$	$\frac{50}{60} = \frac{5}{6}$
	$\wedge$	$\vee$
M2	$\frac{40}{60} = \frac{2}{3}$	$\frac{40}{60} = \frac{2}{3}$
	$\frac{2}{3} > \frac{5}{9}$	$\frac{5}{6} > \frac{2}{3}$

9) (1.5 points, -15% Penalty)

Consider the following Python operations:

```
l1 = [1,2,3]
l2 = [2,4,6]
d = {'mon': l1, 'tue': l2}
res = d.value()
l1.append(4)
d['tue'].append(8)
print(res)
```

$d = \{'mon': [1,2,3], 'tue': [2,4,6]\}$

[1,2,3,4] [2,4,6,8]

- dict_value([[1,2,3,4],[2,4,6,8]]) ✓

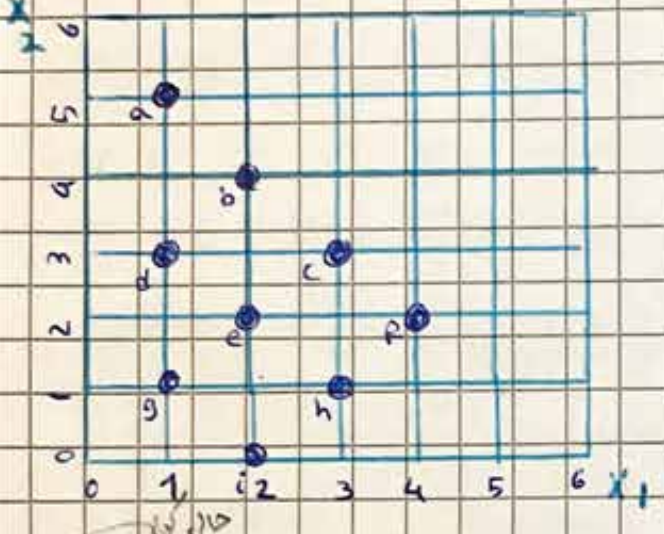
4) (1.5 points - 15% penalty)

The Mean Squared Error between two vectors $z, z' \in R^n$ is defined as

$$MSE(z, z') = \frac{1}{n} \sum_i (z_i - z'_i)^2$$

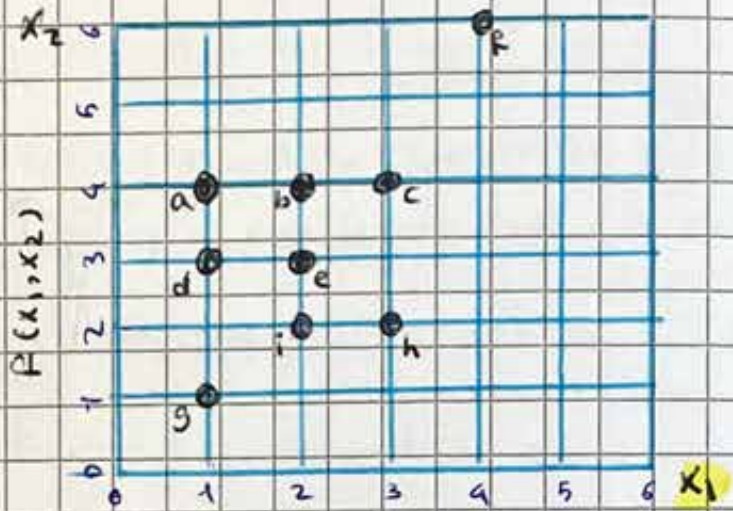
Consider the points in the following below with features x_1, x_2 and ground truth values $f(x_1, x_2)$

Input space / pred



y true

ground truth $f(x_1, x_2)$



Apply the following regression model to all the data points:

$$f'(x_1, x_2) = 0.5 x_1 + 0.5 x_2 + 1$$

Write in the answer box the value of the MSE between the prediction $f'(x_1, x_2)$ and the ground truth $f(x_1, x_2)$

	x_1	x_2	$f(x_1, x_2)$	$f'(x_1, x_2) = 0.5 x_1 + 0.5 x_2 + 1$
a	1	5	4	$\frac{5}{2^{10}} \times 1 + \frac{5}{2^{10}} \times 5 + 1 = 4$
b	2	4	4	4
c	3	3	4	4
d	1	3	3	4
e	2	2	3	3
f	4	2	6	4
g	1	1	1	2
h	3	1	2	3
i	2	0	2	1,5

x / y

$y - \hat{y}$

$f(x_1, x_2) - f'(x_1, x_2)$

$MSE = \frac{1}{n} \left(\sum (\hat{y} - y)^2 \right)$

$MSE = \frac{1}{9} (1 + 4 + 1 + 1 + \frac{1}{4})$

$MSE = \frac{1}{9} (\frac{29}{4}) = \frac{29}{36}$

0 + 0 + 0 + 1 + 4 + 2 + 1 + 1 + 0,25

7,25

(eXam 2021-09-06

12) (2 points)

An itemset is closed if none of its immediate supersets has the same support as the itemset. given the transactional dataset shown in the figure below, apply the apriori algorithm to extract all frequent itemsets.

The value of minsup is 2. support ≥ minsup

an itemset is considered to be frequent if its support count is equal to or higher than the minsup.

Q1) list all frequent itemsets having length 2, along with their support count.

Q2) list all itemsets of length 3 that have been generated by Apriori after the join and prune steps, before counting their support in the database.

Q3) list all frequent itemsets that are not closed, along with their support count.

A1) { list of itemsets w/ support count }

transactions		itemsets	sup
0	BC	close → A	4
1	AE	close → B	7
2	ACD	→ C	6
3	AB	→ D	4
4	BE	→ E	4
5	DE		
6	BCE		
7	BC		
8	BCD		
9	ABCD		

L1

itemsets	sup	
AB	2	
AC	1	x
AD	2	
AE	1	x
BC	5	close
BD	2	not close
BE	2	
CD	3	
CE	1	x
DE	1	x

L2

itemset	sup	
ABC	x	
ABD	1	x
ABE	x	
ACE	x	
ACD	x	
ADE	x	
BCD	2	
BCE	x	
BDE	x	
CDE	x	

L3

itemset	sup
BCD	2

L3

support ≥ 2

A1) { AB:2, AD:2, BC:5, BD:2, BE:2, CD:3 }

A2) { BCD:2 } / ABD, before counting sup ?!

