Machine Learning & Data Mining

Indholdsfortegnelse

Spørgsmål: KDD, Datamining statements, Clust. statements

Fundamentals

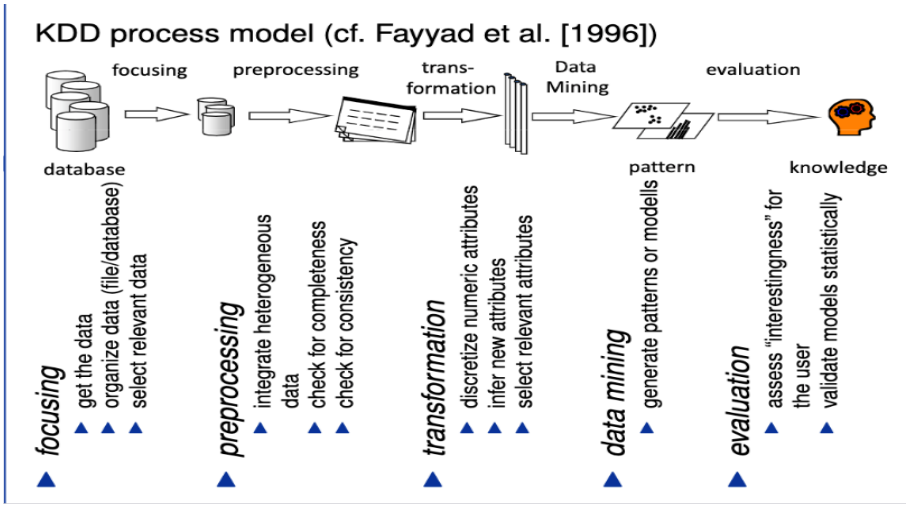
1. Which of the following steps are part of the KDD process (according to the process model by Fayyad et al.)?

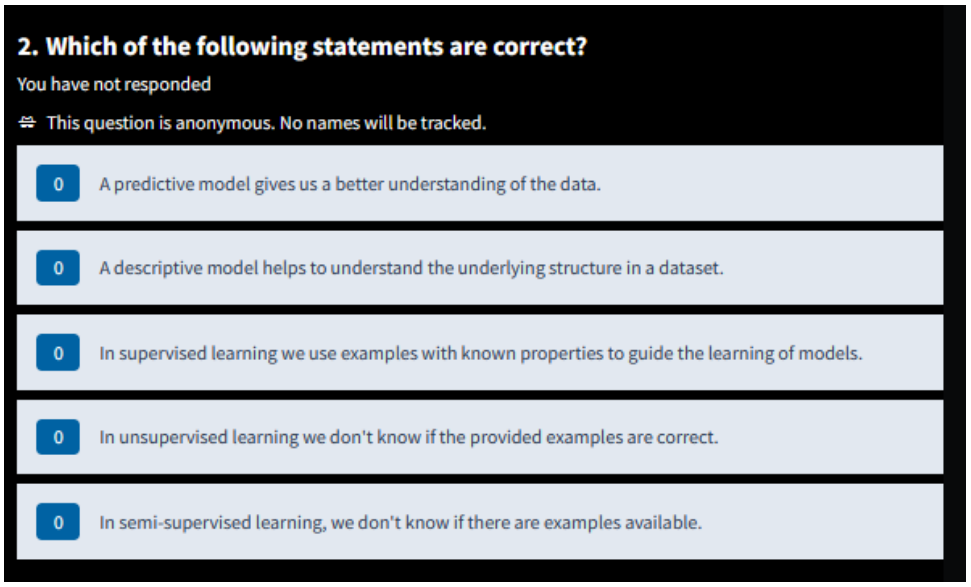
You have not responded

This question is anonymous. No names will be tracked.

- 0 Coloring
- 0 Collecting
- 0 Focusing
- 0 Translation
- 0 Transposing
- 0 Transformation
- 0 Postprocessing
- 0 Evaluation
- 0 Communication
- 0 Selling

Correct Answers: 3, 6 and 8





Correct Answers: 2,3

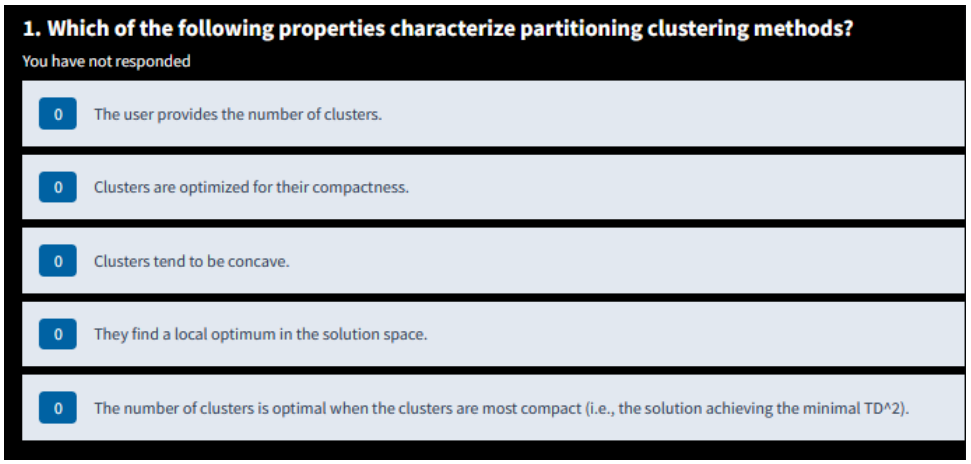
Direkte svar:

Answer 2 true: Descriptive analytics looks at data statistically to tell you what happened in the past.

Answer 3 true:

1. Supervised learning is machine learning approach that is defined by its use of labeled datasets. In supervised learning, the algorithm “learns” from the training dataset by iteratively making.
2. Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets.

What is a dataset with no labels, you ask? Well, it is a dataset with only **features**, and no **target to predict**. For example, if our housing dataset **had no prices**, then it would be an unlabeled dataset.



Correct Answers: 1, 2 and 4

(Hvorfor 5'eren ikke er rigtig – td^2 måler bare hvor kompakte ens clusters er, men sillohoute måler strukturen af clusterne. Derfor kan vi ikke kun være afhængig af TD^2, da den kun måler hvor kompakt clusterne er, men vi skal derfor bruge sillhoute til at fortælle noget om clusternes strukturer)

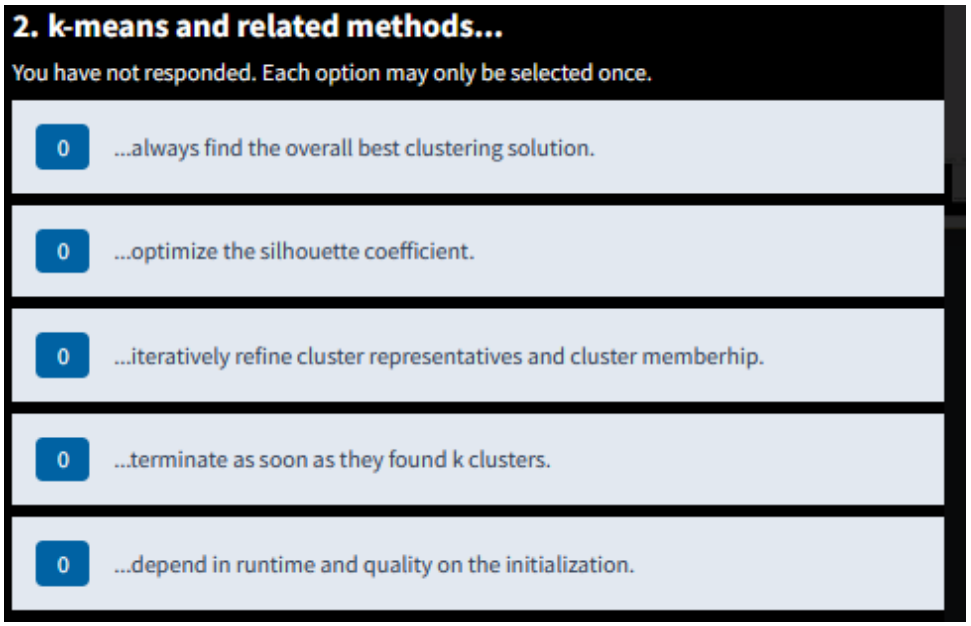
Direkte svar:

Partitional clustering partitions a dataset into k clusters, typically minimizing some cost function (compactness criterion).

Central assumptions for approaches in this family are typically:

- number k of clusters known (Answer 1)
- clusters are characterized by their compactness (Answer 2)
- compactness measured by some distance function (e.g., distance of all objects in a cluster from some cluster representative is minimal)
- criterion of compactness typically leads to convex or even spherically shaped clusters
- Locally optimizing algorithms (Answer 4)

Der henvises til clustering kapitlet i PP og docs dokumentet.



Correct Answers: 3, 5

Direkte svar:

Answer 3

- 1.Choose a first centroid randomly.
- 2.Compute for each point the distance to the closest of the already existing centroids.
- 3.Choose a new centroid from the points with a probability proportional to the squared distance.
- 4.Repeat 2 and 3 until k centroids have been chosen (that's why question 4 is wrong)

Answer 5

Clustering on a small sample usually delivers good **initial** clusters, but some samples are deviating too much from the overall data distribution. Therefore k-means and related methods depends on the runtime and quality of the initialization (Choose of right points as centroids, choosing the wrong ones, would make the process long as we would have to make many more iterations)

Der henvises til clustering kapitlet i PP og docs dokumentet.

1. Which of the following statements about classification are true?

You have not responded. Each option may only be selected once.

0	If a classifier has a stronger bias, the trained hypothesis will be less close to the target function.
0	The weaker the bias, the better a classifier's hypothesis can approximate the target function on the training data.
0	The weaker the bias, the more susceptible some learner is to overfitting.
0	The "apparent classification error" tells us how good a classifier performs on test data.
0	With cross-validation, we can use all available data for training and for testing.

Correct Answers: 2, 3 and 5

Direkte svar:

Right answers

Answer 2 (right): By looking at the below diagram we can see when bias decreases it means we get a more complex model and our predictions error will fall, which can provide a better target function.

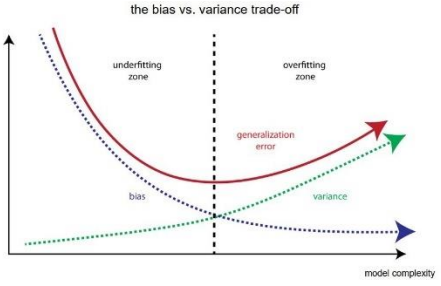
Answer 3 (right): We can see when bias decreases, this means our model is more complex and we are adding more variables which can result in overfitting. The model becomes more complex due to overfitting and gives us a high variance.

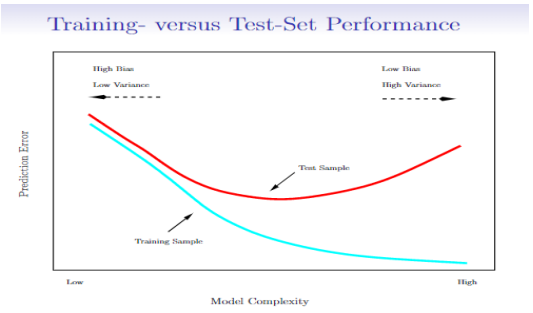
Answer 5 (right): Theoretically, splitting a data set into 50-50 split, using one half of the data to generate a model, and then applying it on the other half, can produce an output (y), which is much different than the output produces if all the data was used.

Wrong answers

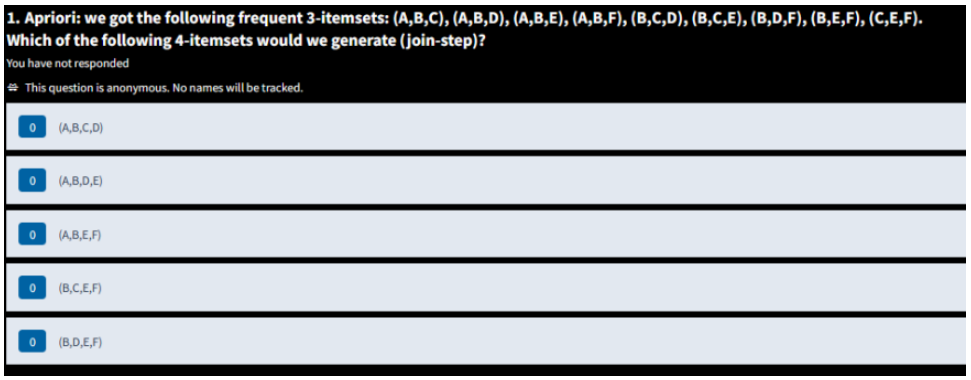
Answer 1 (wrong): If we look at the training-versus Test-Set Performance diagram, we can see when we have high bias on the training sample it results with a high prediction error, which is deviates from the target function.

Answer 4 (wrong): The apparent classification error estimates the misclassification probability by applying the estimated discriminant rule to classify the training set (not test-set)





Spørgsmål: Frequent pattern mining



Correct Answers: 1, 2 and 3

Direkte svar:

Definition:

- **Join Step:** This step generates (K+1) itemset from K-itemsets by joining each item with itself. This means: 3-itemset like (ABC), we can join with (ABD) and from there generate an 4-itemset (ABCD). We already had AB in our 3-itemset (ABD), therefore we only included from 3-itemset (ABD).

Joining: Two frequent (k − 1)-itemsets are joined if they are identical in the first k − 2 items.

Start by testing one of the frequent 4-itemsets and see if it possible to generate this by one of the given 3-itemsets.

From this we can answer:

(ABCD) ->> can be generated from (ABC) + (ABD), Therefore this statement is TRUE.

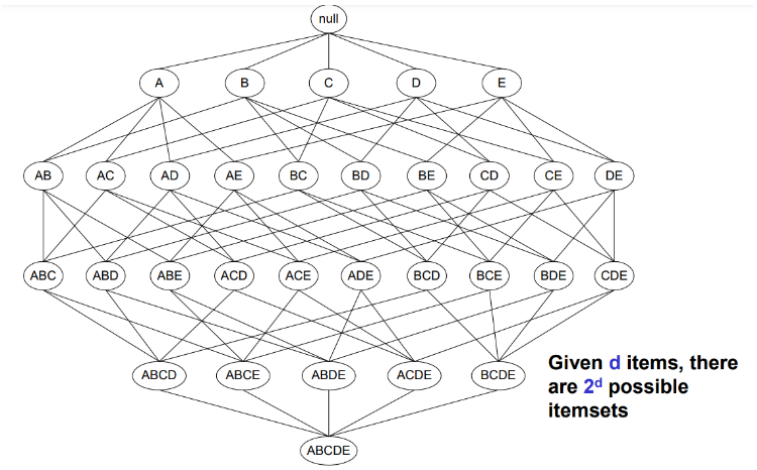
(ABDE) ->> can be generated from (ABD) + (ABE), Therefore this statement is TRUE.

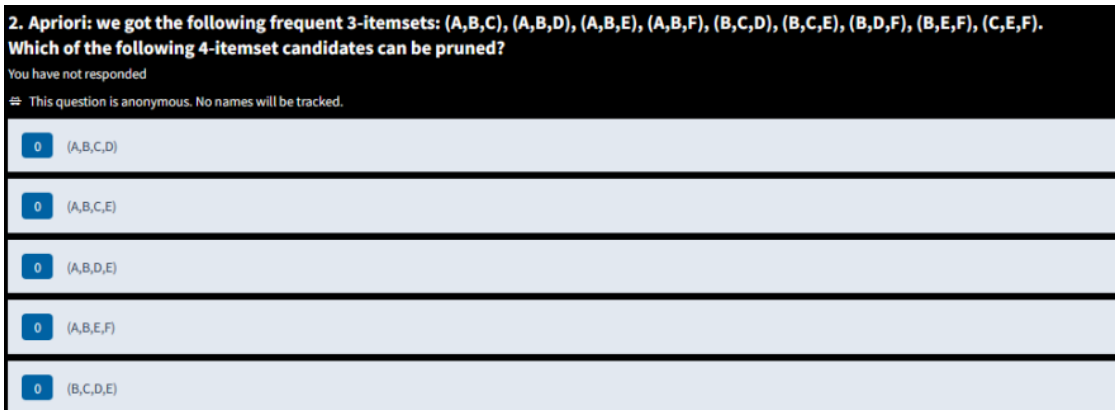
(ABEF) ->> can be generated from (ABE) + (ABF), Therefore this statement is TRUE.

