POLITECNICO
DI TORINO

Data Science Lab

Exercises

DataBase and Data Mining Group

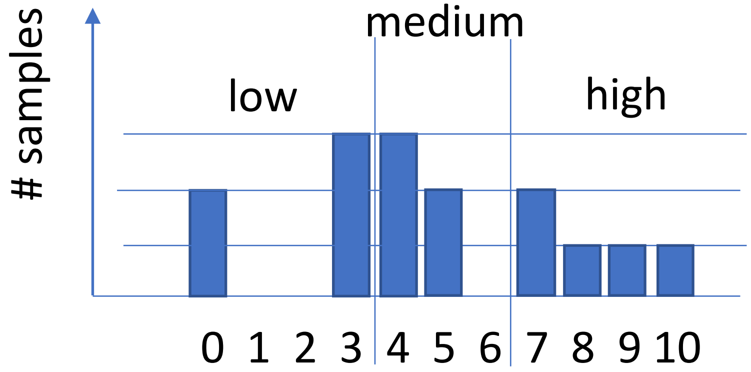
Andrea Pasini, Elena Baralis

1. Preprocessing

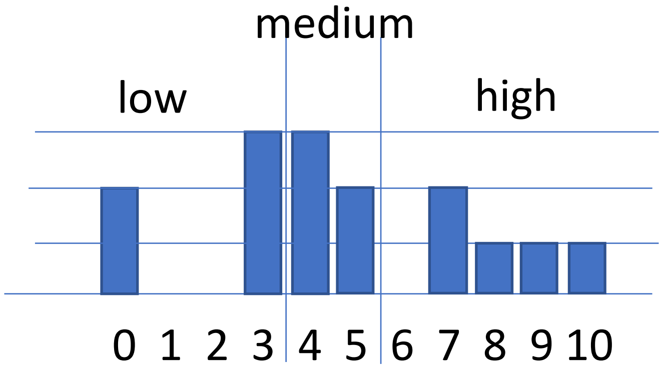
- The following list represents training set values of a specific attribute.
  - [10, 0, 5, 3, 3, 0, 3, 4, 4, 7, 5, 7, 8, 4, 9]
- Use these values to train an equal-frequency based discretization with three bins (low, medium, high). Which statement is correct?
  a) The test vector [1, 7, 9] is discretized 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]

1. Preprocessing

- Solution:
  - Draw the data distribution (5 elements for each bin):



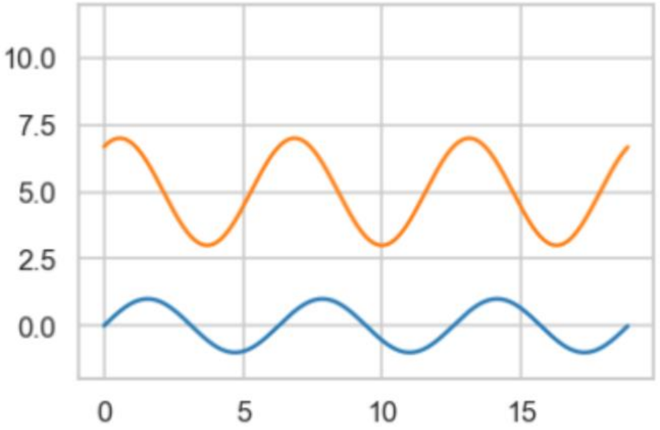
or, equivalently:



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

2. Time series

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



a) Mean, first derivative

b) Mean, percentiles

c) First derivative, percentiles

d) Percentiles, frequency

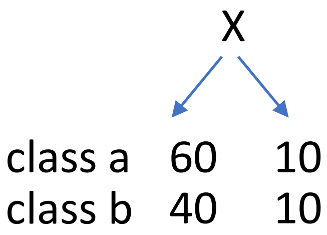
e) All of the pairs above are equivalent for distinguishing between the two series