A3) { AC:1, AE:1, CE:1, DE:1, ABD:1 }

FP-tree DB

①

TID	items
1	A,B → BA
2	B,C,D ✓
3	ACDE ✓
4	ADE ✓
5	ABC → BAC
6	ABCD ✓
7	BC ✓
8	ABC ✓
9	ABD
10	BCE

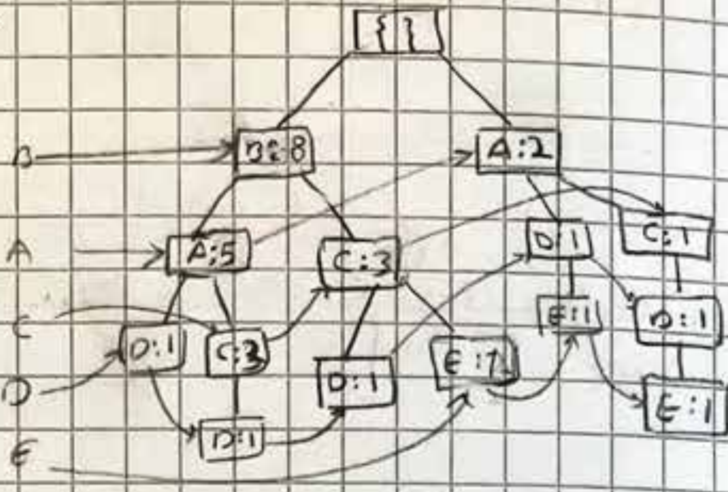
minsup > 1

item	sup
B	8
A	7
C	7
D	5
E	3

	sup
BAC	1
BA	1
BC	1
A	1
AC	1

پیشوندهایی که به D ختم می‌شوند

D , sup(D) = 5

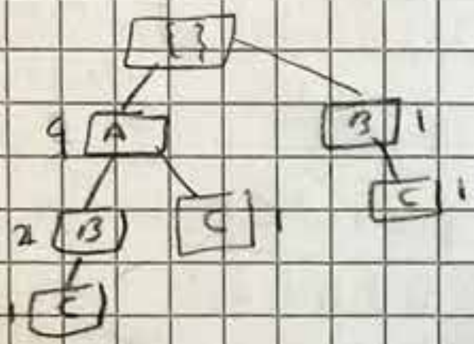


— از ته D وصل کنیم

BAC	1
BA	1
BC	1
A	1
AC	1

⟹

	sup
A	4
B	3
C	3



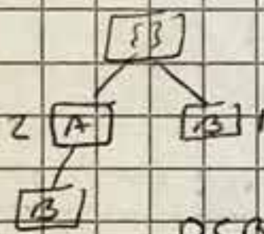
DC , sup(DC) = 3

حذف C

AB	1
B	1
A	1

→

	sup
A	2
B	2



DCB , sup(DCB) = 2

DCB , sup(DCB) = 2   حذف B

DCA , sup(DCA) = 2   حذف A

	sup
A	1

گام ① اول تک عضوی ها را بنویس ← بر اساس sup به صورت نزولی مرتب کن ← اونهایی که همه یکسان دارند
بر اساس حروف الفبا مرتب کن ←

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$ Euclidean distance

$p \mid \begin{matrix} x \\ y \end{matrix}$

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$ Minkowski dist

Similarity Between Binary vectors:

$$SMC = \frac{M_{11} + M_{00}}{M_{00} + M_{01} + M_{10} + M_{11}} = \frac{\text{number of matches}}{\text{number of attributes}}$$ Binary vectors

$M_{00} = P_0 q_0 \;/\; M_{01} = P_0 q_1 \;/\; M_{10} = P_1 q_0 \;/\; M_{11} = P_1 q_1$

(SMC versus Jaccard) $J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$ → no not-both-zero attr

Cosine similarity

$$\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{\|d_1\| \, \|d_2\|}$$

$d_1 =$ ○○○○○○
$d_2 =$ ○○○○○○

$= d_{11} \times d_{21} + \cdots$

$$\cos(\theta) = \frac{x_A x_B + y_A y_B}{\sqrt{x_A^2 + y_A^2} \times \sqrt{x_B^2 + y_B^2}}$$

$\|d_1\| = (d_{11} \times d_{11} + d_{12} \times d_{12} + \cdots)^{1/2}$

$\|d_1\| = (d_1 \times d_1)^{1/2}$

$-1 \leq \cos \leq 1$

Pearson's correlation

$$corr(x,y) = \frac{\text{covariance}(x,y)}{\text{standard\_deviation}(x) * \text{standard\_deviation}(y)} = \frac{S_{xy}}{S_x S_y}$$

$$\text{covariance}(x,y) = S_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$\text{standard\_deviation}(x) = S_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$ انحراف معیار x

$$\text{standard\_deviation}(y) = S_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$$ انحراف معیار y

$\bar{x}$ میانگین x ها

$\bar{y}$ میانگین y ها

classification:

accuracy score : $\dfrac{\text{جمع عناصر روی قطر اصلی}}{\text{مجموع کل عناصر ماتریس}}$ = $\dfrac{\text{عناصر درست‌ها}}{\text{کل عناصر}}$

- عناصر قطر اصلی correct
- سایر عناصر ماتریس incorrect

F1_score (measure) : $\dfrac{2P*R}{P+R}$

(percision) P = $\dfrac{\text{عدد روی قطر اصلی آن ستون}}{\text{مجموع عناصر ستون آن ستون}}$

(Recall) R = $\dfrac{\text{عدد روی قطر اصلی آن سطر}}{\text{مجموع سایر سطر آن سطر}}$

Recall = $\dfrac{\text{No of objects correctly asign to C}}{\text{No of obj belonging to C}}$

Precision : $\dfrac{\text{No of obj correctly assigned to C}}{\text{No of objs assigned to C}}$

Evaluation Regression

n: تعداد نمونه‌ها / تعداد سطر ماتریس / دیتای test انجام بشود

MAE (Mean Absolute Error) = $\frac{1}{n}\sum_i^n |y_i - \hat{y}_i|$ ($\hat{y}$: y pred)

MSE (Mean Squared Error) = $\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$

(tend to penelize less errors close to 0)

$R^2 = 1 - \dfrac{MSE}{std^2}$

$R^2 = \begin{cases} \simeq 1 & \rightarrow \text{good prediction} \\ < 0 \text{ (منفی)} & \rightarrow \text{wrong prediction} \end{cases}$

Minkowski Distance

$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r}$

[illegible]

- r=1 Manhattan distance (L_1 norm)
- r=2 Euclidean distance
- $r \to \infty$ supremum (L_{max} norm, L_∞ norm)

P	x	y
P_1		
P_2		

⇒

	P_1	P_2	P_3
P_1			
P_2			
P_3			

Euclidean : $\sqrt{\sum (x_i - y_i)^2} = d(x,y)$

$\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

```
np.sqrt((x-y)**2).sum(axis=1)
```

shape(x) = (100,50) اگر

shape(y) = (1,50)

⇓

(1,100) = euclidean خروجی

Gini index (classification)

$n = n_1 + n_2 + n_3$

$gini_a = 1 - (\frac{a_1}{n_1})^2 - (\frac{a_2}{n_2})^2 - \dots$

$gini_b = 1 - (\frac{b_1}{n_2})^2 - (\frac{b_2}{n_2})^2 - \dots$

$gini_c = 1 - (\frac{c_1}{n_3})^2 - (\frac{c_2}{n_3})^2 - \dots$

$\Rightarrow Gini\ x = \frac{n_1}{n} \times gini\ a + \frac{n_2}{n} \times gini\ b + \frac{n_3}{n}\ gini\ c$

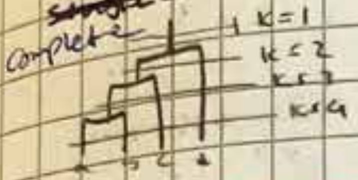
y

در کل در بین x و y خواستیم انتخاب کنیم آنکه عدد gini کمتری دارد مورد قبول است.