(BCEF) ->> it is NOT possible to generate this, Therefore this statement is WRONG.

(BDEF) ->> it is NOT possible to generate this, Therefore this statement is WRONG.

- **It is possible to generate an 4-itemset (BCDE), from (BCD) + (BCE), but it is not possible to generate (BCEF) as (BCE) + (BDF) is not possible as 'BC' is not in (BDF), therefore it not possible to join these two 3-itemsets. Same case with (BDEF). Therefore these 4-itemsets is wrong.**





Correct Answers: All answers

Direkte svar:

Definition: Prune Step: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

Start by testing one of the frequent 4-itemsets and see if it possible to generate this by one of the given 3-itemsets. If is possible to generate this 4-itemsets, it is then also possible to prune it.

"If an itemset can be joined, it can also be pruned"

From this we can answer:

(ABCD) ->> can be generated from (ABC) + (ABD), Then this mean it can be pruned. Therefore this statement is TRUE.

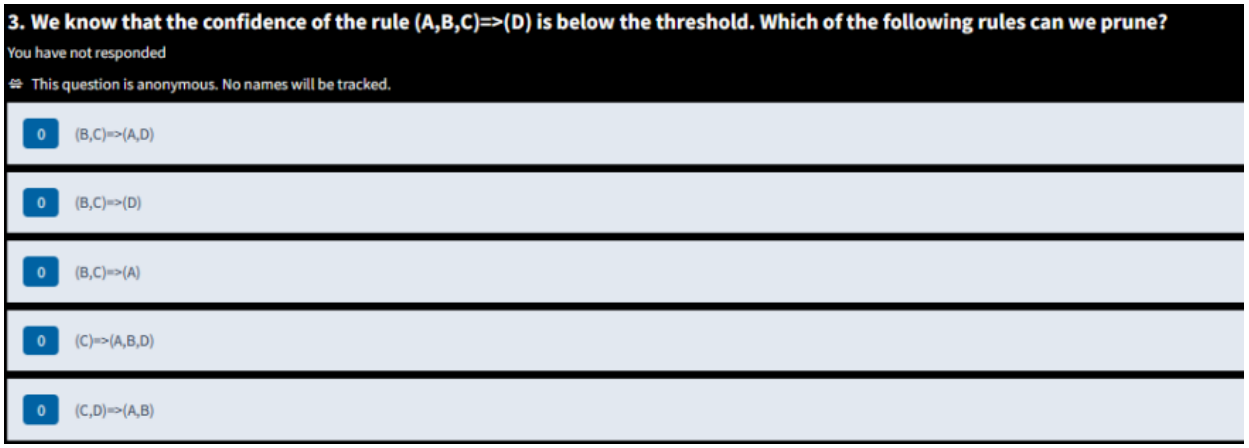
(ABCE) ->> can be generated from (ABC) + (ABE), Then this mean it can be pruned. Therefore this statement is TRUE.

(ABDE) ->> can be generated from (ABD) + (ABE), , Then this mean it can be pruned. Therefore this statement is TRUE.

(ABEF) ->> can be generated from (ABE) + (ABF), , Then this mean it can be pruned. Therefore this statement is TRUE.

(BCDE) ->> can be generated from (BCD) + (BCE), Then this mean it can be pruned. Therefore this statement is TRUE.

Therefore all answers are TRUE.



Correct Answers: 1, 4

Direkte svar:

Definition:

According to theorem 2.1:

Given:

- itemset X
- $Y \subset X, Y \neq \emptyset$

If $\text{conf}(Y \Rightarrow (X \setminus Y)) < c$, then $\forall Y' \subset Y$:

$$\text{conf}(Y' \Rightarrow (X \setminus Y')) < c.$$

hvis: $Y \rightarrow Y - X$ Ikke overholder minimum confidence **c's** threshold, så angiver det at alle regler/subsets for $Y' \rightarrow X - Y'$, Hvor Y' er et subset af Y, heller ikke vil overholde minimum confidence threshold. # Finding the Silhouette score for S2

Therefore:

We have the association rule (A,B,C)=>(D), which we are told is below a threshold.

To know which other rules, which are also below the threshold and which we then can prune, we should look after two things:

1. ALL of the 4 items A,B,C and D should be represented in the association rules
2. Items on the right side of the =>, can not be moved to the left side. In our case D, should always be on the right side.

Knowing these rules we can answer:

(BC) => (AD) have all 4 items of (A,B,C)=>(D) represented and 'D' is not on the left side of =>, therefore this statement is **TRUE.**

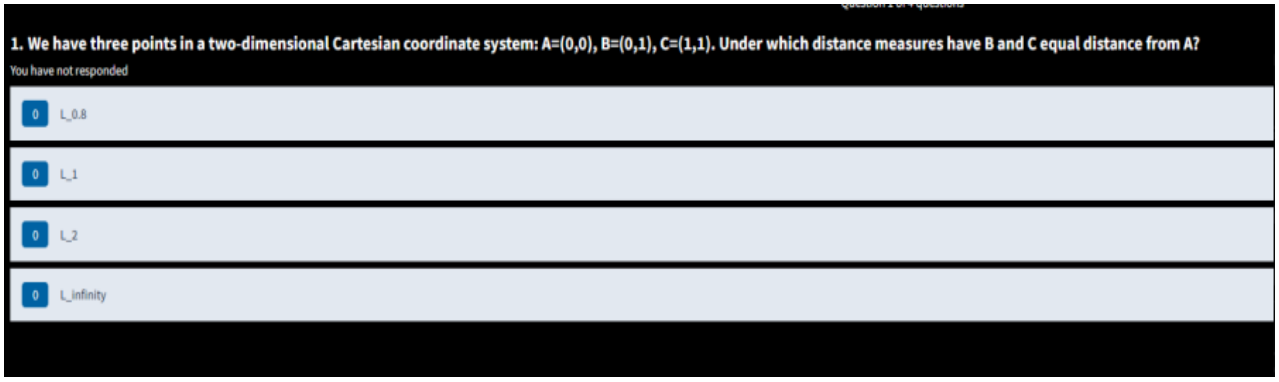
(BC) => (D) have NOT all 4 items of (A,B,C)=>(D) represented, therefore this statement is **WRONG.**

(BC) => (A) have NOT all 4 items of (A,B,C)=>(D) represented, therefore this statement is **WRONG.**

(C) => (ABD) have all 4 items of (A,B,C)=>(D) represented and 'D' is not on the left side of =>, therefore this statement is **TRUE.**

(CD) => (AB) have all 4 items of (A,B,C)=>(D) represented, BUT 'D' is on the LEFT side of => and this is against the rules, therefore this statement is **WRONG.**

Spørgsmål: Distance Related Questions



Correct answer: L_infinity

Correct Answers: 4

Direkte svar:

Please read the question carefully. We need to measure the distance from B to A and C to A, and we choose the method that has equal distance. Therefore, we apply the code in Python or R to calculate the three different distance-methods:

- L_1 norm = Manhattan
- L_2 norm = Euclidean
- L_Infinity = Maximum

R kode:

```
# Definere de 2 distancer som skal måles ved p og q:

p <- c(0,0)

q <- c(1,1)
```

```
#Euclidean (L_2 norm)

dist2 <- sqrt(sum((p-q)^2))

dist2

#Manhattan (L_1 norm)

dist1 <- sum(abs(p-q))

dist1

#Chebyshev or maximum, infinity

dist_max<- max(abs(p-q))

dist_max
```

Hver af disse metoder printer distancerne ud, her skal vi teste på både B's distance til A og C's distance til A. Ved at køre koderne kan vi se at L_infinity viser samme distancer for B og C til A.

Exercise 4-1 **Color-histograms and distance functions (1 point)**

As a warm-up on distance measures: For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean, quadratic form)

$$
\begin{aligned}
\mathrm{dist}_2(p,q) &= \left(|p_1-q_1|^2+|p_2-q_2|^2+|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\mathrm{dist}_1(p,q) &= |p_1-q_1|+|p_2-q_2|+|p_3-q_3| \\
\mathrm{dist}_\infty(p,q) &= \max(|p_1-q_1|,|p_2-q_2|,|p_3-q_3|) \\
\mathrm{dist}_w(p,q) &= \left(w_1|p_1-q_1|^2+w_2|p_2-q_2|^2+w_3|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\mathrm{dist}_M(p,q) &= \left((p-q)M(p-q)^{\mathrm{T}}\right)^{\frac{1}{2}}
\end{aligned}
$$

2. We have three points in a two-dimensional Cartesian coordinate system: A=(0,0), B=(0,2), C=(1,1). Under which distance measures have B and C equal distance from A?

You have not responded

- 0 L_0.8
- 0 L_1
- 0 L_2
- 0 L_infinity

Correct answer: L_1 (manhatten)

Direkte svar:

Please read the question carefully. We need to measure the distance from B to A and C to A, and we choose the method that has equal distance. Therefore, we apply the code in Python or R to calculate the three different distance-methods:

- L_1 norm = Manhattan
- L_2 norm = Euclidean
- L_Infinity = Maximum

R kode:

Definere de 2 distancer som skal måles ved p og q:

p <- c(0,0)

q <- c(1,1)

#Euclidean (L_2 norm)

dist2 <- sqrt(sum((p-q)^2))

dist2

#Manhattan (L_1 norm)

dist1 <- sum(abs(p-q))

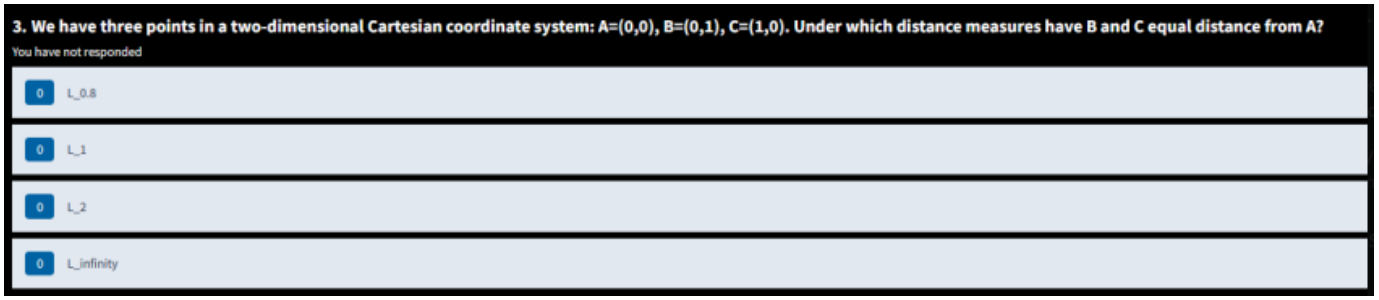
dist1

#Chebyshev or maximum, infinity

dist_max<- max(abs(p-q))

dist_max

Hver af disse metoder printer distancerne ud, her skal vi teste på både B's distance til A og C's distance til A. Ved at køre koderne kan vi se at L_2 viser samme distancer for B og C til A.



Correct answer: L_1, L_2 and L_infinity

Direkte svar:

Please read the question carefully. We need to measure the distance from B to A and C to A, and we choose the method that has equal distance. Therefore, we apply the code in Python or R to calculate the three different distance-methods:

- L_1 norm = Manhattan
- L_2 norm = Euclidean
- L_Infinity = Maximum

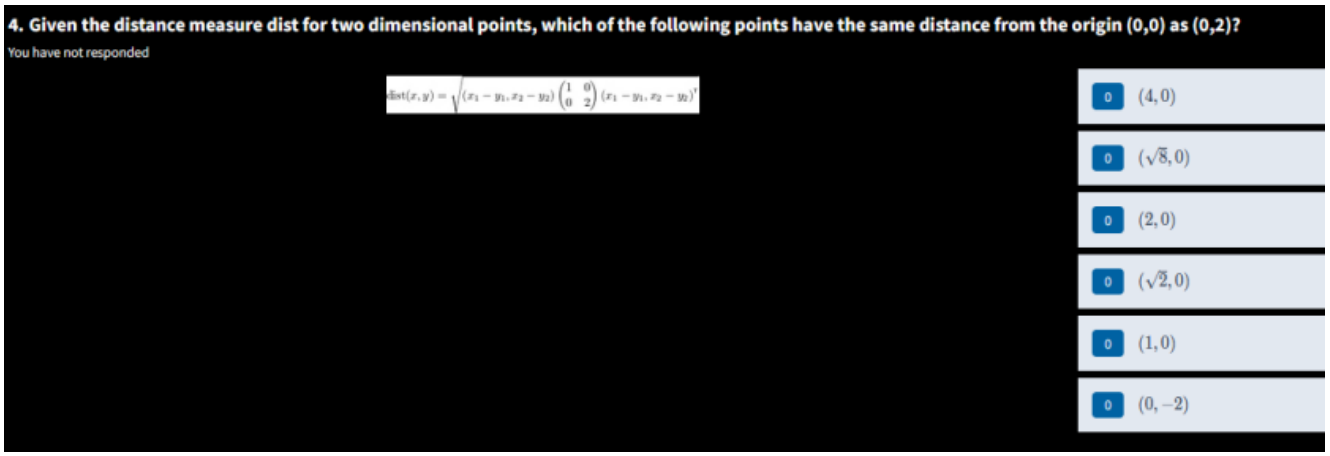
R kode:

Definere de 2 distancer som skal måles ved p og q:

p <- c(0,0)

```
q <- c(1,1)

#Euclidean (L_2 norm)

dist2 <- sqrt(sum((p-q)^2))

dist2

#Manhattan (L_1 norm)

dist1 <- sum(abs(p-q))

dist1

#Chebyshev or maximum, infinity

dist_max<- max(abs(p-q))

dist_max
```

Hver af disse metoder printer distancerne ud, her skal vi teste på både B's distance til A og C's distance til A. Ved at køre koderne kan vi se at L_infinity, L_2 og L_1 alle viser samme distancer for B og C til A.



Correct Answers: 2,6

Direkte svar:

Svar 0,-2 er korrekt da 0.2 til 0,0 har samme afstand som fra 0,-2 til 0,0.

Brug nedenstående kode:

Skift heletiden point 1 ud fra statements. Point 2 er origin distance som vi måler afstanden til.

Det kun punkt (sqrt(8), 0) og (0, -2) som har samme afstand til origin (0,0) som punkt (0,2)

Quadratic kode I R:

```
point1 <- c(0,2)

point2 <- c(0,0)

p <- matrix((point1),1,2, TRUE)
```

o <- matrix((point2),1,2, TRUE)

M

M <- matrix(c(2,0,0,4), 2,2,TRUE)

sqrt((p-o)%*%M%*%t(p-o))

Direkte svar:

P,o=(0,2), (0,0) = 4 (de andre afstande skal være 4 som denne)

(4,0), (0,0) = 5,6

(sqrt(8),0), (0,0)=4

(8,0), (0,0)=11,3

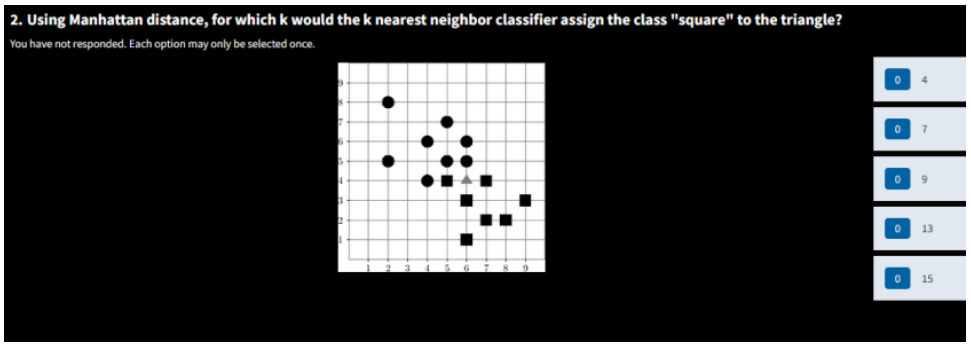
(sqrt(2),0) , (0,0)=2

(1,0) , (0,0)=1,41

(0,-2) , (0,0)**=4**

Som det ses af vores distances foroven, så har vi den originale p's afstand fra o er 4. de øvrige punkter hvor afstanden er det samme er markeret med fed gul foroven. Ved at anvende r koden fås kombinationerne (sqrt(8),0) og (0,-2), som de rigtige svar I denne opgave. Anvend formlen for quadratic distance fra exercise 4.