2. Time series

- Solution: b)
- Mean and percentiles are significant since they both present different values for the two series
- The derivative of a time series is still a time series (not a feature)
- Frequency is equal for the two time series, hence not important

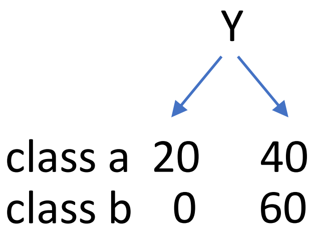
3. Classification

- 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



X

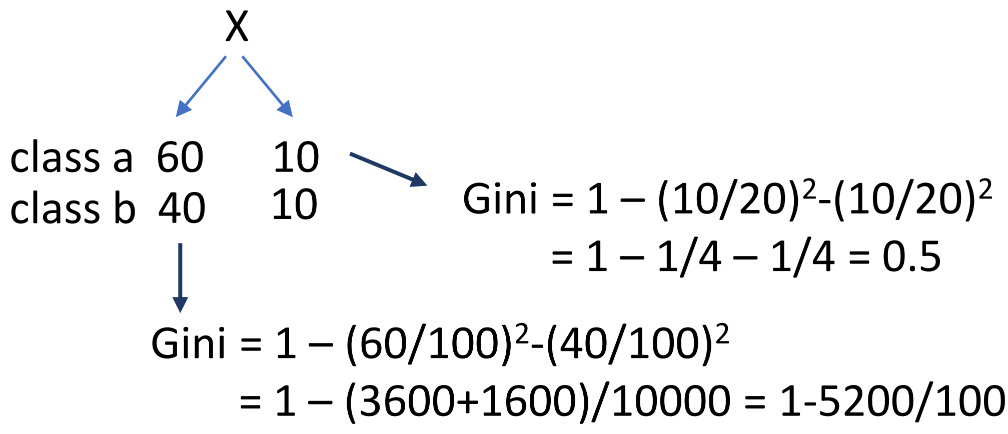
class a	60	10
class b	40	10



Y

class a	20	40
class b	0	60

3. Classification



X

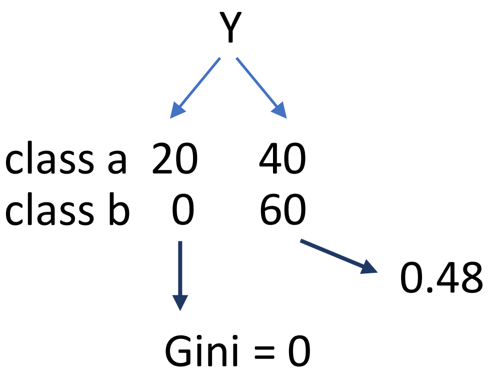
class a 60 10

class b 40 10

Gini = 1 – $(10/20)^2$-$(10/20)^2$
= 1 – 1/4 – 1/4 = 0.5

Gini = 1 – $(60/100)^2$-$(40/100)^2$
= 1 – (3600+1600)/10000 = 1-5200/10000 = 1 - 0.52 = 0.48

→ Gini(X) = 100/120*0.48 + 20/120*0.5 = **58/120** = 0.4833



Y

class a 20 40

class b 0 60

0.48

Gini = 0

→ Gini(Y) = 20/120*0 + 100/120*0.48 = **48/120** = 0.4

-> Correct answer is b. The algorithm will choose split Y (0.4<0.4833)

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

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

4. Hierarchical clustering

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

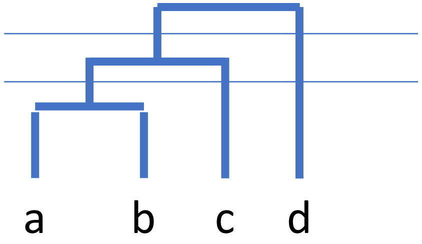
merge a, b

min(ac, bc)

	ab	c	d
ab	0	**2**	**3**
c		0	3
d			0

min(ad, bd)

merge ab, c



k=2: abc, d

k=3: ab, c, d

- Correct answer: e) all the statements are correct