Hierarchical Clustering:

single-linkage(min) → در ماتریس دنبال کمترین مقدار باشیم ← نزدیکترین فاصله

~~single~~ complete-linkage(Max) → در ماتریس دنبال بیشترین مقدار باشیم



max or complete linkage → بیشترین فاصله / دورترین

Naive Bayes classifier

$x \in R^N \quad X = (x_1, x_2, \dots, x_n)$

$$p(class \mid X) = p(x_1 \mid class) * p(x_2 \mid class) * \dots * p(x_n \mid class) * p(class)$$

$p(x_1 \mid class)$: تعداد ان کلاس با x_1 / تعداد کل کلاس مورد نظر

$p(class)$: تعداد کل کلاس مورد نظر / جمع همه اشکال

Manhatan distance

$$dist(P_1, P_2) = |x_1 - x_2| + |y_1 - y_2|$$

DBSCAN
- core : Points with atleast minpoints neighbors within a distance ε
- border: at least a core point in the neighborhood
- noise : neither core nor border

$$A \Rightarrow B$$

support

$$sup = \frac{\#itemsets \text{ تعداد تکرار}}{\text{تعداد کل سطرها}} = \frac{\#\{A,B\}}{|T|}$$

confidence

$$conf = \frac{\text{itemset تعداد تکرار آیتم}}{\text{تعداد آیتم مورد نظر}} = \frac{sup(A,B)}{sup(A)}$$

→ $sup \geq minsup$

$conf \geq minconf$

Frequent Itemsets is an itemset whose support is greater than or equal to a minsup threshold

---

Rand Index

$$\text{Rand Index} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{TN + TP}{TN + FP + TP + FN}$$

دسته‌بندی واقعی		نتیجه خوشه بندی برای زوج (x_i, x_j)	
در دو دسته مجزا	در یک دسته		
FP f_{01}	TP f_{11}	در یک خوشه	خوشه بندی
TN f_{00}	FN f_{10}	در دو خوشه مجزا	

y_{pred} [, ,]

y_{true} [, ,]

در محاسبه RI جفت‌های داده را می‌سازیم

	y_t	y_p	TP	TN
0,1	0			
0,2				
0,3				
1,2				
1,3				
2,3				

→ هر عدد را با بقیه [illegible]

$TP = 1 \leftarrow y_{true} = 1, y_p = 1$

$TN = 1 \leftarrow y_{true} = 0, y_p = 0$

$0 : TN = TP \leftarrow y_{true} \neq y_p$

در RI تعداد کلاسترها hyper parameter نیست

Apriori Algorithm

Example DB

1	{A, B}
2	{B, C, D}
3	{A, C, D, E}
4	{A, D, E}
5	{A, B, C}
6	{A, B, C, D}
7	{B, C}
8	{A, B, C}
9	{A, B, D}
10	{B, C, E}

C_1 (حذف آیتم‌های کم‌تکرار)

	items	sup
1	A	7
2	B	8
3	C	7
4	D	5
5	E	3

L_1

C_2 (حذف آیتم‌های کم‌تکرار)

	items	sup
1	A, B	5
2	A, C	4
3	A, D	4
4	A, E	2
5	B, C	6
6	B, D	3
7	B, E	1
8	C, D	3
9	C, E	2
10	D, E	2

L_2

C_3 (حذف آیتم‌های کم‌تکرار)

	items	sup
1	A, B, C	3 ✓
2	A, B, D	2 ✓
3	A, B, E	
4	A, C, D	2 ✓
5	A, C, E	1
6	A, D, E	2 ✓
7	B, C, D	2 ✓
8	B, C, E	1
9	B, D, E	
10	C, D, E	1

L_3 ✓

minsup > 1

C_4

items	sup
A, B, C, D	1
A, B, D, E	
A, C, D, E	

$\Rightarrow L_4 = \emptyset$

* An Itemset is closed if none of its immediate supersets has the same support as the itemset.

accuracy = تعداد درست پیش‌بینی‌شده / تعداد کل

Precision (A) = تعداد درست پیش‌بینی‌شده A / تعداد پیش‌بینی‌شده A

recall = تعداد درست پیش‌بینی‌شده A / تعداد واقعی A

y_{true} [, , , ,]

y_{pred} [, , , ,]

		Predicted Class: +	Predicted Class: −
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	−	f_{-+} (FP)	f_{--} (TN)

← confusion Matrix (binary classifier)

$$accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

for positive class:

$$Precision(P) = \frac{a}{a+c} = \frac{TP}{TP + FP}$$

$$Recall(r) = \frac{a}{a+b} = \frac{TP}{TP + TN}$$

$$F\text{-measure}(F) = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

FP-tree:

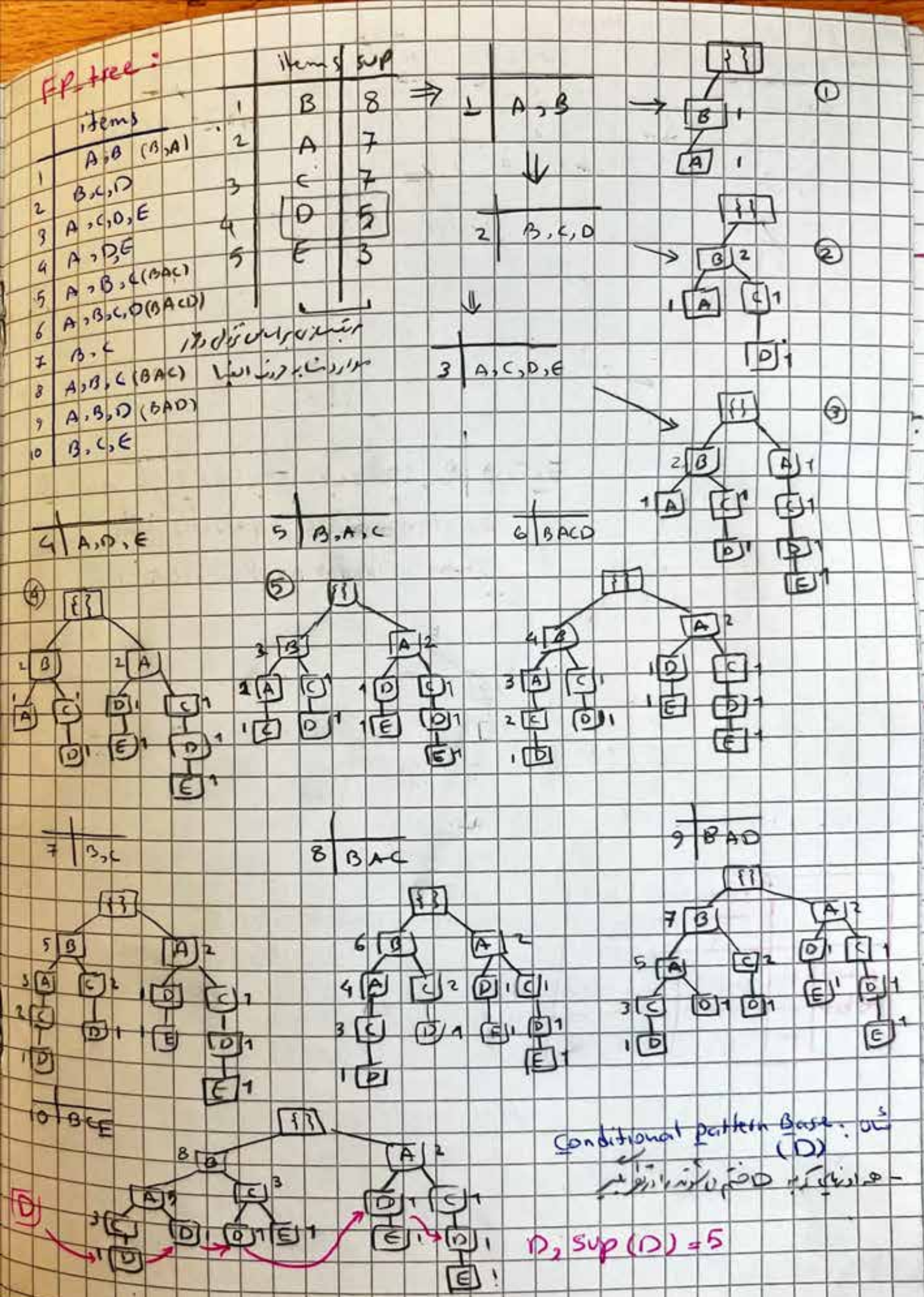
	items
1	A, B (B, A)
2	B, C, D
3	A, C, D, E
4	A, D, E
5	A, B, C (BAC)
6	A, B, C, D (BACD)
7	B, C
8	A, B, C (BAC)
9	A, B, D (BAD)
10	B, C, E