Spørgsmål: Distance + KNN + DBSCAN



Correct Answers: 1, 3 and 4

Direkte svar:

How to answer this question:

Start med at kigge på hvert statement og se om hvert k, classificere trekanten til classen firkant.

K = 4, her er de 4 tætteste på trekanten, firkanter (3 firkanter og 1 cirkel)

K = 7, her er de 7 tætteste på trekanten, cirkler (3 firkanter og 4 cirkler)

K = 9, her er de tætteste på trekanten, firkanter (5 firkanter og 4 cirkler)

K = 13, her er de tætteste på trekanten, firkanter (7 firkanter og 6 cirkler)

K = 15, her er de tætteste på trekanten, cirkler (7 firkanter og 8 cirkler)

1.

Start med at tælle alle observationer og dermed afgøre hvilken figur der fremtræder fleste gange i nærheden af "trekanten" (ubekendte feature). Herfra undersøger man de figurer, der fremtræder indenfor den givne afstand, og udvælger den figur, der optræder flest gange.

Firkanter afstande	Cirkler afstande
1	1
1	2
1	2
3	2
3	4
4	4
4	5
	8

Exercise 6-3:

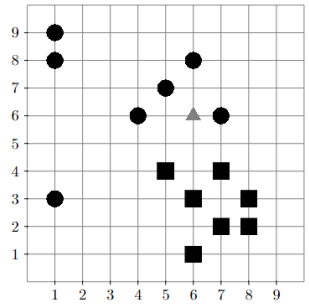
Exercise 6-3 Nearest neighbor classification (1 point)

The 2D feature vectors in the figure below belong to two different classes (circles and rectangles). Classify the object at $(6, 6)$ — in the image represented using a triangle — using k nearest neighbor classification. Use Manhattan distance (L_1 norm) as distance function, and use the non-weighted class counts in the k-nearest-neighbor set, i.e. the object is assigned to the majority class within the k nearest neighbors. Perform kNN classification for the following values of k and compare the results with your own "intuitive" result.

(a) $k = 4$

(b) $k = 7$

(c) $k = 10$



Ved brug af manhatten distance, som metode:

K = 4, her er de 4 tætteste på trekanten, cirkler (0 firkanter og 4 cirkel)

K = 7, her er de 7 tætteste på trekanten, cirkler(3 firkanter og 4 cirkel)

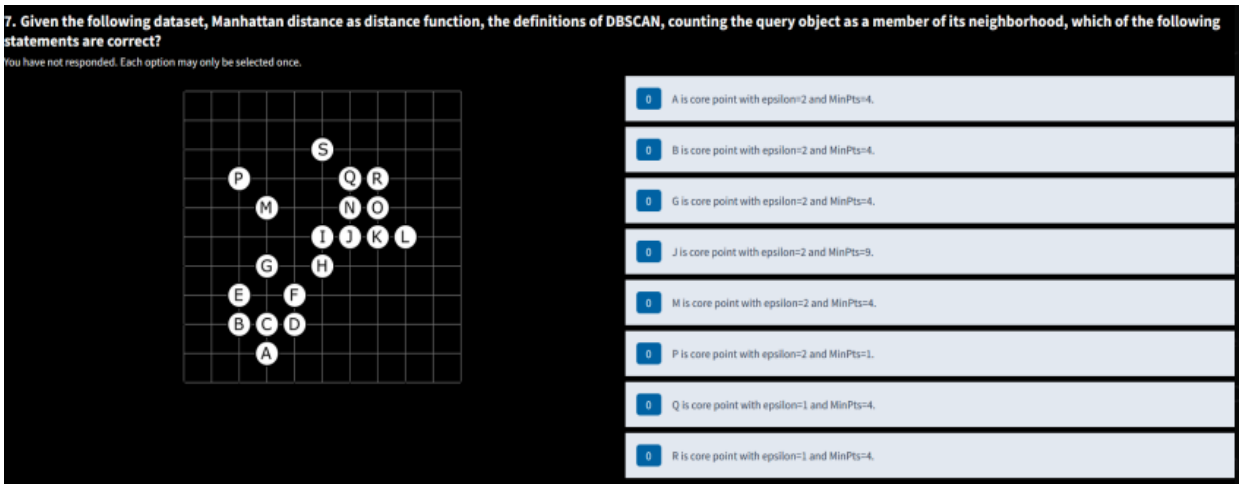
K = 10, her er de 10 tætteste på trekanten, firkanter (6 firkanter og 4 cirkel)

Altså kan vi se at når k bliver sat op til 10, vil man kunne klassificere trekanten til classen 'firkant'.

ANTAL

Cirkler: 1, 2, 2, 2, 7, 8, 8

Firkanter: 3,3,3,5,5,6,8



Correct Answers: 1, 2, 3 and 6

Direkte svar:

Definition:

Episilon : It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the ***k-distance graph***.

MinPts: Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. The minimum value of MinPts must be chosen at least 3

Tag en statement af gangen og tjek om den er korrekt

1. Det første statement: Hvis man antager A som sit core point med episilon=2 (epsilon er den maksimale afstand som er tilladt fra A altså A's radius.)
   Dvs. hvis epsilon er 2, så benyttes manhatten distance til at måle afstanden mellem core point og de tætteste punkter. (her må max afstanden være 2 eller mindre fra punkt A.)
2. MinPts er det minimale antal punkter som skal være I punkt A's nabolag.
   Fx hvis MinPoints = 4, så skal der være mindst 4 punkter fra punkt A (A inkl.) før det kan blive betragtet som et cluster, hvor A er inkluderet selv.

Konklusion: Epsilon = 2, MinPoints = 4 betyder derfor at der skal være mindst 4 punkter med en afstand af 2 eller mindre fra punkt A, før A kan blive betragtet som en Corepoint. (Her er A selv inkl. I MinPoints)

Answers:

A is a core point with epsilon = 2 & Minpts = 4 →TRUE, fordi B, C, D og A er inden for epsilon 2 fra A.

B is a core point with epsilon = 2 & Minpts = 4 →TRUE, fordi A, C, D, E og B er inden for epsilon 2 fra B.

G is a core point with epsilon = 2 & Minpts = 4 →TRUE, fordi H, F, C, E og G er inden for epsilon 2 fra G.

J is a core point with epsilon = 2 & Minpts = 9 →WRONG, fordi der kun er 8 punkter som er inden for epsilon 2 fra J, men Minpts var 9.

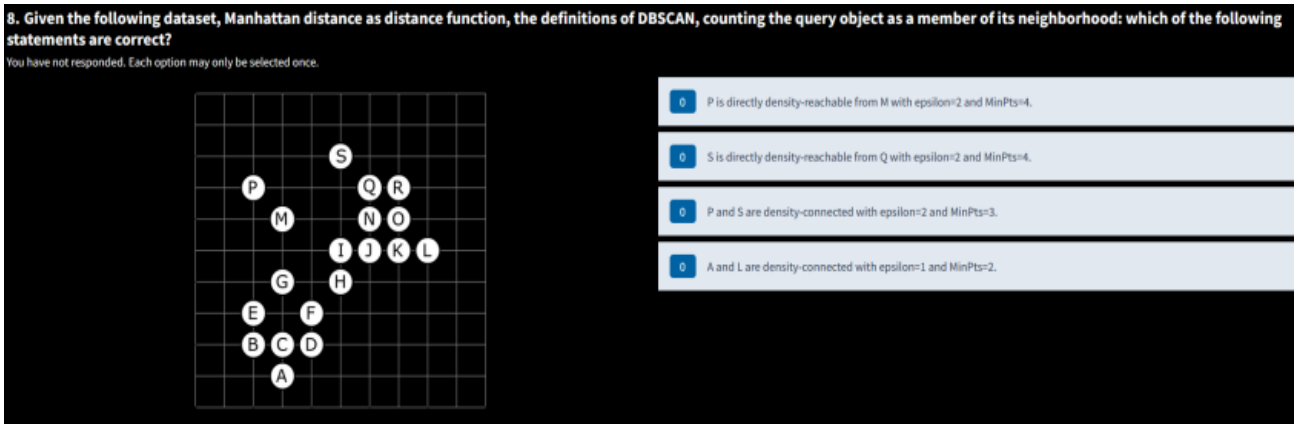
M is a core point with epsilon = 2 & Minpts = 4 →WRONG, fordi der kun er 3 punkter som er inden for epsilon 2 fra M, men Minpts var 4.

P is a core point with epsilon = 2 & Minpts = 1 →TRUE, fordi M og P er inden for epsilon 2 fra P.

Q is a core point with epsilon = 1 & Minpts = 4 →WRONG, fordi der kun er 3 punkter som er inden for epsilon 1 fra Q, men Minpts var 9.

P is a core point with epsilon = 2 & Minpts = 1 →TRUE, fordi og G er inden for epsilon 2 fra G.

R is a core point with epsilon = 1 & Minpts = 4 →WRONG, fordi der kun er 3 punkter som er inden for epsilon 1 fra R, men Minpts var 9.



Correct Answers: 2,3

Direkte svar:

Definition:

Epsilon : It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the ***k-distance graph***.

MinPts: Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. The minimum value of MinPts must be chosen at least 3.

Directly density reachable: An object (or instance) A is directly density reachable from object P if A is within the ε-Neighborhood of P and P is a core object.

Density connectivity: Object A is density-connected to object P w.r.t ε and *MinPts* if there is an object O such that both P and A are density-reachable from O w.r.t ε and *MinPts*.

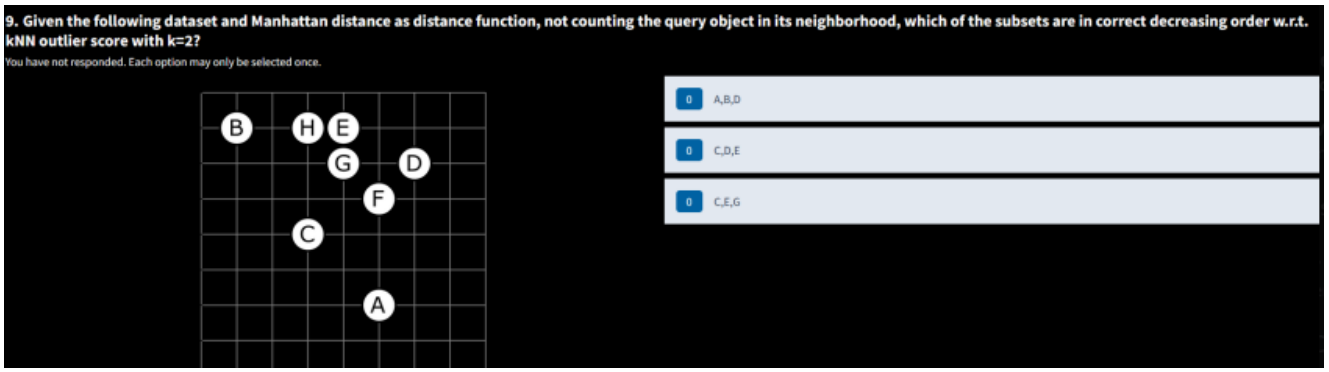
Right:

Answer 2:

S er density reachable for cluster Q, siden den ligger indenfor en radius på epsilon 2 fra Q, hvor der er en manhattan distance på 2 fra S til Q. Samtidig har Q minPts >= 4.

Answer 3:

P og S er density connected hvis, der er core-points der connecter de to punkter. Core-points skal opfylde følgende krav i denne case: epsilon=2 og MinPts>=3.



Correct Answers: 1, 2

Direkte svar:

K = 2

We have different subsets as statements and are told that these need to be in decreasingly ordered list, according to their KNN-outlier scores. (here there 2-nearest neighbors)

We start by taking every statement one by one and calculating their distance to their 2-nearest neighbors .We do this for every components in the statements and see if the distances is in decreasingly ordered list.

A,B,D = (TRUE)

A distance to 2-nearest neighbors, C and F = 4 and 3

B distance to 2-nearest neighbors, H and E = 2 and 3

D distance to 2-nearest neighbors, G and F = 2 and 2

Therefore the distance between the components 2-nearest neighbors is:

- A = 4 + 3 =7
- B = 2 +3 = 5
- D = 2+2 = 4

Therefore this statement is TRUE as the distances is in decreasingly ordered list (7,5,4)

C,D,E = (TRUE)

C distance to 2-nearest neighbors, H and G = 3 and 3

D distance to 2-nearest neighbors, G and F = 2 and 2

E distance to 2-nearest neighbors, H and G = 1 and 1

Therefore the distance between the components 2-nearest neighbors is:

- C = 3 + 3 = 6
- D = 2 + 2 = 4
- E = 1 + 1 = 2

Therefore this statement is TRUE as the distances is in decreasingly ordered list (6,4,2)

C,E,G = (WRONG)

C distance to 2-nearest neighbors, H and G = 3 and 3

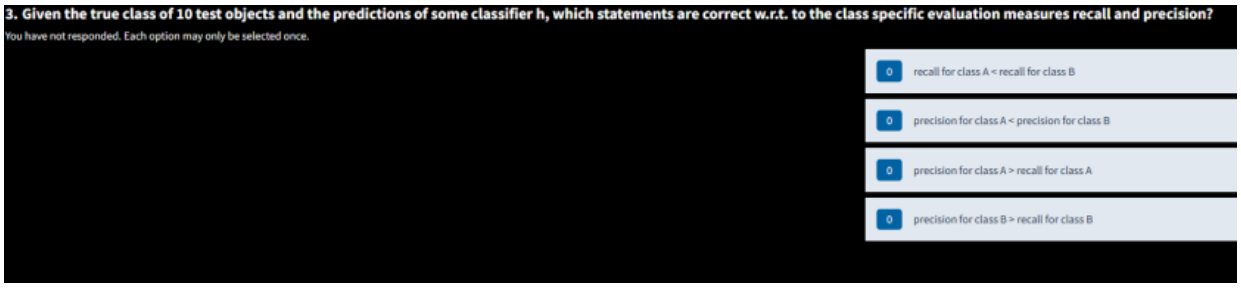
E distance to 2-nearest neighbors, H and G = 1 and 1

G distance to 2-nearest neighbors, E and D = 1 and 2

Therefore the distance between the components 2-nearest neighbors is:

- C = 3 + 3 = 6
- E = 1 + 1 = 2
- G = 1 + 2 = 3
- Therefore this statement is WRONG as the distances is NOT in decreasingly ordered list (6,2,3)

Spørgsmål: Classification



o	true class (f (o))	prediction (h(o))
o_1	A	A
o_2	B	A
o_3	B	B
o_4	B	B
o_5	A	B
o_6	A	B
o_7	A	A
o_8	A	A
o_9	A	A
o_{10}	B	B

Correct Answer: 1 and 3

Direkte svar:

A:

True Positive (prediction true, actual true - Both are A) = 4

True Negative (predition false, actual false - both that are not A) = 3

False Positive (prediction true, actual false - Prediction is A but true is B) = 1

False Negative (predicted false, actual true - Prediction B but true should be A) = 2

Precision = TP / (TP + FP) = 4 / (4 + 1) = 0.8

Recall = TP / (TP + FN) = 4 / (4 + 2) = 0.66

B:

True Positve (prediction true, actual true, Both are B) = 3

True Negative (predition false, actual false, Both are not B) = 4

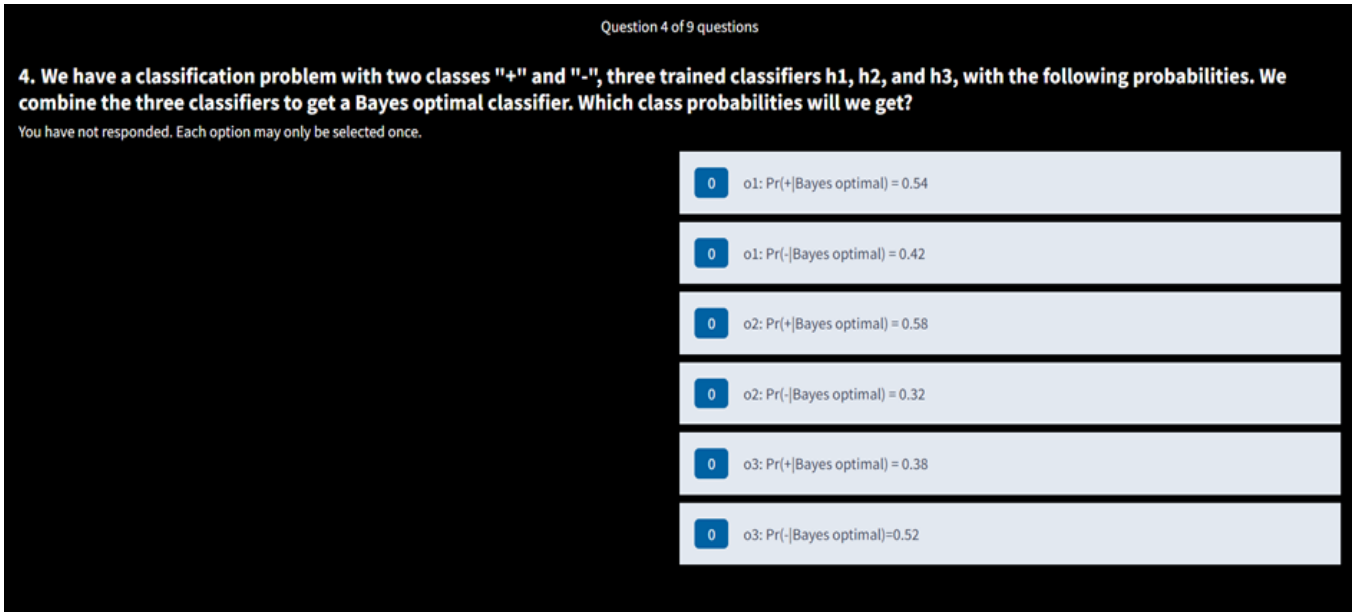
False Positive (prediction true, actual false - Prediction is B but true is A) = 2

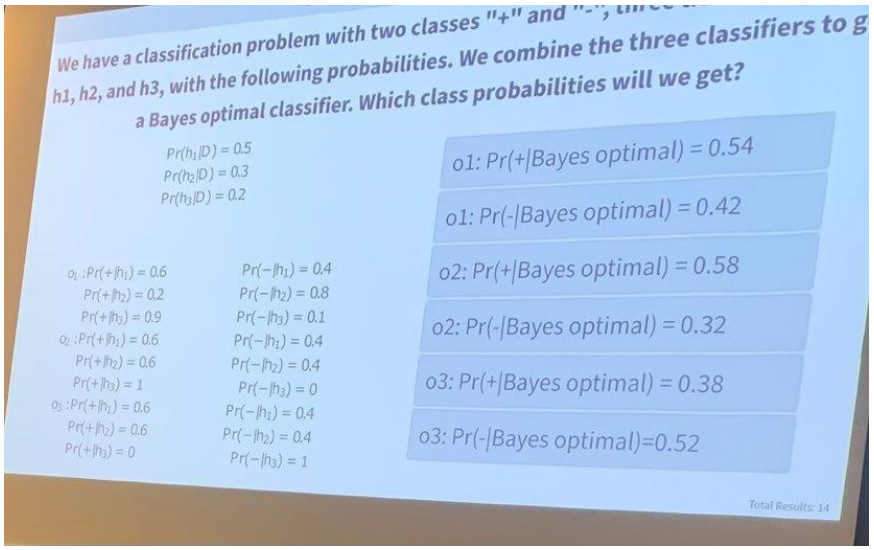
False Negative (predicted false, actual true - Prediction A but true should be B) = 1

Precision = TP / (TP + FP) = 3 / (3 + 2) = 0.6

Recall = TP / (TP + FN) = 3 / (3 + 1) = 0.75

Der er også en python kode:





Correct Answers: 1, 4 and 6

Direkte svar:

For at løse denne opgave, bedes du gange de angivet probabilities såsom 0.5, 0.3 og 0.2 (sandsynligheder) med de tilsvarende + (0.6, 0.2 og 0.9) probabilities og se om det samlet tal giver den angivet probabilities i svarmulighederne.

o1 +

0.5 x 0.6 = 0.3

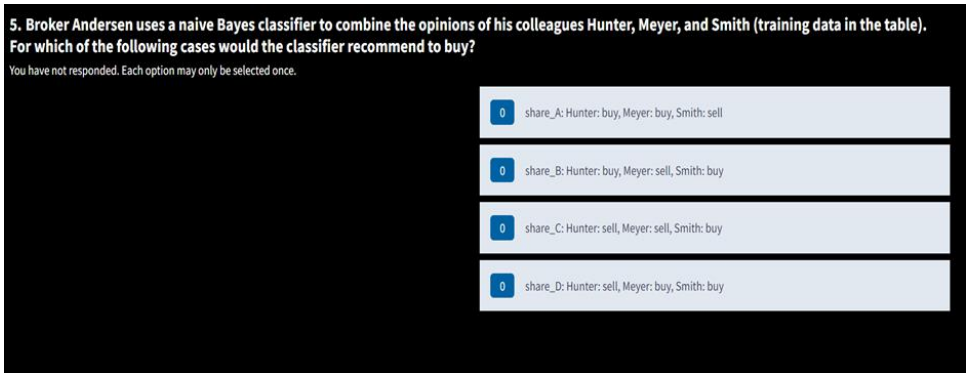
0.3 x 0.2 = 0.06

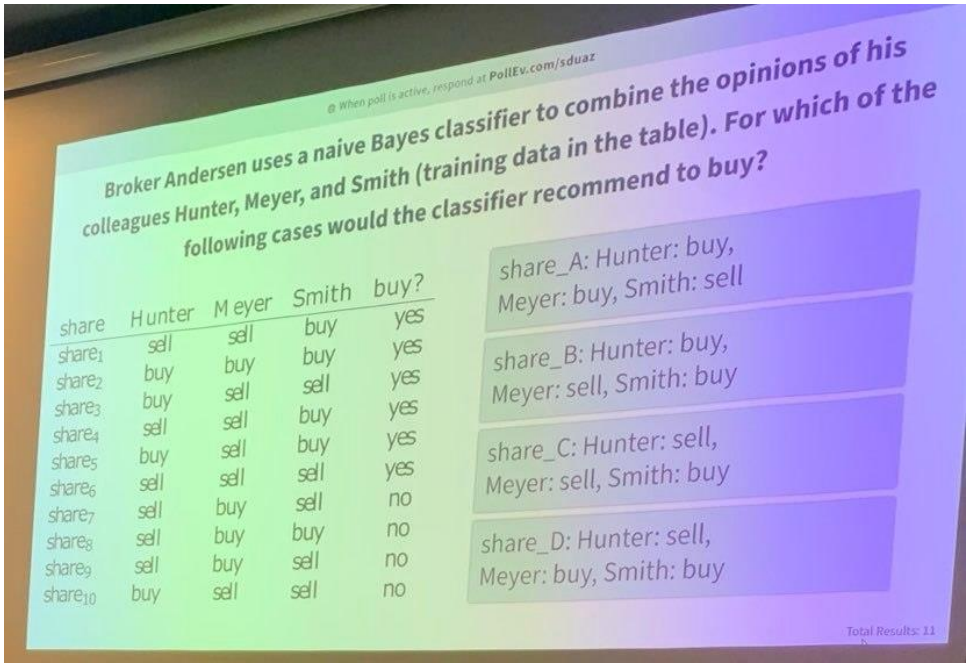
0.2 x 0.9 = 0.18

I alt = **0.54**

Der er en R-kode:

Spørgsmål: Probability





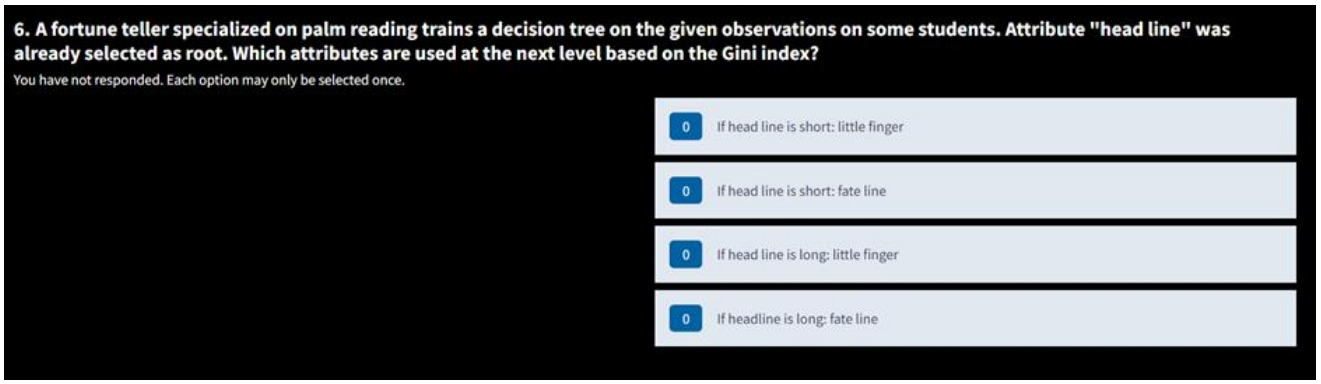
Correct Answers: 2, 3

Direkte svar:

Kig hvis den angivet kombination eksisterer i træningssættet, såfremt den eksisterer i træningsættet vil den anbefale at buy. Sørg for at tjekke hvis samme kombination forekommer flere gange med forskellige svar (derved udregn conditional probability)

Spørgsmål: Decision Tree med Gini Index

student	little f nger	head line	fate line	passed exam
1	straight	long	invisible	yes
2	bent	long	invisible	yes
3	straight	long	deep	no
4	straight	long	invisible	no
5	straight	long	invisible	yes
6	bent	long	invisible	yes
7	bent	short	deep	no
8	straight	short	deep	no
9	bent	short	invisible	no
10	straight	short	deep	yes



Correct Answers: 1, 4

Direkte svar:

Man skal finde gini index for hver gruppe i nævnte kolonner, dvs. for attribute ”little finger” finder vi for både ”straight” og ”bent” i forhold til grupperne ”long” og ”short” i attribute ”headline”, og for attribute ”fateline” finder vi for ”invisible” og ”deep” i forhold til ”long” og ”short” i attribute headline. Vi vælger den kombination med den laveste gini index.

Little finger, headline short

gini index for little finger bent, headline short, passed exam = 1 - ((0/2)^2 + (2/2)^2 = 0

- (0/2) = Antal gange “bent” I “little finger” = “yes” i ”exam passed” (Man kigger kun på de rækker som optræder i forhold til ”short” i ”headline” kolonnen) / antal gange ”bent” i ”little finger” optræder i forhold til ”short” i ”headline”
- (2/2) = antal gange “bent” I “little finger” = “no” i ”exam passed” (Man kigger fortsat kun på de rækker som optræder i forhold til ”short” i ”headline” kolonnen) / antal gange ”bent” i ”little finger” optræder i forhold ”short” i ”headline”

gini index for little finger straight, headline short, passed exam = 1 - ((1/2)^2 + (1/2)^2 = 0.5

- Samme som ovenstående, men her kigger man på ”straight” i ”little finger”.

gini index for little finger, headline short = (2/4) * 0 + (2/4) * (0.5) = 0.25

- (2/4) = antal gange “bent” I “little finger” optræder i forhold til ”short” i ”headline” / antal alle værdier i little finger (både ”bent” og ”straight”) som optræder i forhold til ”short” i ”headline”.
- (2/4) = antal gange “straight” I “little finger” optræder i forhold til ”short” i ”headline” / antal alle værdier i little finger (både ”bent” og ”straight”) som optræder i forhold til ”short” i ”headline”.

Fate line, headline short

gini index for fateline deep, headline short, passed exam = 1 - ((1/3)^2 + (2/3)^2 = 0.44

gini index for fateline invisible, headline short, passed exam = 1 - ((0/1)^2 + (1/1)^2 = 0

gini index for fateline, headline short , passed exam = (3/4) * (0.44) + (1/4) * 0 = 0.33

Little finger, headline long

gini index for little finger straight, headline long, passed exam = 1 - ((2/4)^2 + (2/4)^2) = 0.5

gini index for little finger bent, headline long, passed exam = 1 - ((2/2)^2 + (0/2)^2) = 0

gini index for little finger, headline long, passed exam = (4/6) * (0.5) + (2/6) * 0 = 0.33

Fate line, headline long

gini index for fateline invisible, headline long, passed exam = 1 - ((4/5)^2 + (1/5)^2 = 0.32

gini index for fateline deep, headline long, passed exam = 1 - ((0/1)^2 + (1/1)^2 = 0

gini index for fateline, headline long , passed exam = (5/6) * (0.32) + (1/6) * 0 = 0.2666

Svar 1 og 4 er korrekte da de har den laveste gini index.

Python:

```
"""
Created on Sun Jun 19 22:01:14 2022

@author: gamer
"""

def g(a,b):
    sum = (a+b)**2
    return 1 - (((a**2)/sum) + ((b**2)/sum))

def gini(a,b,c,d):
    #print(a+b)
    #print(a+b+c+d)
    #print(g(a,b))
    return (((a+b)/(a+b+c+d))*g(a,b)) + (((c+d)/(a+b+c+d)) * g(c,d))

print(gini(4,1,0,1))

#short, lil finger 0.25
#short, fate 0.333
#long, lil finger 0.333
#long, fate 0.266
```

Indsæt røde tal i gini() function

MCQ Alternative spørgsmål

Questions 1: APRIORI algorithm (Yousef)

Question 1)

Given the items *I =[A,B,C,D,E,F,G,H,I]*
And the set of transactions T:

Trans ID	Items
1	A B C E G H I
2	A B D E F H I
3	A B D E H
4	A B E F H
5	A B E H
6	A D F G I
7	A F I
8	B C D E G I
9	C G I
10	D E F G H I
11	D G I
12	F

QUESTION
For the minimum support of 3, we already detemined the frequent 3 –items.

Which of the following 4 itemssets are preliminary candidates in the next sequens

(i.e. after the merging step but before pruning

Answers (10)
ABDE
ABEH
BEHI
CDGI
CEGI
DEFI
DEGI
DEHI
EFHI
EGHI

Svar til opgaven

Materiale: Exercise 2 og 3 har lignende opgaver til at besvare disse typer opgave (APRIORI)

Direkte svar:

Give the items I = [A,B,C,D,E,F,G,H,I]
And the set of transactions T:

Trans ID	Items
1	A B C E G H I
2	A B D E F H I
3	A B D E H
4	A B D E F H
5	A B E H
6	A D F G I

7	A F I
8	B C D E G I
9	C G I
10	D E F G H I
11	D G I
12	F

For the minimum support of 3, we already determined the frequent 3-items. Which of the following 4 itemssets are preliminary candidates in the next sequens.

For the minimum support of 3, we already determined the frequent 3-itemsets with the APRIORI algorithm:

$$L^3 = \{ABE, ABH, AEH, AFI, BDE, BEH, BEI, CGI, DEH, DEI, DFI, DGI, EFH, EGI, EHI\}$$

Which of the following 4-itemsets are preliminary candidates in the next step of APRIORI (i.e., after the merging step but before pruning)?