	items	sup
1	B	8
2	A	7
3	C	7
4	D	5
5	E	3

حذف آیتم‌های کمتر از min_sup

مرتب‌سازی براساس تکرار آیتم‌ها

1	A, B
2	B, C, D
3	A, C, D, E
4	A, D, E
5	B, A, C
6	BACD
7	B, C
8	BAC
9	BAD
10	BCE



Conditional pattern Base: مثلا (D)

- همه آیتم‌هایی که به D ختم می‌شوند را در نظر بگیر

*

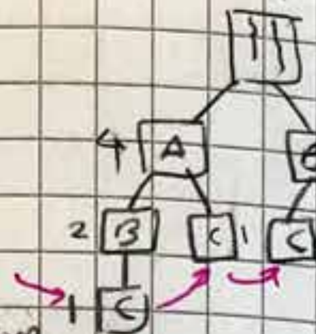
D در پایگاه آیتم ها / items	sup
BAC	1
BA	1
BC	1
A	1
AC	1

دوباره از اول ⇒

	items	sup
1	A	4
2	B	3
3	C	3

→ A حال باید حذف شود

هر دو یکسان



آنهایی که به C ختم می‌شوند را رسم کن

$DC , sup(DC) = 3$

	items	sup
1	AB	1
2	A	1
3	B	1

⇒

		sup
1	A	2
2	B	2

$DCB , sup(DCB) = 2$ آنهایی که به B ختم می‌شوند را رسم کن

		sup
1	A	1

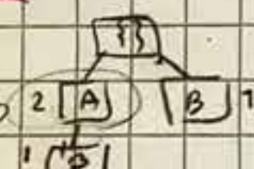
empty = DCB − CPB , از اینجا مشخص می‌شود که آیتم A باید از مجموعه حذف شود

از بین B و A را انتخاب کنیم چون تعداد sup یکسان داشتند

$DCA , sup(DCA) = 2$

		sup
	B	1

→



باقی می‌ماند B | 1 → DCA − CPB = empty

حال برمی‌گردیم به عقب ← (B) به این مرحله ← چون B و C هر دو sup=3 بودند یکی را انتخاب کردیم B و C را انتخاب می‌کنیم

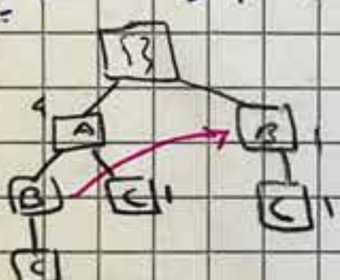
BAC	1
BA	1
BC	1
A	1
AC	1

→

A	4
B	3
C	3

$DB , sup(DB) = 3$

→



→

	sup
A	1

Correlation
- Covariance → same unit measurment / two variable / must be normalized
- Pearson → حساس به نویز / linear corr برای
- Spearman
  - 1 : 2 var correlated by an increasing monotonic func
  - -1 : Function is decreasing monotonic
  - 0 : there isn't a monotonic func correlation

Support Vector Regression

میانگین خطای مطلق :

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

میانگین مربعات خطا :

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

خطای استاندارد باقیمانده :

$$RSE = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

number of sample : n ← هرچقدر تعداد RSE کمتر مدل بهتر

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - FVU = 1 - \frac{MSE}{\sigma^2 (std)^2}$$

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

$$TSS = \sum_i (y_i - \bar{y}_i)^2$$

$$\bar{y} = \frac{1}{n} \sum_i y_i$$

predicted value : $\hat{y}$ / actual value : y

MAE, MSE هرچقدر مقدارشون به صفر نزدیکتر باشه، آن مدل بهتر است.

R^2 =
- $R^2 = 1$ | perfect linear relationship between x & y | 100% of the y variation is explained by variation in x
- R^2 close to 1 | very good linear relationship between x, y | good perdiction
- $0 < R^2 < 1$ | weaker linear relationship | A portion of the variation in y is not explained by variation in x
- $R^2 = 0$ | No linear relationship | the value of y doesn't depend on the value of x

It's upper bound +1
It has no lower bound ($-\infty$)
It is used as a regression evaluation metric

Micro average scores:

$$\text{micro-}P = \frac{\text{total\_TP}}{\text{total\_TP} + \text{total\_FP}}$$

$$\text{micro-}r = \frac{\text{total\_TP}}{\text{total\_TP} + \text{total\_FN}}$$

$$\text{micro-}f1 = \text{micro-}P = \text{micro-}r$$

Correlation or lift:

$$\text{Correlation} = \frac{P(A,B)}{P(A)P(B)} = \frac{\text{Conf}(r)}{\text{sup}(B)}$$

$r: A \Rightarrow B$

$$\text{correlation} \begin{cases} > 1 & \rightarrow \text{positive correlation} \\ = 1 & \rightarrow \text{statistical independence} \\ < 1 & \rightarrow \text{negative correlation} \end{cases}$$

$$\text{lift}(A,B) = \frac{C(A \rightarrow B)}{\text{sup}(B)} = \frac{S(A \cup B)}{S(A) \times S(B)}$$

ROC

- TPR (True Positive Rate) (on the y-axis) $= \frac{TP}{TP + FN}$
- FPR (False Positive Rate) (on the x-axis) $= \frac{FP}{FP + TN}$

The output of TF-IDF is:

A sparse matrix, since most terms will not occur in most documents

t

$W_{F_1} = \frac{11-2}{3} = \frac{9}{3} = 3$

$W_{F_2} = \frac{16-10}{3} = \frac{6}{3} = 2$

Bin 3 Bin 3 Bin 3

F_1: 2 — 5 — 8 — 11 (L | M | H)

F_2: 10 — 12 — 14 — 18

	F_1		F_2	
X_1	6	M	13	M
X_2	2	L	11	L
X_3	2	L	10	L
X_4	11	M	16	H

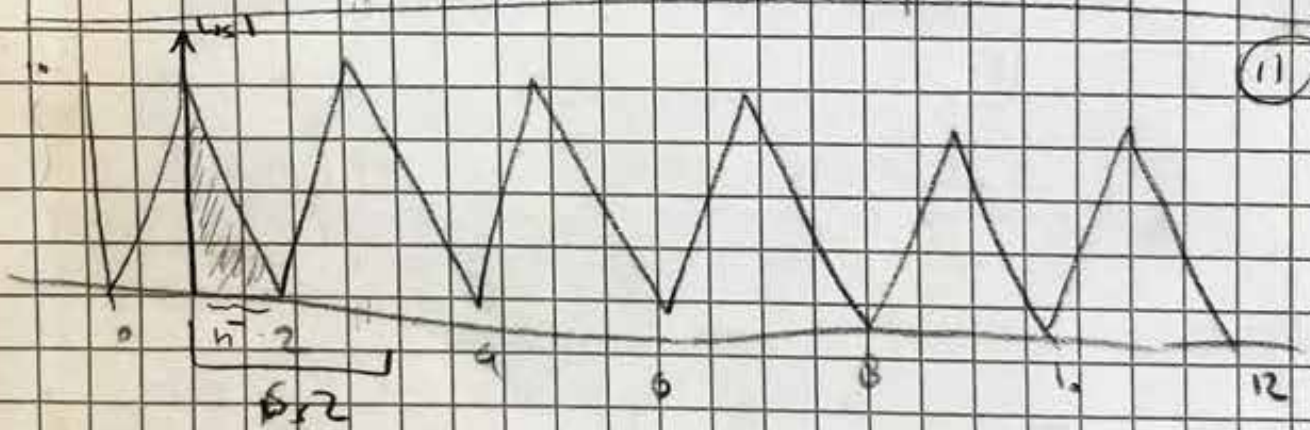
$X_1 = [M, M]$

$X_2 = [L, L]$

	age	Gender	occupation
0	23	M	Student
1	27	M	Consult
2	34	F	manager
3	44	F	manager
4	54	M	manager
5	29	M	Freelancer
6	21	F	Student

			age
Student	M	1	21
	F	1	
Consult	M	1	—
Freelancer	M	1	—
manager	F	2	34
	M	1	44
			54

→ means $= \frac{34+44+54}{3} = 44$



step = distance you move at each iteration

W = window size should be the piece you observe

W = 1 ST = 1 df.area.mean() = 5

		alpha	area	day	C_1	c_1	C_2	c_2
x1	high	0.7	5	wed	80	0	5	15
x2	low	0.8	10	mon	30			
x3	low	0.4	50	Fri				
x4	high	0.6	20	wed	30	5	2	70
					5/10	0/15	5/10	15/15 =1

$[c_2, c_1, C_2, C_2]$

highest average

$$X = \begin{matrix} a \\ b \\ c \\ d \end{matrix}\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 4 & 1 \\ 2 & 3 \end{bmatrix} \quad \begin{matrix} x_1 & x_2 & x_1x_2 & x_1^2 & x_2^2 \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 2 & 1 & 4 \\ 2 & 1 & 2 & 4 & 1 \\ 4 & 1 & 4 & 16 & 1 \\ 2 & 3 & 6 & 4 & 9 \end{bmatrix} \times \begin{bmatrix} 1 \\ -1 \\ 0.5 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1-2+1 \\ 2-1+1 \\ 4-4+2 \\ 2-3+3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 2 \\ 2 \end{bmatrix} \quad x_{poly}$$

$$+ \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = y_{pred}$$

(a) جواب

$$y_{true} \; [0 \;\; 1 \;\; 2 \;\; 3 \;\; 4]$$
$$[1 \;\; 3 \;\; 2 \;\; 3 \;\; 2]$$

$$y_{pred} \; [1 \;\; 1 \;\; 2 \;\; 3 \;\; 2]$$

-8 جواب

	y_{true}	y_{pred}	TP	TN	FP	FN
(0,1)	0	1	0	0	1	
(0,2)	0	0	0	1	0	
(0,3)	0	0	0	1	0	
(0,4)	0	0	0	1	0	
(1,2)	1	0	0	1	0	
(1,3)	1	0	0	1	0	1
(1,4)	1	0	0	1	0	
2,3	0	0	0	1	0	
2,4	1	1	1	0	0	
3,4	0	0	0	1	0	
			1	8	1	1

$$FM = \frac{1}{\sqrt{(1+1)(1+1)}} = \frac{1}{2}$$

$y_{true}=1$: FN
$y_{pred}=1$: FP

$$X = \begin{bmatrix} 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{bmatrix} \qquad Z = \begin{bmatrix} & & & \\ & & & \\ & & & \end{bmatrix}$$

(9 جواب)

$$Y = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 4 & 6 & 8 \\ 4 & 8 & 12 & 16 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} 3 & 4 \\ 7 & 10 \end{bmatrix}$$

2020-9-2

$y = [1\ 6\ 8]$ $N=3$

$y_{pred1} = [1\ 5\ 6]$

$y_{pred2} = [1\ 6\ 6]$

$\theta_1 = [1/3\ \ -2/3\ \ 1\ \ -1/3]$

$Q_2 = [4/3\ \ -6/3\ \ 2/3\ \ -3/3]$

$MSE(y, y_{pred})$

$\frac{1}{3}\left[(1-1)^2 + (6-5)^2 + (8-6)^2\right] + \left(\frac{1}{3} + \frac{-2}{3} + \frac{3}{3} + \frac{-1}{3}\right)$

$\frac{5}{3} + \frac{7}{9} \cdot (4) = (Y, Y_{pred1})$

$(Y, Y_{pred2})/(\theta_2) =$

$inter(b) = \frac{1}{2}(1+3) = \frac{4}{2}$ (20 min)

$inter(e) = \frac{1}{1}(1) = 1$

$extra(b), (C_2 \text{ نسبت}) = \frac{1}{2}(4+8) = 6$ $(C_3 \text{ نسبت}) = \frac{1}{2}(6+8) = 7$

$Sill(b) = \frac{4-4}{\max(4,4)} = \frac{4}{6} = \frac{2}{3}$

$extra(e), (C_2 \text{ نسبت}) = \frac{1}{2}(7+9) = 8$ $(C_1 \text{ نسبت}) = \frac{1}{3}(6+6+6) = 6$

$Sill(e) = \frac{4-1}{\max(6,1)} = \frac{5}{6}$

2022/1/20 ✓

X_0	X_1	X_2
c	b	x
a	a	z
c	c	y
a	a	y

	X_0-a	X_0-c	X_1-a	X_1-b	X_1-c	X_2-x	X_2-y	X_2-z	
x_1	0	1	0	1	0	1	0	0	← P
x_2	1	0	1	0	0	0	0	1	← q
x_3	0	1	0	0	1	0	1	0	
x_4	1	0	1	0	0	0	1	0	