1. ABDE
2. ABEH
3. BEHI
4. CDGI
5. CEGI
6. DEFI
7. DEGI
8. DEHI
9. EFHI
10. EGHI

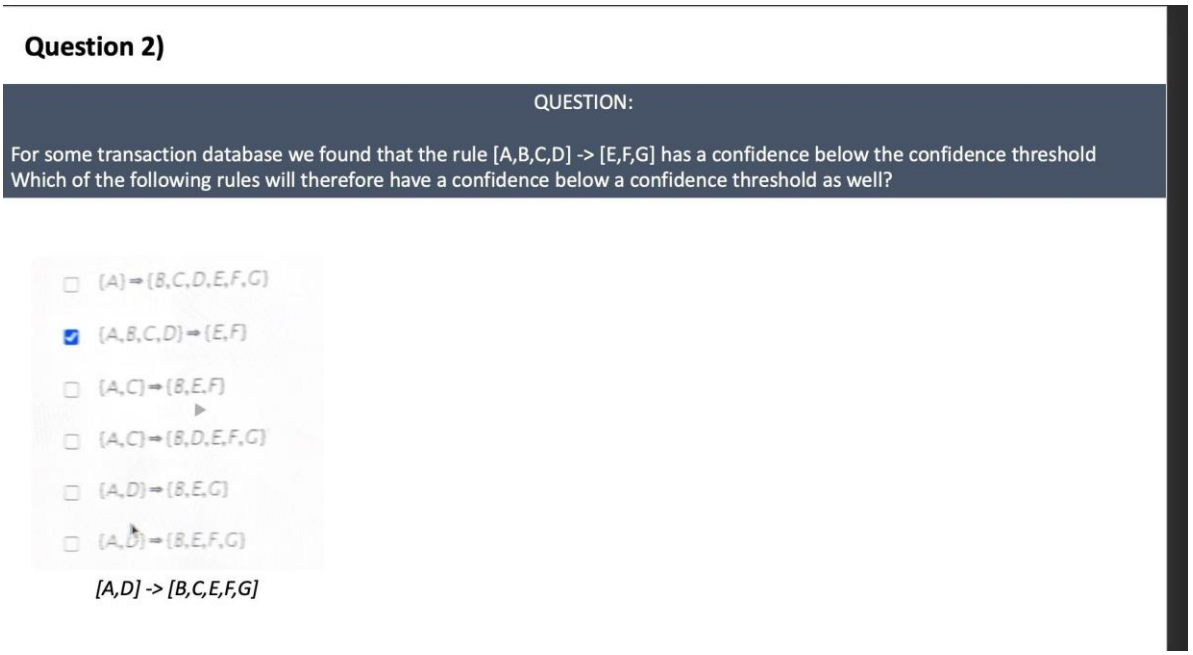
Solution:
Generate candidate set C_4 using L_3. Condition of joining two itemsets is that they should have $(k-2)$ elements in common, which is 2 in this case.

Candidate set C_4 would be

$$L_4 = \{ABEH, BEHI, DEHI\}$$

Thus itemsets 2, 3, 8 are preliminary candidates in step 4.

Question 2: Which of the following rules will therefore have a confidence below a confidence threshold as well (Yousef, skal laves sammen)



Svar til opgaven

Materiale: I forhold til threshold, så kig i følgende dokumenter:

- Machinelearning.docx
- Exercise 3
- Exercise 13

Direkte svar:

1 & 4

Question 3: One-dimensional dataset k-means

Question 3)

We have the following one-dimensional dataset:

ID	VALUE
A	1
B	3
C	5
D	7
E	10
F	11
G	12

Answers:

S1 is better that S2, in terms of TD^2

S2 is better that S3, in terms of TD^2

S3 is better that S1, in terms of TD^2

S1 and S3 are equally good in terms of TD^2

QUESTION

In three attempts, k-means delivered the following three clustering solutions.

S1 = [A,B,C,],[D,E,F,G]
S2 = [A,B],[C,D],[E,F,G]
S3 = [A,B,C,D],[E,F,G]

We want to compare the solutions using TD^2.

Which of the following statement are correct?

Svar til opgaven

I forhold til TD^2, så kig i følgende dokumenter:

- Exercise 5 omkring clustering: k-means and Silhouette
- Exercise 4 omkring k-means 1-dimensional Example

Direkte svar:

To solve this question we will use the TD^2 formula

The formula for calculating TD^2 $\approx SSQ(\mu_1, p_1) + SSQ(\mu_1, p_2) + \cdots + SSQ(\mu_1, p_n) = TD^2$

Start by calculating the mean of the different clusters: S1(A,B,C) $\approx \frac{1+3+5}{3} = 3$

The following code is a function made for both one-dimensional and two-dimensional datasets. To find the TD^2 in a one-dimensional dataset you plot your mean x-value in x and ordinary x-value in y.

In the two-dimensional function you plot your mean x-value in x and ordinary x-value in y. And plot your mean y-value in z and ordinary y value in d.

```
one_dim <- function(x,y)
  abs(x-y)^2

two_dim <- function(x,y,z,d)
  (abs(x-y)^2)+(abs(z-d)^2)
```

The following will show how s1 will look like:

```
x1 <- one_dim(3,1)
x2 <- one_dim(3,3)
x3 <- one_dim(3,5)



x4 <- one_dim(10,7)
x5 <- one_dim(10,10)
x6 <-one_dim(10,11)
x7 <- one_dim(10,12)
```

Then we need to sum all the values together to get the TD^2:

```
s1 <- sum(x1,x2,x3,x4,x5,x6,x7)
```

Answer for question 3:

TD^2 for S1 = 22

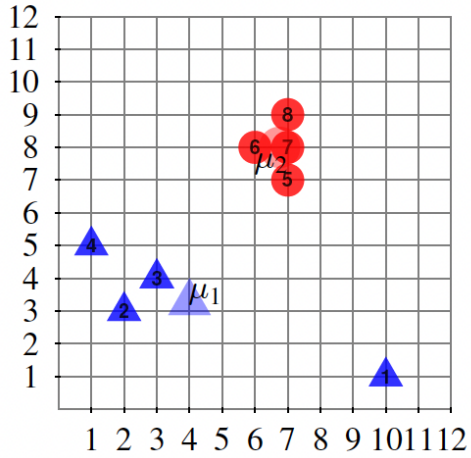
TD^2 for S2 = 6

TD^2 for S3 = 22

a. Then S1 is **NOT** better than S2, This is a false statement
b. S2 is better than S3, This is a true statement
c. S3 is **NOT** better than S1, This is a false statement
d. S1 and S3 is equally good, This is a true statement

Example from teachers slides:

k-means Clustering – Quality



First solution: $TD^2 = 61\frac{1}{2}$

$SSQ(\mu_1, p_1) = |4-10|^2 + |3.25-1|^2 = 36 + 5\frac{1}{16} = 41\frac{1}{16}$
$SSQ(\mu_1, p_2) = |4-2|^2 + |3.25-3|^2 = 4 + \frac{1}{16} = 4\frac{1}{16}$
$SSQ(\mu_1, p_3) = |4-3|^2 + |3.25-4|^2 = 1 + \frac{9}{16} = 1\frac{9}{16}$
$SSQ(\mu_1, p_4) = |4-1|^2 + |3.25-5|^2 = 9 + 3\frac{1}{16} = 12\frac{1}{16}$
$TD^2(C_1) = 58\frac{3}{4}$

$SSQ(\mu_2, p_5) = |6.75-7|^2 + |8-7|^2 = \frac{1}{16} + 1 = 1\frac{1}{16}$
$SSQ(\mu_2, p_6) = |6.75-6|^2 + |8-8|^2 = \frac{9}{16} + 0 = \frac{9}{16}$
$SSQ(\mu_2, p_7) = |6.75-7|^2 + |8-8|^2 = \frac{1}{16} + 0 = \frac{1}{16}$
$SSQ(\mu_2, p_8) = |6.75-7|^2 + |8-9|^2 = \frac{1}{16} + 1 = 1\frac{1}{16}$
$TD^2(C_2) = 2\frac{3}{4}$

Question 4: In the three attempts, k-means delivered the following three clustering solutions

Question 4)

QUESTION

In three attempts, k-means delivered the following three clustering solutions:

S1 = [A,B,C,],[D,E,F,G]
S2 = [A,B],[C,D],[E,F,G]
S3 = [A,B,C,D],[E,F,G]

We want to compare the solutions simplified Silhouette.
Which of the following statement are correct?

Answers options:

S1 is better that S2, in terms of simplified Silhouette

S2 is better that S3, in terms of simplified Silhouette

S3 is better that S1, in terms of simplified Silhouette

S1 and S3 are equally good in terms of in terms of simplified Silhouette

Svar til opgaven

Materiale: I forhold til three clustering solution, så kig i følgende dokumenter:

- Exercise 5 omkring clustering: k-means and Silhouette

Direkte svar:

To solve this question we will use the formula of the simplified silhouette:

Simplified Silhouette Coefficient [Rousseeuw, 1987]: Points

- let $a(o)$ be the distance between o and its "own" cluster representative
- let $b(o)$ be the distance between o and the closest "foreign" cluster representative
- the silhouette of o is given by

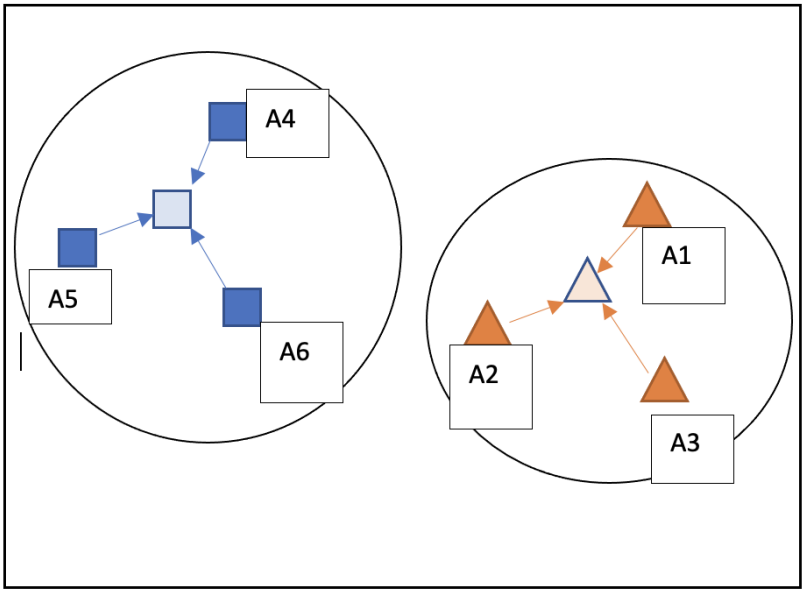
$$s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))}$$

- it holds that $-1 \leq s(o) \leq 1$
- $s(o) \approx -1, 0, 1$: bad, indifferent, good assignment of o

As we first need to define a and b, we will start by defining a.

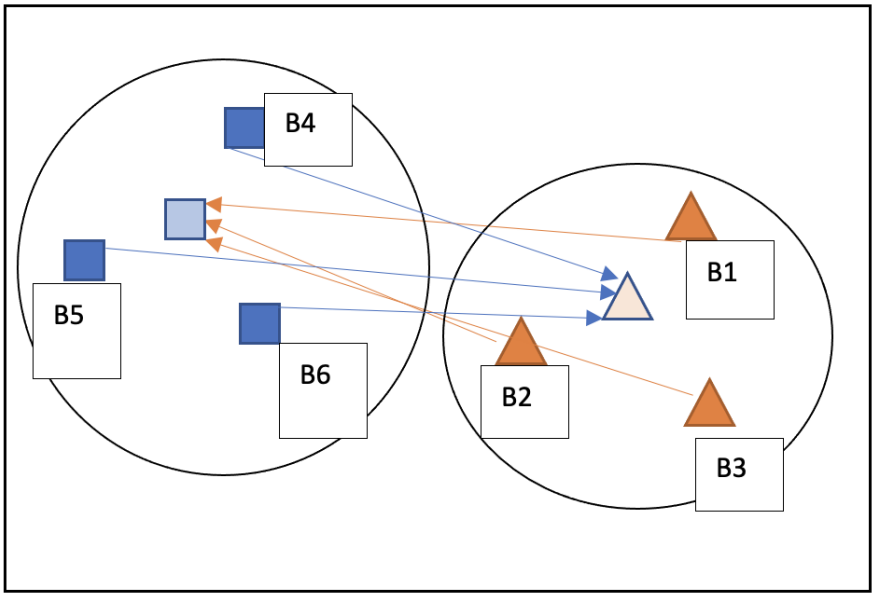
a is defined by the distance between every datapoint to the mean of its own cluster.

An example is provided in the following:



Then b is the distance between every datapoint to the closest “foreign” cluster mean. That means that we have to calculate the distance for every datapoint to the foreign clusters mean.

An illustration is created in the following:



Now that we have defined what a and b is, we will start by calculating the different a's and b's, and doing it for every datapoint. To calculate the distance between the datapoints we will use the following function:

```
one_dim <- function(x,y)
  sqrt(abs(x-y)^2)

two_dim <- function(x,y,z,d)
  sqrt((abs(x-y)^2)+(abs(z-d)^2))
```

We will use S1 as an example in our R code, so it will be more understandable:

So first we have to calculate our distance for a1, b1, a2, b2,…, a7, b7.

The R code is shown below:

**NOTE that we use the function from question 3 (Euclidian distance)

```
# S1

# 1. Cluster
a1 <- one_dim(3,1)
b1 <- one_dim(10,1)
a2 <- one_dim(3,3)
b2 <- one_dim(10,3)
a3 <- one_dim(3,5)
b3 <- one_dim(10,5)

# 2. Cluster
a4 <- one_dim(10,7)
b4 <- one_dim(3,7)
a5 <- one_dim(10,10)
b5 <- one_dim(3,10)
a6 <- one_dim(10,11)
b6 <- one_dim(3,11)
a7 <- one_dim(10,12)
b7 <- one_dim(3,12)
```

**NOTE HOW THE MEAN CHANGES FROM a4, AS WE HAVE A DIFFERENT MEAN IN THE OTHER CLUSTER. ALSO THAT THE MEAN VALUES CHANGED SO THAT WE HAVE THE FOREIGN CLUSTERS MEAN IN b.

Now that we have our a and b, we can now use the formular that we saw from the beginning, that is the following:

$$S(o) = \frac{b(o) - a(o)}{\max(b(o), a(o)}$$

We can use the following function, to find every silhouette score:

```
sil <- function(a,b)
  (b-a)/max(a,b)
```

As we insert every clusters a's and b's in the function, we will take the average of all our silhouette scores:

```
# Calculating the silhouette score for S1
sil1 <- sil(a1,b1)
sil2 <- sil(a2,b2)
sil3 <- sil(a3,b3)
sil4 <- sil(a4,b4)
sil5 <- sil(a5,b5)
sil6 <- sil(a6,b6)
sil7 <- sil(a7,b7)

# Finding the Silhouette score for S1
(sil1+sil2+sil3+sil4+sil5+sil6+sil7)/7

```

```
[1] 0.7543651
```

We have now calculated the simplified silhouette for S1. We can now calculate the simplified Silhouette with the same method and we will get the following answers:

$$Silhouette\ score\ S1 = 0.75$$

$$Silhouette\ score\ S2 = 0.73$$

$$Silhouette\ score\ S3 = 0.77$$

a. Then S1 is better than S2, This is a true statement
b. S2 is **NOT** better than S3, This is a false statement
c. S3 is better than S1, This is a true statement
d. S1 and S3 is **NOT** equally good, This is a false statement

(Question 5: EM-clustering: which of the following statements are true (Majid)

Question 5)

QUESTION
EM-CLUSTERING: Which of the following statements are true

Answers options:

EM- clustering makes use of Bayes rule?
EM- clustering assumes independence between attributes?
EM- clustering is a generalization of k-means?
The principles of EM clustering is to find the MAP hypothesis?