M_{00} تعداد P و q هردو صفر
M_{01} تعداد q هایی که صفر است
M_{10} تعداد P هایی که صفر است
M_{11} تعداد P و q هردو یک

$$j(x_1,x_2) = \frac{M_{11}(x_1,x_2)}{M_{11}(x_1,x_2) + M_{01}(x_1,x_2) + M_{10}(x_1,x_2)}$$

$$j(x_1,x_2) = \frac{0}{0+3+3} = 0 \quad [M_{11}=0,\ M_{00}=1,\ M_{10}=3,\ M_{01}=3]$$

$$j(x_3,x_4) = \frac{1}{2+2+1} = \frac{1}{5} \quad [M_{11}=1 \quad M_{00}=3 \quad M_{10}=2 \quad M_{01}=2$$

ACD
CD
AC
C
BD
ABCD (CDAB)
CDE
ABC
AC
D
CD
AB
ABDE

	sup
A	7
B	5
C	8
D	8
E	2

	sup
C	8
D	8
A	7
B	5
E	2

CDA	1

B-CPB

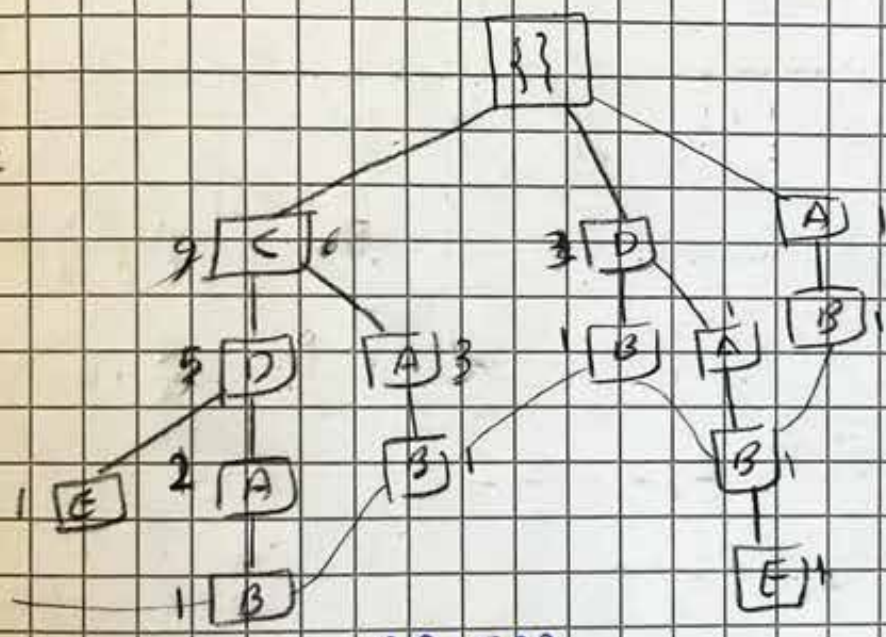
CDA	1
CA	1
D	1
DA	1
A	1

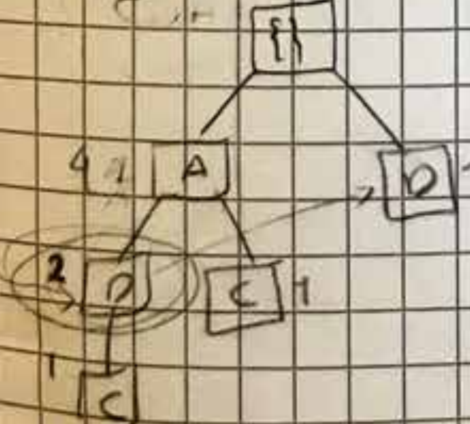
A1

B-CHT ✓

A	4
D	3
C	2

A2





BD-CPB

A	2

→ minsup D

AB

x	y	z	ground truth		
2	1	3	Accept	1+9+4 = 14 ≠ 15	X
0	2	-1	Accept	9+4+4 = 15	✓
-3	1	1	Accept	36+9+0 = 45	✓
1	3	-2	reject	4+1+9 = 14	X
0	0	0	Accept	9+16+1 = 26	✓

$R^2 = (x-3)^2 + (y-4)^2 + (z+1)^2$ (4 سؤال

اگر $r^2 > 15$ بود Accept
اگر $r^2 < 15$ بود Reject

$P(\text{Accept}) = \frac{3}{3}$ $\quad$ $P(\text{reject}) = \frac{1}{2}$ ← تعداد پیش بینی reject

$R(\text{Accept}) = \frac{3}{4}$ $\quad$ $R(\text{reject}) = \frac{1}{1}$ ← تعداد واقعی reject

17 سؤال

X	Y
2	-3
3	-4
2	-3
1	-2

$RSS = \frac{1}{n} \sum_i (y'_i - y_i)^2$

$RSS = \frac{1}{4}\left[(2m+3)^2 + (3m+4)^2 + (2m+3)^2 + (m+2)^2\right]$

$= 4m^2 + 12m + 9 + 9m^2 + 24m + 16 + 4m^2 + 12m + 9 + m^2 + 4m + 4$

$= 18m^2 + 52m + 38 = 0$

مشتق → $36m + 52 = 0 \quad 36m = -52 \quad m = \frac{-52}{36} = -1.44$

(8 سؤال

X_1	X_2
-6.7	-11.4
-3.3	-9.8
9.7	4.1
-4.9	-11.4
-4.1	-7.6
11.2	3.8
-4.2	-9.8
-6.6	-9.8
8.5	3.5
-4.7	-9.1

10 نمونه

knn k=3

نقطه

$3 \times 10 = 30$

$20 \times 10 \times 3 = 600$

$5 \times 20 \times 10 \times 3 = 3000$

تعداد، iteration، نقطه، نمونه

level = 0

high	width	weight
4	4	2
1	2	4
4	3	1
3	4	4
3	3	2
3	2	1
3	4	3
2	2	1

type: dataFrame

groupby high

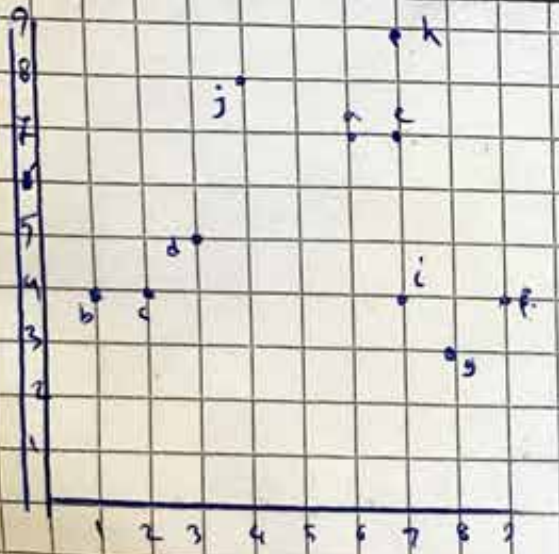
high		count
4	4	1
	3	1
1	2	1
2	2	1
3	2	1
	3	1
	4	2

unstack →

high \ width	2	3	4
1	1	NaN	NaN
2	1	NaN	NaN
3	1	1	2
4	NaN	1	1

← width

↑ index



a	B	0
e	C	0
h	B	0
i	C	1
g	C	1
f	C	1
b	B	2
c	C	2
d	B	2
j	N	-1

	x	y	z	L	
a	5	4	2	B	0
b	4	3	2	C	0
c	1	2	4	C	1
d	2	2	4	B	1
e	3	2	2	C	0
f	1	1	5	C	1
g	2	1	2	B	0
h	3	1	5	B	1
i	5	1	4	N	-1

(a,b) (1,1,-1) < 2 ✓
(a,c) (0,3,2) < 5 ×
(a,e) (2,2,0) < 4 ×
(b,e) (1,1,0) < 2 ✓
(b,g) (2,2,0) < 4 ×
(e,g) (1,1,0) < 2 ✓
(e,d) (1,0,2) < 3 ×
(e,h) (0,1,3) < 4 ×
(d,g) (0,1,2) < 3 ×
(d,c) (1,0,-1) < 1 ✓
(d,f) (1,1,1) < 3 ×
(d,h) (1,1,1) < 3 ×
(c,f) (0,1,1) < 2 ✓
(c,g) (1,1,2) < 4 ×
(c,e) (2,0,2) < 4 ×
(c,f) (0,1,1) < 2 ✓
(f,h) (2,0,0) < 2 ✓
(g,f) (1,0,3) < 4 ×
(g,h) (1,0,3) < 4 ×

① ✓

② ✓

Ytrue = [B B C B A B C C]

Ypred = [A B C A A B B A]

accuracy = 4/8

precision(A) = 1/4 (1,4) (4,2) (5,0)

recall(A) = 1/1

③ ✓

	C_1	C_2	C_3	
a(3,5)	3	4	7	C_1
b(5,4)	4	3	4	C_2
c(2,3)	2	3	6	C_1
d(3,3)	3	2	5	C_2
e(4,3)	4	1	4	C_2
f(6,3)	6	3	4	C_2
g(5,2)	6	1	2	C_2
h(0,1)	4	5	6	C_1
i(2,0)	5	4	3	C_3

$C_1 = \{a, c, h\}$

$C_2 = \{b, d, e, f, g\}$

$C_3 = \{i\}$

$C_3(x,y) = (2,0)$

$C_1(x,y) = \frac{1}{3}\begin{pmatrix}(3,5)\\(2,3)\\(0,1)\end{pmatrix} = \frac{1}{3}(5,9) = (\frac{5}{3}, \frac{9}{3})$

$C_2(x,y) = \frac{1}{5}\begin{pmatrix}(5,4)\\(3,3)\\(4,3)\\(6,3)\\(5,2)\end{pmatrix} = (\frac{23}{5}, \frac{15}{5})$

Minkowski Distance:

$r=1$ Manhatan distance

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots + |x_{ip}-x_{jp}|$$

$r=2$ Euclidean distance

$$d(i,j) = \sqrt{|x_{i1}-x_{j1}|^2 + |x_{i2}-x_{j2}|^2 + \cdots + |x_{ip}-x_{jp}|^2}$$

$r \to \infty$ supremum distance

$$d(i,j) = \lim_{r\to\infty} \left(\sum_{k=1}^{n} |x_{ik}-x_{jk}|^r \right)^{1/r} = \max_{k}^{n} |x_{ik}-x_{jk}|$$

$\hookrightarrow \max\{|x_{i1}-x_{j1}|, |x_{i2}-x_{j2}|\}$

& manhatan

Point	x	y
P1	0	2
P2	2	0
P3	3	1
P4	5	1

$r=1$ (L1)

	P1	P2	P3	P4
P1	0	4	4	6
P2	4	0	2	4
P3	4	2	0	2
P4	6	4	2	0

$d(P_1,P_2) = |x_2-x_1| + |y_2-y_1| = |2-0| + |0-2| = 4$

$d(P_1,P_3) = |x_3-x_1| + |y_3-y_1| = 1$

$d(P_1,P_4) = |x_4-x_1| + |y_4-y_1| = 1$

$d(P_2,P_3) = |x_3-x_2| + |y_3-y_2| = 1$

$d(P_2,P_4) = |x_4-x_2| + |y_4-y_2| = 1$

$d(P_3,P_4) = |x_4-x_3| + |y_4-y_3| = 1$

$r=2$

L_2	P_1	P_2	P_3	P_4
P_1	0			
P_2		0		
P_3			0	
P_4				0

$$d(P_1,P_2) = \sqrt{|x_2-x_1|^2 + |y_2-y_1|^2} = \quad \text{Euclidean}$$

$$= \quad \checkmark$$

$r \to \infty$

L_∞	P_1	P_2	P_3	P_4
P_1	0	2	3	
P_2	2	0		
P_3	3		0	
P_4				0

$$d(P_1,P_2) = \max\{|x_2-x_1|, |y_2-y_1|\}$$
$$= \{2,2\} = 2$$
$$d(P_1,P_3) = \max\{|x_3-x_1|, |y_3-y_1|\}$$
$$= \{3,1\} = 3$$

---

① SMC : (versus Jaccard)

Similarity between binary vectors

M_{01} $p=0, q=1$ تعداد موارد که M_{00} $p=0$ $q=0$

M_{10} $p=1, q=0$ M_{11} $p=1$ $q=1$

$p = 0\ 1\ 0\ 0\ 0\ 0\ 1\ 0$

$q = 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 1$

② Cosin Similarity

$$\cos(d_1,d_2) = (d_1 \cdot d_2) / \|d_1\| \, \|d_2\|$$

$$\sqrt{(d_1 \times d_1)} \times \sqrt{d_2 \times d_2}$$

$d_1 = 3\ 2\ 0\ 2\ 1\ 1$

$d_2 = 1\ 1\ 5\ 4\ 0\ 1$

$$\cos(d_1,d_2) = \frac{3+2+0+8+0+1}{\sqrt{3\times3+2\times2+0\times0+2\times2+1\times1+1\times1} \times \sqrt{1\times1+1\times1+5\times5+4\times4+0\times0+1\times1}}$$

$$= \frac{14}{\sqrt{19} \times \sqrt{44}}$$

classification techniques:

1- Decision trees
2- classification rules
3- Association Rules
4- Neural Networks
5- Naïve Bayes and Bayesian Networks
6- K-Nearest Neighbours (K-NN) → Euclidean
7- Support Vector Machines (SVM)

clustering Algorithms:

1- K-means and its variants → SSE
2- Hierarchical clustering
3- Density-based clustering

① evaluating k-means:

SSE (Sum of squared Error)

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

one way to reduce SSE is to increase K, the number of clusters
(A good clustring with smaller K can have a lower SSE than a poor clustring with higher K)

Running K-means with K=10 on the same dataset will typically produce lower mean SSE

Rand Index: $\dfrac{f_{00}^{TN} + f_{11}^{TP}}{f_{00} + f_{01}^{FP} + f_{10}^{FN} + f_{11}}$

TN f_{00} number of pairs of objects having a different class and a different cluster (0, 0)
FP f_{01} — — different class (0), same cluster (1)
FN f_{10} — — same class (1), different cluster (0)
TP f_{11} — — same class (1), same cluster (1)

```
numbers = [1, -8, 5, -2, 5]
```