Svar til opgaven

I forhold til EM-clustering, så kig i følgende dokumenter:

- Exercise 9: EM-Clustering, Density Estimation, DBSCAN, Comparison of Clusterings
- Ellers kig i hans powerpoint for teoretisk forklaring af EM-clustering

Direkte svar:

- **EM-clustering makes use of Baye's rule?**
  True
- **EM-clustering assumes independence between attributes?**
  True (Variablerne skal være ens, men være uafhængige af hinanden)
- **EM-clustering is a generalization of k-means?**
  True
  - **The principles of EM-clustering is to find the MAP hypothesis**

  False

Question 6: Forskellige muligheder

Question 6)

QUESTION
Which of the following statements are correct

Answers options:

The number of parameters to describe a d-dimensional normal distribution grows quadratically with the number of dimensions?
In 10-fold cross validation, each object is used exactly ten times for testing
A non-parametric clustering algorithm is an algorithm that does not require any user-specified parameters.
A larger decision tree has a higher tendency to overfit than a smaller decision tree.
Neural Nets and Decision Trees search heuristically for separation boundaries and cannot guarantee to find the best solution

Svar til opgaven

Materiale: I forhold til three clustering solution, så kig i følgende dokumenter:

- Kig i slides eller teoretisk baggrund for de respektive områder.

Direkte svar:

Statement 1: Correct, hver gang vi tilføjer en variable bliver correlation matrix større (flere variabler man skal samligne, derfor bliver den kvadratisk større)

Statement 2: Wrong, I 10-fold tester vi 9 gange ud af de 10 folds.

Statement 3: Wrong, Oftes skal der specificeres nearest neighbors distance measure og nogle gange antalet af. Clusters(K-means).

Statement 4: Correct,

Statement 5 = Correct, boundaries kommer I et hierarki. (eksempel: I den første gren (step 1), der bliver lavet efter roden (step 0), kan der ifølge gini index beregning mene at mænd giver den laveste score, så kvinder bliver slet ikke taget med I første gren (step 1), men I grenen efterfølgende (step 2) kan det f.eks. handle om makeup, havde vi taget kvinder I step 1 istedet for mænd, havde gini index beregningen for step 2 været lavere. Med andre ord, vi kunne havde bygget en model, der var bedre, hvis vi så på step 2 istedet for step 1, men grundet hierarki ser vi kun fra step 0 -> step 1 -> step 2 osv.

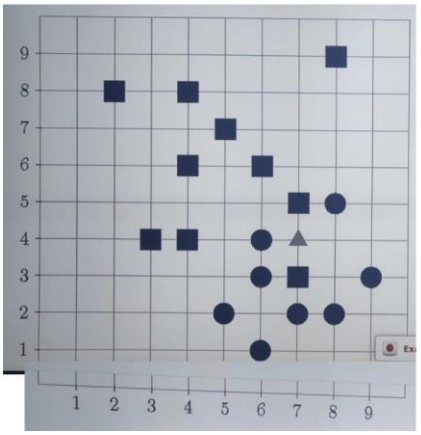
Question 7: two-dimensional dataset - using manhattan distance

Question 7)

QUESTION

We have the following two-dimensional dataset:

Using Manhattan distance, for which choices of k would the KNN classifier classify the query point (triangle) as square?



Answers options:

- *3*
- *6*
- *10*
- *13*
- *15*
- *17*
- *18*

Svar til opgaven

Materiale: I forhold til manhattan distance, så kig i følgende dokumenter:

- Exercise 6: Classification Evaluation, k-Nearest Neighbor Classification, Introduction to R. (Exercise 6-3 Nearest neighbor classification)

Alternativ:

- Exercise 10: Shared Nearest Neighbors, Hierarchical Clustering, Outlier Detection
- Exercise 9: EM-Clustering, Density Estimation, DBSCAN, Comparison of Clusterings

Direkte svar:

How to answer this question:

Start med at kigge på hvert statement og se om hvert k, klassificere trekanten til klassen firkant.

K = 3, her er de 3 tætteste på trekanten, firkanter (2 firkanter og 1 cirkel)

K = 6, her er de 6 tætteste på trekanten, cirkler (2 firkanter og 4 cirkler)

K = 10, her er de tætteste på trekanten, cirkler (3 firkanter og 7 cirkler)

K = 13, her er de tætteste på trekanten, cirkler (5 firkanter og 8 cirkler)

K = 15, her er de tætteste på trekanten, cirkler (7 firkanter og 8 cirkler)

K = 17, her er de tætteste på trekanten, firkanter (9 firkanter og 8 cirkler)

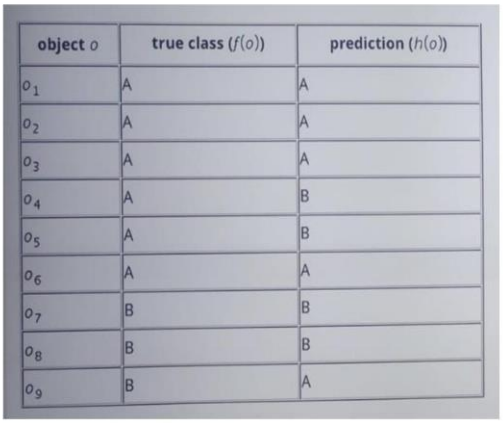
K = 18, her er de tætteste på trekanten, firkanter (10 firkanter og 8 cirkler)

Question 8: Evaluation measures recall and precision

QUESTION

Question 8)

Given the true class of 10 test objects and the predictions of some classifier h, which statement are correct w.r.t. to the class specific evaluation measures recall and precision?



object o	true class ($f(o)$)	prediction ($h(o)$)
o_1	A	A
o_2	A	A
o_3	A	A
o_4	A	B
o_5	A	B
o_6	A	A
o_7	B	B
o_8	B	B
o_9	B	A

Answers options:

- *Recall for class A > recall for class B*
- *Precision for class A > precision for Class B*
- *Precisions for class A > recall for class A*
- *Precision for class B > recall for class B*

Svar til opgaven

Materiale: I forhold til evaluation measures, så kig i følgende dokumenter:

- Exercise 6-1 Measure for Evaluation of Classifiers

Direkte svar:

Løsning med Python

Nederst I koden indsætter man alle rækker således:

```
# Question 8 - MCQ
classifier(["a","b"],[["a","a"],["a","a"],["a","a"],["a","b"],["a","b"],["a","a"],["b","b"],["b","b"],["b","a"]])
```

classifier(["a","b"],[["a","a"],["a","a"],["a","a"],["a","b"],["a","b"],["a","a"],["b","b"],["b","b"],["b","a"]])

I den første parantes (den røde), skriver man de grupper der er findes i tabellen, dvs ”a” og ”b”.
Derefter indtastes alle rækker i tabellen således: [blå, grøn]
Blå = venstre kolonne (true class)
Grøn = højre kolonne (prediction)

Output:

```
Precision: {'a': 0.8, 'b': 0.5}
Recall: {'a': 0.6666666666666666, 'b': 0.6666666666666666}
(2 * 0.6666666666666666 * 0.8) / (0.6666666666666666 + 0.8)
F1 = 0.7272727272727272, a
(2 * 0.6666666666666666 * 0.5) / (0.6666666666666666 + 0.5)
F1 = 0.5714285714285715, b
```

Precision for A = 0.8
Precision for B = 0.5
Recall for A = 0.66
Recall for B = 0.66
F1-score for A = 0.72
F1-score for B = 0.57

Svar:

Recall for class A > recall for class B

- False

Precision for class A > precision for class B

- True

Precision for class A > recall for class B

- True

Precision for class B > recall for class B

- False

Question 9: Classification problem with bayes optimal

QUESTION

Question 9)

We have classification problem with two classes "+" and "-", three trained classifier h_1, h_2, and h_3, with the following probabilities of the classifiers, given the training data D:

$Pr(h_1|D) = 0.4$
$Pr(h_2|D) = 0.1$
$Pr(h_3|D) = 0.5$

For the tree test instances o_1, o_2, o_3, the classifier give the following class probabilities

o_1:
$Pr(+|h_1) = 0.7, Pr(-|h_1) = 0.3$
$Pr(+|h_2) = 0.2, Pr(-|h_2) = 0.8$
$Pr(+|h_3) = 0.5, Pr(-|h_3) = 0.5$

o_2:
$Pr(+|h_1) = 0.8, Pr(-|h_1) = 0.2$
$Pr(+|h_2) = 0.6, Pr(-|h_2) = 0.4$
$Pr(+|h_3) = 0.5, Pr(-|h_3) = 0.5$

o_3:
$Pr(+|h_1) = 0.6, Pr(-|h_1) = 0.4$
$Pr(+|h_2) = 0.6, Pr(-|h_2) = 0.4$
$Pr(+|h_3) = 0.5, Pr(-|h_3) = 0.5$

Answers options:

- o_1: Pr(+ | **Bayes optimal = 0.55**
- o_1: Pr(- | **Bayes optimal = 0.5**
- o_2: Pr(+ | **Bayes optimal = 0.53**
- o_2: Pr(- | **Bayes optimal = 0.37**
- o_3: Pr(+ | **Bayes optimal = 0.65**
- o_3: Pr(- | **Bayes optimal = 0.45**

Svar til opgaven

I forhold til classification problem with two classes, så kig i følgende dokument:

- Exercise 8-1 Bayes Optimal

Direkte svar:

I denne opgave man anvende de 3 ovenstående probabilities med deres tilsvarende udfald for at udregne Bayes optimal. Eksempel: gøres ved at man tager Probability af h1 og ganger den med h1 for + og -. Så for at udregne Probability for + i første instance (o1), skal man man gange udfaldet med probability med dens tilsvarende classifier (h1, h2 eller h3) – Altså:

Pr(h1 |D) = 0.4

O1+:

Man tager Pr(h1 | D) * Pr(+ | h1).

Derefter tager man summen af de udregnede tal for at få Bayes optimal:

Eksempel:

O1: Pr(+ | Bayes optimal) kan udregnes således:

O1+:

$$0.4 * 0.7 = 0.28$$

$$0.1 * 0.2 = 0.02$$

$$0.5 * 0.5 = 0.25$$

Bayes optimal: 0.28 + 0.02 + 0.25 = 0.55

Altså er O1: Pr(+ | Bayes optimal) = 0.55

Når man skal udregne det for o1-, så bruger man værdierne for Pr(- | h1, h2, eller h3), og ganger værdien med den tilsvarende training data værdierne:

O1-:

$$0.4 * 0.3 = 0.12$$

$$0.1 * 0.8 = 0.08$$

$$0.5 * 0.5 = 0.25$$

Bayes optimal: 0.12 + 0.08 + 0.25 = 0.45.

Altså er Altså er O1: Pr(- | Bayes optimal) = 0.45.

Samme metode anvendes for resterende, beregninger fra ovenstående spørgsmål gives forneden:

O2+:

$$0.4 * 0.8 = 0.32$$

$$0.1 * +0.6 = 0.06$$

$$0.5 * 0.5 = 0.25$$

$$Bayes: 0.32 + 0.06 + 0.25 = 0.63$$

O2-:

$$0.4 * 0.2 = 0.08$$

$$0.1 * 0.4 = 0.04$$

$$0.5 * 0.5 = 0.25$$

$$Bayes: 0.37$$

O3+:

$$0.4 * 0.6 = 0.24$$

$$0.1 * 0.6 = 0.06$$

$$0.5 * 0.5 = 0.25$$

$$Bayes: 0.55$$

O3-:

$$0.4 * 0.4 = 0.16$$

$$0.1 * 0.4 = 0.04$$

$$0.5 * 0.5 = 0.25$$

$$Bayes: 0.45$$

Derfor er de korrekte svar: 1, 4 og 6.

Kode:

bayes optimal classifier

```{r}
# when h1 = 0.5, h2= 0.3 and h3 = 0.2
# checking for o1 +
x1 <- 0.5 * 0.6
x2 <- 0.3 * 0.2
x3 <- 0.2 * 0.9
sum(x1, x2, x3)

# checking for o1 -
x1 <- 0.5 * 0.4
x2 <- 0.3 * 0.8
x3 <- 0.2 * 0.1
sum(x1, x2, x3)

# checking for o2 +
x1 <- 0.5 * 0.6
x2 <- 0.3 * 0.6
x3 <- 0.2 * 1
sum(x1, x2, x3)

# checking for o2 -
x1 <- 0.5 * 0.4
x2 <- 0.3 * 0.4
x3 <- 0.2 * 0
sum(x1, x2, x3)

# checking for o3 +
x1 <- 0.5 * 0.6
x2 <- 0.3 * 0.6
x3 <- 0.2 * 0
sum(x1, x2, x3)

# checking for o3 -
x1 <- 0.5 * 0.4
x2 <- 0.3 * 0.4
x3 <- 0.2 * 1
```

Question 10: Naive Bayes

QUESTION

Question 10)

Navigating the society out of the lock-down, health inspector Mortensen shall give his opinion on allowing or forbidding several social activities. He takes orientation on the recommendations of medical, economical, and social science experts Olsen, Frandsen, and Jensen.

As these three are not always in agreement, Mortensen wants to use a naive Bayes classifier to combine their opinions. As training data, he uses the following observations from the past:

activity	Olsen	Frandsen	Jensen	allow?
$activity_1$	forbid	forbid	allow	yes
$activity_2$	forbid	forbid	forbid	yes
$activity_3$	forbid	forbid	allow	yes
$activity_4$	allow	forbid	forbid	yes
$activity_5$	allow	forbid	allow	yes
$activity_6$	allow	allow	allow	yes
$activity_7$	allow	allow	forbid	no
$activity_8$	forbid	allow	forbid	no
$activity_9$	forbid	allow	allow	no
$activity_{10}$	allow	forbid	forbid	no

Answers options

activity	Olsen	Frandsen	Jensen
$activity_A$	allow	allow	forbid

activity	Olsen	Frandsen	Jensen
$activity_B$	allow	forbid	forbid

activity	Olsen	Frandsen	Jensen
$activity_C$	forbid	forbid	allow

activity	Olsen	Frandsen	Jensen
$activity_D$	forbid	allow	allow

Svar til opgaven

Materiale: I forhold til naive bayes classifier, så kig i følgende dokument:

- Exercise 8-2 Naive Bayes

Direkte svar:

I denne opgave er den helt rigtige metode at træne en naive bayes classifier med træningssættet og køre denne på vores test set. Dog skal vi opstille sandsynligheder og bruge bayes classifier til at vurdere svarmuligheder (se exercise 8.2). Det kan nemt aflæsses, hvis der ikke er to ens scenarier, dog kan man også lave beregningen. Derfor har vi vores prior sandsynligheder fra træningssættet:

$$\Pr(yes) = \frac{6}{10}$$

$$\Pr(no) = \frac{4}{10}$$

$$\Pr(Olsen = Allow|Yes) = \frac{3}{5}$$

$$\Pr(Olsen = Allow|No) = \frac{2}{5}$$

$$\Pr(Olsen = Forbid|Yes) = \frac{3}{5}$$

$$\Pr(Olsen = Forbid|No) = \frac{2}{5}$$

$$\Pr(Frandsen = Allow|Yes) = \frac{4}{5}$$

$$\Pr(Frandsen = Allow|No) = \frac{1}{5}$$

$$\Pr(Frandsen = Forbid|Yes) = \frac{2}{5}$$

$$\Pr(Frandsen = Forbid|No) = \frac{3}{\mathbf{5}}$$

$$\Pr(Jensen = Allow|Yes) = \frac{1}{4}$$

$$\Pr(Jensen = Allow|No) = \frac{3}{4}$$

$$\Pr(Jensen = Forbid|Yes) = \frac{5}{6}$$

$$\Pr(Jensen = Forbid|No) = \frac{1}{6}$$

Derefter ganger vi dem sammen for vores aktiviteter, og får svaret:

Activity A=No det kan aflæses og også ses af beregningen forneden:

$\Pr(yes) * \Pr(Olsen = Allow|Yes) * \Pr(Frandsen = Allow|Yes) * \Pr(Jensen = Forbid|Yes) \Rightarrow$
$\frac{6}{10} * \frac{3}{5} * \frac{4}{5} * \frac{5}{6} = 0{,}24$
$\Pr(No) * \Pr(Olsen = Allow|No) * \Pr(Frandsen = Allow|No) * \Pr(Jensen = Forbid|No) \Rightarrow$
$\frac{4}{10} * \frac{2}{5} * \frac{1}{5} * \frac{1}{6} = 0{,}0053$

Conditional probability for "no":

$$\frac{0{,}24}{0{,}24 + 0{,}0053} \approx \underline{\underline{0.98}}$$

Activity B=No ikke så nemt at aflæse da der er to ens, men beregningen siger, at der er 88% sandsynlighed for det er No

$\Pr(yes) * \Pr(Olsen = Allow|Yes) * \Pr(Frandsen = Forbid|Yes) * \Pr(Jensen = Forbid|Yes) \Rightarrow$
$\frac{6}{10} * \frac{3}{5} * \frac{2}{5} * \frac{5}{6} = 0{,}12$
$\Pr(no) * \Pr(Olsen = Allow|no) * \Pr(Frandsen = Forbid|no) * \Pr(Jensen = Forbid|no) \Rightarrow$
$\frac{4}{10} * \frac{2}{5} * \frac{3}{5} * \frac{1}{6} = 0{,}016$

Conditional probability for "no":

$$\frac{0{,}12}{0{,}12 + 0{,}016} \approx \underline{\underline{0.88}}$$

Activity C= Yes eftersom sandsynligheden for no er 33%, må sandsynligheden for yes være 77%, derfor er aktivitet c = yes.

$$\Pr(yes) * \Pr(Olsen = Forbid|Yes) * \Pr(Frandsen = Forbid|Yes) * \Pr(Jensen = Allow|Yes) \Rightarrow$$
$$\frac{6}{10} * \frac{3}{5} * \frac{2}{5} * \frac{1}{4} = 0{,}036$$
$$\Pr(No) * \Pr(Olsen = Forbid|No) * \Pr(Frandsen = Forbid|No) * \Pr(Jensen = Allow|No) \Rightarrow$$
$$\frac{4}{10} * \frac{3}{5} * \frac{2}{5} * \frac{3}{4} = 0{,}072$$

Conditional probability for "no":

$$\frac{0{,}036}{0{,}036 + 0{,}072} \approx \underline{\underline{0{,}33}}$$

Activity D= No 75% sandsynlighed for no.

$$\Pr(yes) * \Pr(Olsen = Forbid|Yes) * \Pr(Frandsen = Allow|Yes) * \Pr(Jensen = Allow|Yes) \Rightarrow$$
$$\frac{6}{10} * \frac{3}{5} * \frac{4}{5} * \frac{1}{4} = 0{,}072$$

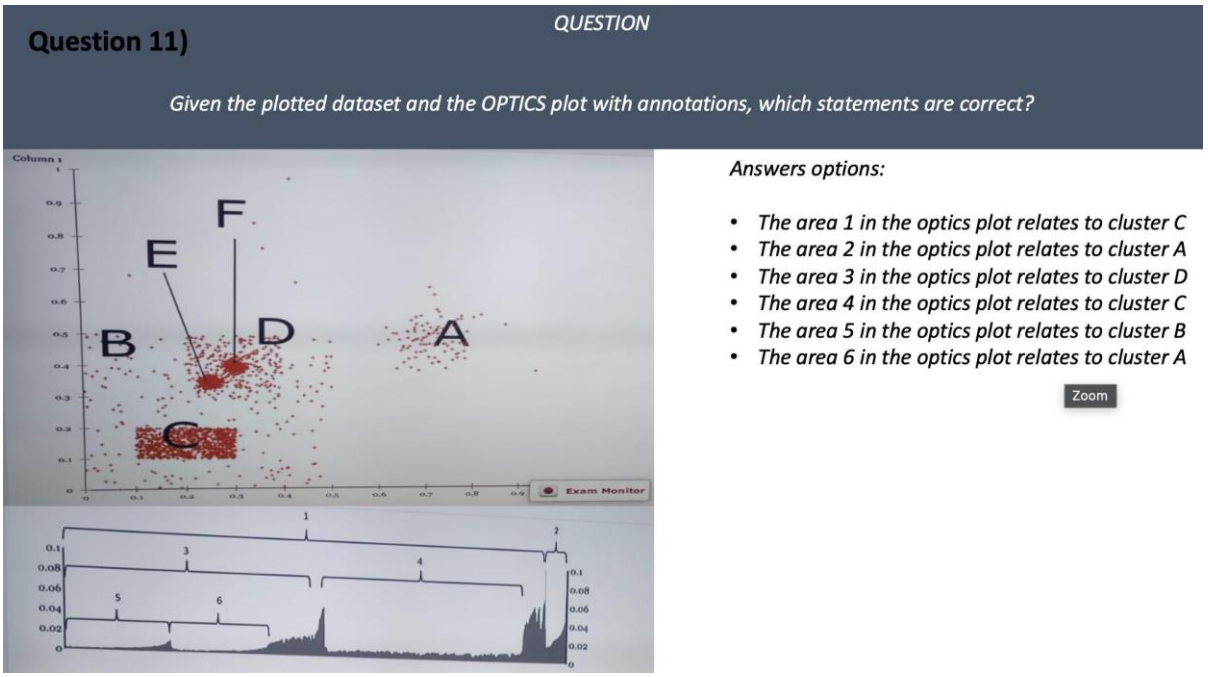
$$\Pr(No) * \Pr(Olsen = Forbid|No) * \Pr(Frandsen = Allow|No) * \Pr(Jensen = Allow|No) \Rightarrow$$
$$\frac{4}{10} * \frac{2}{5} * \frac{1}{5} * \frac{3}{4} = 0{,}024$$

Conditional probability for "no":

$$\frac{0{,}072}{0{,}072 + 0{,}024} \approx \underline{\underline{0{,}75}}$$

Svar: Acitivty C da conditional probability for "No" er 0.33% og derfor 0.67% for "Yes".

Question 11: OPTICS and plotted data



Svar til opgaven

Materiale: I forhold til OPTICS and plotted data, så kig i følgende dokument:

- Exercise 10-2 OPTICS Plot

Direkte svar:

Korrekte svar: 2, 3, 4

Skemaet forneden viser et diagram med en masse søjler deri. Disse søjler indikere afstanden fra punkt til punkt. Jo mindre disse søjler er, jo mere dense vil et cluster være. Dvs punkterne er meget tætte på hinanden. Ydermere ville store spikes indikere, at der er tale om et nyt cluster. Hvis et område indeholder flere spikes, betyder det, at der er flere clusters inde i et større cluster. Man kan tolke mellem to lignende clusters ved at se hvor dense de er, hvilket gøres ved at se på skemaet dernede: Det mest dense cluster har de korteste søjler.

I ovenstående opgave:

Area 1 afdækker det største område i skemaet, hvori man kan se flere areas nedenfor. Dette indikerer, at dette cluster har mange andre clusters i sig. På grafen kan man se at B er meget stort, og at der er 4 andre clusters under den, derfor er B = 1

Area 2 viser et cluster, der er adskilt fra de andre, og at dette cluster ikke er så dense, pga den store længde af søjler i dette areal. Det kan ses at A er adkilst fra de andre clusters og at punkterne ikke er tæt på hinanden, derfor er A = 2

Area 3 viser at der er 2 andre dense clusters inde i den. På grafen kan de ses at det gælder for D, hvor E og F er under den. Derfor er D = 3

Area 4 viser en dense cluster pga den korte søjlelængde. Derudover er der ikke nogen areas under den, dvs ingen cluster i den. Det kan ses på C at punkterne er tætte på hinanden og at der ikke er andre clusters inde i den, derfor er C = 4.

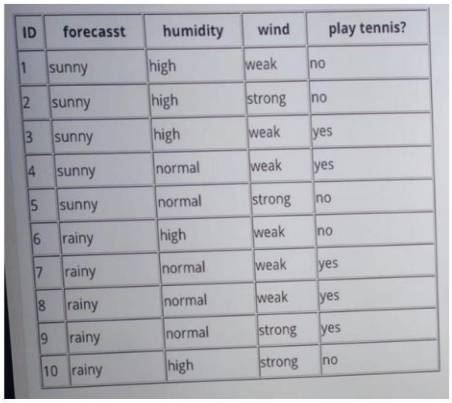
Area 5 og 6 ligner meget hinanden, de begge er meget dense pga den korte søjlelængde, og de begge kan ses under area 3. Kigger man på grafen, ser man at E og F er under D, og at de begge er meget dense. For at kunne skelne mellem 5 og 6, så kigger man på deres density. Det kan ses at E er mere dense end F, derfor må det være det area, som har den korteste søjlelængde være E. Det kan ses at area 6 har de korteste søjler, og derfor er den mest dense. Derfor er E = 6 og F = 5.

Question 12: Decision trees

QUESTION

Question 12)

A decision tree is being trained on the below data set. As root of the three, the attributes "forecast" was already selected Which attributes are selected as test nodes at the next level based on the Gini index?



ID	forecasst	humidity	wind	play tennis?
1	sunny	high	weak	no
2	sunny	high	strong	no
3	sunny	high	weak	yes
4	sunny	normal	weak	yes
5	sunny	normal	strong	no
6	rainy	high	weak	no
7	rainy	normal	weak	yes
8	rainy	normal	weak	yes
9	rainy	normal	strong	yes
10	rainy	high	strong	no

Answers options:

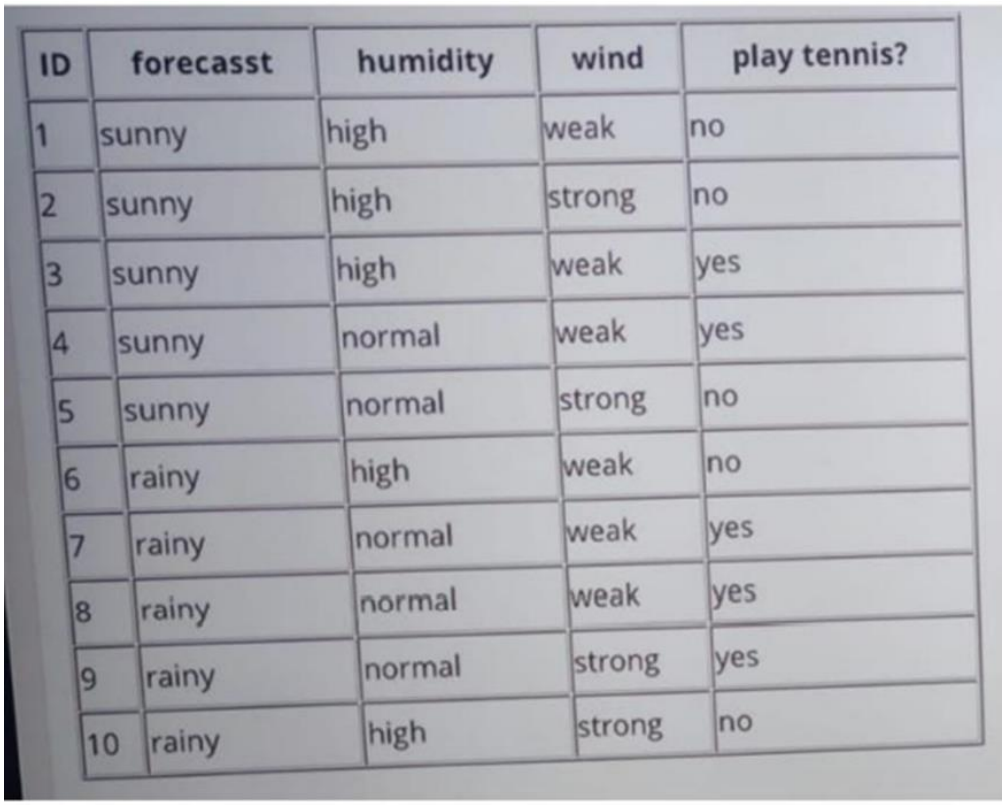
- *For the branch of forecast = sunny, we test wind*
- *For the branch of forecast = sunny, we test humidity*
- *For the branch of forecast = rainy, we test wind*
- *For the branch of forecast = rainy, we test humidity*

Svar til opgaven

I forhold til decision tree opgaver, så kig i følgende dokument:

- Exercise 12-1 Decision trees

Direkte svar:



ID	forecasst	humidity	wind	play tennis?
1	sunny	high	weak	no
2	sunny	high	strong	no
3	sunny	high	weak	yes
4	sunny	normal	weak	yes
5	sunny	normal	strong	no
6	rainy	high	weak	no
7	rainy	normal	weak	yes
8	rainy	normal	weak	yes
9	rainy	normal	strong	yes
10	rainy	high	strong	no

Svar 1,4

Wind, forecast sunny

Wind weak, forecast sunny, play tennis yes+no = 1 - ((2/3)^2 + (1/3)^2 = 0.44444

Wind strong, forecast sunny, play tennis yes+no = 1 - ((0/2)^2 + (2/2)^2 = 0

Forecast sunny, wind = (3/5) * (0.44) + (2/5) * 0 = 0.26

Humidity, forecast sunny

Humidity high, forecast sunny, play tennis yes+no = 1 - ((1/3)^2 + (2/3)^2 = 0.44444

Humidity normal, forecast sunny, play tennis yes+no = 1 - ((1/2)^2 + (1/2)^2 = 0.5

Forecast sunny, humidity high = (3/5) * (0.4444) + (2/5) * (0.5) = 0.4666

Wind, forecast rainy

Wind weak, forecast rainy, play tennis yes+no = 1 - ((2/3)^2 + (1/3)^2 = 0.44444

Wind strong, forecast rainy, play tennis yes+no = 1 - ((1/2)^2 + (1/2)^2 = 0.5

Forecast rainy, wind = (3/5) * (0.4444) + (2/5) *(0.5) = 0.46666

Humidity, forecast rainy

Humidity high, forecast rainy, play tennis yes+no = 1 - ((0/2)^2 + (2/2)^2 = 0

Humidity normal, forecast rainy, play tennis yes+no = 1 - ((3/3)^2 + (0/3)^2 = 0

= 0 (perfekt score)

Svarene 1 (wind, forecast sunny) og 4 (humidity, forecast rainy) har de laveste gini index og derfor de korrekte svar.

Question 13: Distance measure (dist)

Question 13)

QUESTION

Given the distance measure dist:

$$dist(x,y) = \sqrt{(x_1 - y_1, x_2 - y_2)\begin{pmatrix}2 & 0\\ 0 & 4\end{pmatrix}(x_1 - y_1, x_2 - y_2)^T}$$

For two-dimensional points and the point p=(0,2),
Which of the following points have the same distance as p from the origin (0,0)?

Answers options:

- *(4,0)*
- $(\sqrt{8}, 0)$
- *(8,0)*
- $(\sqrt{2}, 0)$
- $(\sqrt{2}, \sqrt{3})$
- *(1,0)*
- *(0,-2)*

Svar til opgaven

Materiale: I forhold til distance measure opgaver, så kig i følgende dokument:

- Exercise 4: Distance Measures, k-means

Quadratic kode I R:

```
point1 <- c(0,2)

point2 <- c(0,0)

p <- matrix((point1),1,2, TRUE)

o <- matrix((point2),1,2, TRUE)

M

M <- matrix(c(2,0,0,4), 2,2,TRUE)

sqrt((p-o)%*%M%*%t(p-o))
```

Direkte svar:

P,o=(0,2), (0,0) = 4

(4,0), (0,0) = 5,6

(sqrt(8),0), (0,0)=4

(8,0), (0,0)=11,3

(sqrt(2),0) , (0,0)=2

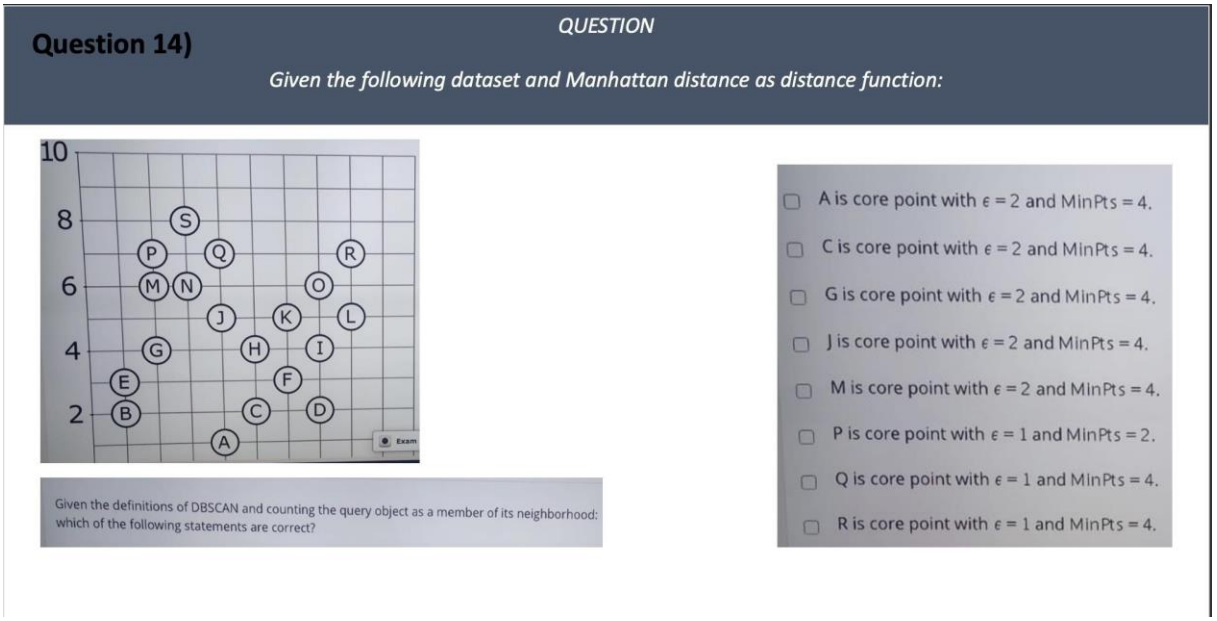
(sqrt(2),sqrt(3)) , (0,0)**=4**

(1,0) , (0,0)=1,41

(0,-2) , (0,0)**=4**

Som det ses af vores distances foroven, så har vi den originale p's afstand fra o er 4. de øvrige punkter hvor afstanden er det samme er markeret med fed foroven. Ved at anvende r koden fås kombinationerne (sqrt(8),0) og (0,-2) og (sqrt(2), sqrt(3)), som de rigtige svar I denne opgave. Anvend formlen for quadratic distance fra exercise 4.