Filter: filter the element of a list based on a condition

فیلتر امکان چک کردن یک شرط را روی تمام اعضای یک لیست فراهم می‌کند.
فیلتر کن با توجه به شرطی که به آن میدهم

```
neg = list(filter(lambda x: x<0 , numbers))
neg → [-8, -2]
```

map: map each element of a list with a new value

```
sqrd = list(map(lambda x: x**2 , neg))
sqrd = [64, 4]
```

مپ این امکان را به ما می‌دهد که یک تابع را روی تمام المان‌های یک لیست اعمال کنید.

```
map(function, list)
```

مپ کن این فانکشن را روی تک تک عناصر لیست

مثال: R = List(map(lambda x: 'big' if x>10 else 'small', number))

('big': اگر شرط برقرار بود — x>10: شرط — 'small': اگر شرط برقرار نبود — number: لیست ورودی)

```
filter(function, list)
```

* فیلتر مانند مپ فانکشن را دونه دونه روی المان‌های لیست اعمال می‌کنه اگر True بود برش می‌گردونه اگر False بود از روش رد میشه.

partitional clustering

k-means

$$C_x = \frac{\sum x_i}{n} \qquad C_y = \frac{\sum y_i}{n}$$

محاسبه C مرکز کلاستر C_i

در الگوریتم k-means هدف این است که مراکز خوشه‌ای انتخاب شوند که مقدار زیر مینیمم شود.

$$\sum_{i=1}^{N} \min \| x^{(i)} - c_j \|^2 \rightarrow$$

زمان به بیان فاصله هر x از نزدیک‌ترین مرکز خوشه‌ای

می‌توانیم از رابطه اقلیدسی هم استفاده کنیم

$$\Rightarrow$$

یا چون با یک Rand Index برای اندازه‌گیری شباهت خوشه‌بندی‌ها را می‌سنجد

میانگین: $\bar{X} = \frac{\text{مجموع مقادیر } x}{\text{تعداد } n} = m$

x	y
x_1	x_2
a	d
b	e
c	f

واریانس: $V(x) = \frac{1}{\text{تعداد} - 1} \left[(a-m)^2 + (b-m)^2 + (c-m)^2 \right]$

کوواریانس: $COV(x_1, x_2) = \frac{1}{n-1} \left[(a-m)(d-m) + (b-m)(e-m) + (c-m)(f-m) \right]$

$$cov(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

در آرایه‌ها [سطر ، ستون ، عمق]

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

axis

np.sort(a1, axis = 1) — اعداد را در هر آرایه مرتب می‌کند

np.sort(a1, axis = 0) — مثل قبلی می‌باشد ولی در ترتیب اعداد هیچ تغییری نمی‌دهد.

np.sort(a1, axis = None) — کل آرایه را در یک لیست یک‌بعدی به ترتیب مرتب می‌کند

a.sum(axis = 0) — mean(axis=0) میانگین هر ستون — جمع هر ستون را برمی‌گرداند

a.min(axis = 1) — min هر سطر را برمی‌گرداند

np.sort(a1, axis = -1) — مرتب سازی در طول سطر

axis = -1, 1 : در طول سطر

axis = 0 : در طول ستون

accessing Series Elements

1) by index

```
S1 = pd.Series({'a1':1, 'a2':2, 'a3':3})
S1 = pd.Series([1, 2, 3], index=['a1', 'a2', 'a3'])
```

- Explicit → print(S1.loc['شماره داده شده در index'])
- Implicit → print(S1.iloc['شماره اندیس داده‌ها'])

به این طریق کار، slicing انجام می‌شود.

```
print(S1.loc['a1']) = 1
print(S1.iloc[1]) = 2
print(S1.loc['a1':'a2']
```

explicit → Stop element included

```
print(S1.iloc[1:3])
```

Implicit → Stop element excluded

```
print(S1.loc[[a1, a2]])
```

a1	1
a2	2

Q: A decision tree trained on a dataset with categorical features only:

A: Each internal node of a decision tree can have at most C children, where C is equal to the number of distinct values of the attribute associated with the internal node.

$RI(y_{true}, y_{pred}) = \frac{a+b}{\binom{n}{2}}$ $\quad [\quad]_{n \times m}$

unlabeled dataset X (n rows, m columns) and run a cluster analysis using DBSCAN

* The number of clusters is not a hyper parameter that can be fine-tuned (based on the knee of the curve of the sum of squared errors (SSE))

* The number of clusters extracted is dependent of the hyperparameter ϵ (DBSCAN)

* DBSCAN may extract at most (n) cluster

* You can use the RI to assess the quality of the resulting classification (not clusters)

Feature → x_0 x_1 x_2 ...

$= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{n \times m}$ ← sample

n → sample

m → feature

Apriori Algorithm   min sup ≥ 2

12 سوال
6-9-2021

	transaction
0	BC
1	AE
2	ACD
3	AB
4	BE
5	DE
6	BCE
7	BC
8	BCD
9	ABCD

	Item	Sup
C	A	4
C	B	7
C	C	6
C	D	4
C	E	4

	Item	sup
	AB	2
N	AC	2
N	AD	2
	AE	1 ×
C	BC	5
N	BD	2
	BE	2
C	CD	3
	CE	1 ×
	DE	1 ×

sup بزرگتر مساوی با ۲

Item	① after prune	② sup بعد
ABC		1 ×
ABD		1 ×
ABE	×	
ACD		2 •
ACE	×	
BCD		2 •
BCE	×	
BDE	×	
CDE	×	

ABCD ×
ABCE ×
ABDE ×
ACDE ×
BCDE ×

⇓

L = ∅

q1 = list all frequent Itemsets having length 2 along with the support count

A1 : { AB:2 , AC:2 , AD:2 , BC:5 , BD:2 , BE:2 , CD:3 }

q2 = list all itemsets of length 3 that have been generated by Apriori after the join and prune steps. (before counting their support in the database)

A2 = { ABC , ABD , ACD , BCD }

نکته: تمام حالتهای ممکن ۳ تایی را می‌سازیم (join) بعد به مرحله قبل آن نگاه می‌کنیم آنهایی که minsup کمتر داشته‌اند
در حالت دوتایی، در این مرحله از ترکیبات ۳ تایی حذف می‌شوند (prune)

q3 = list all frequent itemsets that are **not closed** along with their support count.

A3 = { AC:2 , AD:2 , BD:2 }

① حذف مرحله‌ای که از آیتم‌ست‌ها ...
② در ست نهایی sup بعد از حساب کردن که باقی مانده نگاه می‌کنیم (حالا می‌خواهیم طول ۳ تایی‌ها را داشته باشیم) [•]
③ به ترکیبات سازنده هر مرحله قبل آن نگاه می‌کنیم [مثلاً ACD ← AC, AD, CD] و sup آنها را مقایسه می‌کنیم
sup(ACD) = 2 ← sup(AC) = 2 , sup(AD) = 2 , sup(CD) = 3 اگر sup دو طرف یکسان بود که آن آیتم‌ست
not closed
در غیر اینصورت closed
④ این مرحله را تا رسیدن به اولین آیتم‌ست‌های با طول 1 (A, B, C, ...) ادامه می‌دهیم [illegible].

نکتہ نہ آید سہ سعدی راست ۔ [سعدی، سنہ وفات]