Question 14: Manhattan distance as distance function



Svar til opgaven

Materiale: I forhold til manhattan distance opgaver, så kig i følgende dokument:

- Exercise 4: Distance Measures, k-means
- Exercise 6-3 Nearest neighbor classification (Der benyttes Manhattan distance som distance function!)

Direkte svar:

Definition:

Episilon : It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the ***k-distance graph***.

MinPts: Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. The minimum value of MinPts must be chosen at least 3.

Vi starter med at tage et statement ad gangen. Det første statement fortæller følgende:

A is a core point with **ε = 2** and minPts = 4 og at vores distance measure er **Manhattan**

Dvs. at vi kan bevæge os i en radius af **ε 2** og dermed opfylde minPts 4. Vi kan i ovenstående statement se at vi kun kan bevæge os til punkt c og dermed stopper kæden. Dvs. vi ender med minPts på 2 inklusive A som core point. Så dette statement er forkert.

Hvorimod vi tager det andet statement:

C is a core point with **ε = 2** and minPts = 4 og at vores distance measure er **Manhattan**

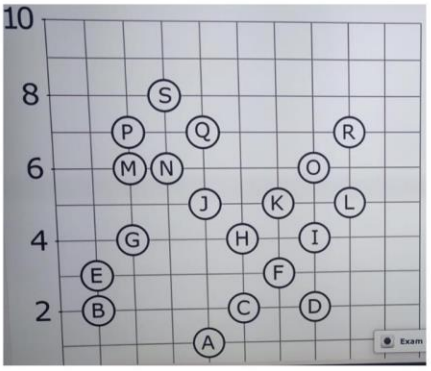
Der kan vi fra punkt C med **ε = 2** ramme både H, F og D. Som giver os en total minPts på 4 og indenfor vores **ε = 2.** Så dette statement er korrekt.

Question 15: Manhattan distance as distance function

Question 15)

QUESTION

Given the following dataset and Manhattan distance as distance function:



Given the definitions of DBSCAN and counting the query object as a member of its neighborhood: which of the following statements are correct?

Your answer:

- ☐ P is directly density-reachable from N with $\epsilon = 2$ and MinPts = 4.
- ☐ S is directly density-reachable from M with $\epsilon = 2$ and MinPts = 4.
- ☐ A and L are density-connected with $\epsilon = 2$ and MinPts = 2.
- ☐ B and R are density-connected with $\epsilon = 2$ and MinPts = 2.

Svar til opgaven

I forhold til manhattan distance opgaver, så kig i følgende dokument:

- Exercise 4: Distance Measures, k-means
- Exercise 6-3 Nearest neighbor classification (Der benyttes Manhattan distance som distance function!)

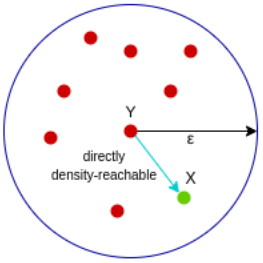
Direkte svar: 1, 3, 4 er rigtige

"*Counting the query object as a member of its neighborhood*" betyder at man tæller punktet som en del af dets eget neighborhood.
"*Not counting the query object as member of its neighborhood*" = man tæller ikke punktet som en del af dets eget neighborhood, fx hvis S har epsilon = 2 og MinPts = 3, så vil den ikke være et core point da kun P og Q er anset som points.

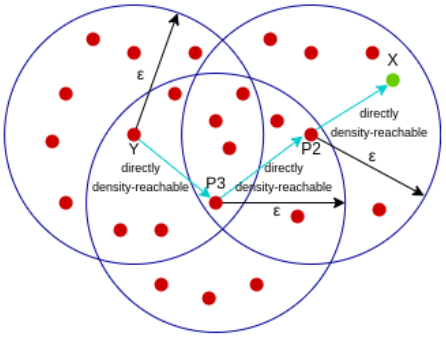
Corepoint = et punkt som opfylder både Epsilon og MinPts.

Eksempler visualiseret med punkter X og Y:

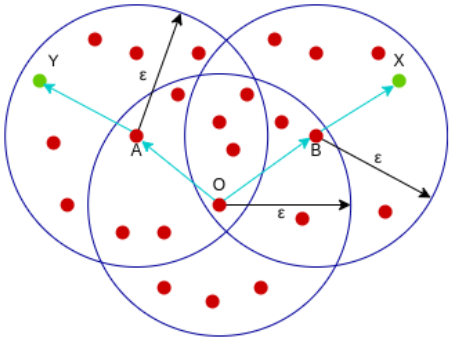
Directy density-reachable: Hvis et punkt er indenfor et andet core-points epsilon radius. X er inde i Y's nabolag:



Regel for **density-reachable**: To punkter er koblet sammen igennem en række af core-points som er directly density reachable til hinanden. X kan nås fra Y da P3 er inde i Y's nabolag, og P2 er inde i P3's nabolag, og X er inde i P2's nabolag. (Arthur fokusere ikke på denne)



Regel for **density-connected**: For at finde ud af om X og Y er density-connected, kræves det at der er et core-point mellem X og Y som er density-reachable til både X og Y. O er density-reachable til både X og Y.

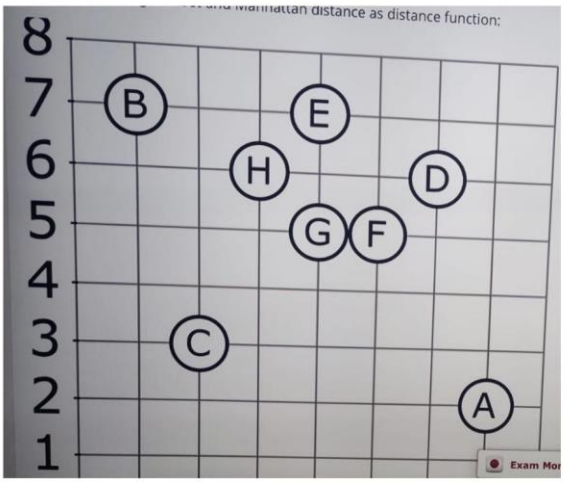


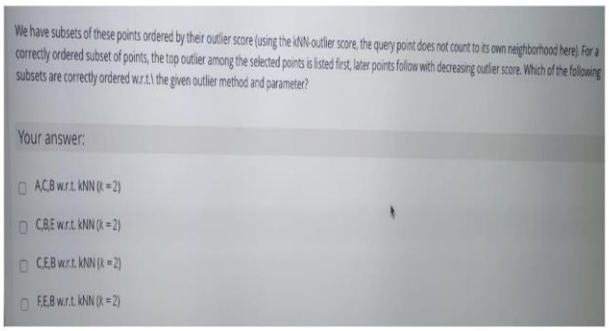
Question 16: Manhattan distance as distance function

Question 16)

QUESTION

Given the following dataset and Manhattan distance as distance function:





Svar til opgaven

Materiale: I forhold til manhattan distance opgaver, så kig i følgende dokument:

- Exercise 4: Distance Measures, k-means
- Exercise 6-3 Nearest neighbor classification (Der benyttes Manhattan distance som distance function!)

Direkte svar:

<u>Given outlier method = KNN and Manhatten as distance function</u>

We have different subsets as statements and are told that these need to be in decreasingly ordered list, according to their KNN-outlier scores. (here there 2-nearest neighbors)

We start by taking every statement one by one and calculating their distance to their 2-nearest neighbors .We do this for every components in the statements and see if the distances is in decreasingly ordered list.

A,C,B = (TRUE)

A distance to 2-nearest neighbors, D and F = 5 and 5

C distance to 2-nearest neighbors, H and G = 4 and 4

B distance to 2-nearest neighbors, E and H = 3 and 3

Therefore the distance between the components 2-nearest neighbors is:

- A = 5 + 5 = 10
- C = 4 +4 = 8
- B = 3+3 = 6

Therefore this statement is TRUE as the distances is in decreasingly ordered list (10,8,6)

C,B,E = (TRUE)

C distance to 2-nearest neighbors, H and G = 4 and 4

B distance to 2-nearest neighbors, E and H = 3 and 3

E distance to 2-nearest neighbors, H and G = 2 and 2

Therefore the distance between the components 2-nearest neighbors is:

- C = 4 + 4 = 8
- B = 3 + 3 = 6
- E = 2 + 2 = 4

Therefore this statement is TRUE as the distances is in decreasingly ordered list (8,6,4)

C,E,B = (WRONG)

C distance to 2-nearest neighbors, H and G = 4 and 4

E distance to 2-nearest neighbors, H and G = 2 and 2

B distance to 2-nearest neighbors, E and H = 3 and 3

Therefore the distance between the components 2-nearest neighbors is:

- C = 4 + 4 = 8
- E = 2 + 2 = 4
- B = 3 + 3 = 6

Therefore this statement is WRONG as the distances is NOT in decreasingly ordered list (8,4,6)

F,E,B = (WRONG)

F distance to 2-nearest neighbors, G and D = 1 and 2

E distance to 2-nearest neighbors, H and G = 2 and 2

B distance to 2-nearest neighbors, E and H = 3 and 3

Therefore the distance between the components 2-nearest neighbors is:

- F = 1 + 2 = 3
- E = 2 + 2 = 4
- B = 3 + 3 = 6

Therefore this statement is WRONG as the distances is NOT in decreasingly ordered list (3,4,6)

Question 17: ROC and AUC

Question 17) *QUESTION*

In a dataset with ten points (A,B,C,D,E,F,G,H,I,J), A and B are labeled outliers.

Four outlier detection methods, $m_1,\ldots,m_4$, deliver the following rankings (from left-to-right: top-rank to bottom-rank):

method	ranking
m_1	C,D,A,E,F,B,G,H,I,J
m_2	J,A,D,E,F,G,B,H,I,C
m_3	I,D,A,E,F,G,B,H,C,J
m_4	I,J,E,A,B,F,G,H,C,D

Based on ROC AUC as evaluation measure, which of the following statements is correct?

- ☐ m_1 and m_2 perform equally well.
- ☐ m_2 is better than m_3.
- ☐ m_3 is better than m_4.
- ☐ m_2 is better than m_4.

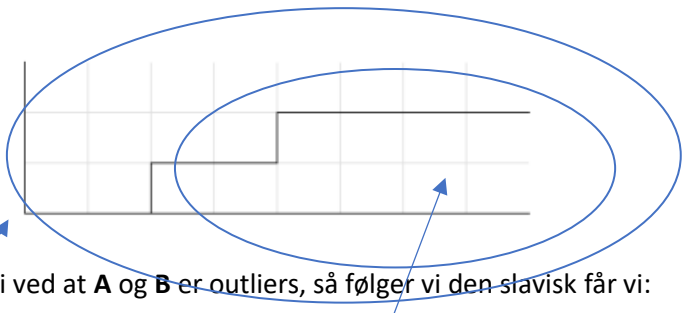
The correct statements are statements 1 and 2.

ROC curves: Recall for ROC curves: **(HOLD ØJE MED DENNE ER, DEN ER ANDERLEDES I EXERCISE 9-3)**

- each TP in the ranking: one step up
- each FP in the ranking: one step to the right
- comparison of two rankings: area under the curve (ROC AUC)

Vi tæller blot indtil vi når A og B, rammer vi f.eks. A går vi op, rammer vi derefter et vilkårligt andet bogstav (undtagen A og B), går vi til højre:

method	ranking
m_1	C,D,A,E,F,B,G,H,I,J



Se på eksemplet ovenover, vi ved at **A** og **B** er outliers, så følger vi den slavisk får vi:

1. $C\ er\ FP \rightarrow 1\ skridt\ til\ højre$
2. $D\ er\ FP \rightarrow 1\ skridt\ til\ højre$
3. $A\ er\ TP \rightarrow 1\ skridt\ op$
4. $E\ er\ FP \rightarrow 1\ skridt\ til\ højre$
5. $F\ er\ FP \rightarrow 1\ skridt\ til\ højre$
6. $B\ er\ TP \rightarrow 1\ skridt\ op$
7. $G\ er\ FP \rightarrow 1\ skridt\ til\ højre$
8. $H\ er\ FP \rightarrow 1\ skridt\ til\ højre$
9. $I\ er\ FP \rightarrow 1\ skridt\ til\ højre$
10. $J\ er\ FP \rightarrow 1\ skridt\ til\ højre$

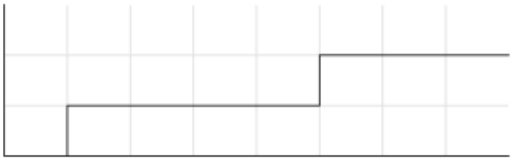
Herefter tæller vi mængden af firkanter til højre, hvor vi alt har **10**, og vi tæller det samlede antal af firkanter i hele figuren, hvilket er **24** (I dette tilfælde er det dog forkert, y-aksen angiver antallet af TRP

= $\frac{TP}{TP+FN} = \frac{2}{2+8} = 0{,}2$ dvs. 2 i Y-aksen, og x-aksen angiver $FPR = \frac{FP}{TN+FP} = \frac{8}{\mathbf{2+8}} = 0{,}8$, dvs. 8 hen af x-aksen, ganger vi $2 \cdot 8 = 16$ mulige firkanter, men her har de valgt $3 \cdot 8 = 24$ mulige firkanter, pointen bliver dog præcis det samme). vi kan nu beregne ROC AUC for $m1$

$$For\ m1: \frac{10}{24} \approx 0.41667$$

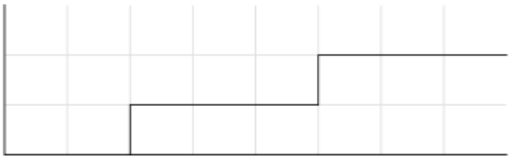
$$for\ m2: \frac{10}{24} \approx 0.41667:$$

m_2	J,A,D,E,F,G,B,H,I,C



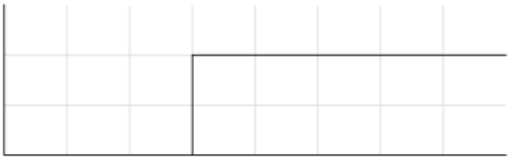
$$for\ m3: \frac{9}{24} \approx 0.375:$$

m_3	I,D,A,E,F,G,B,H,C,J



$$for\ m4: \frac{10}{24} \approx 0.41667$$

m_4	I,J,E,A,B,F,G,H,C,D



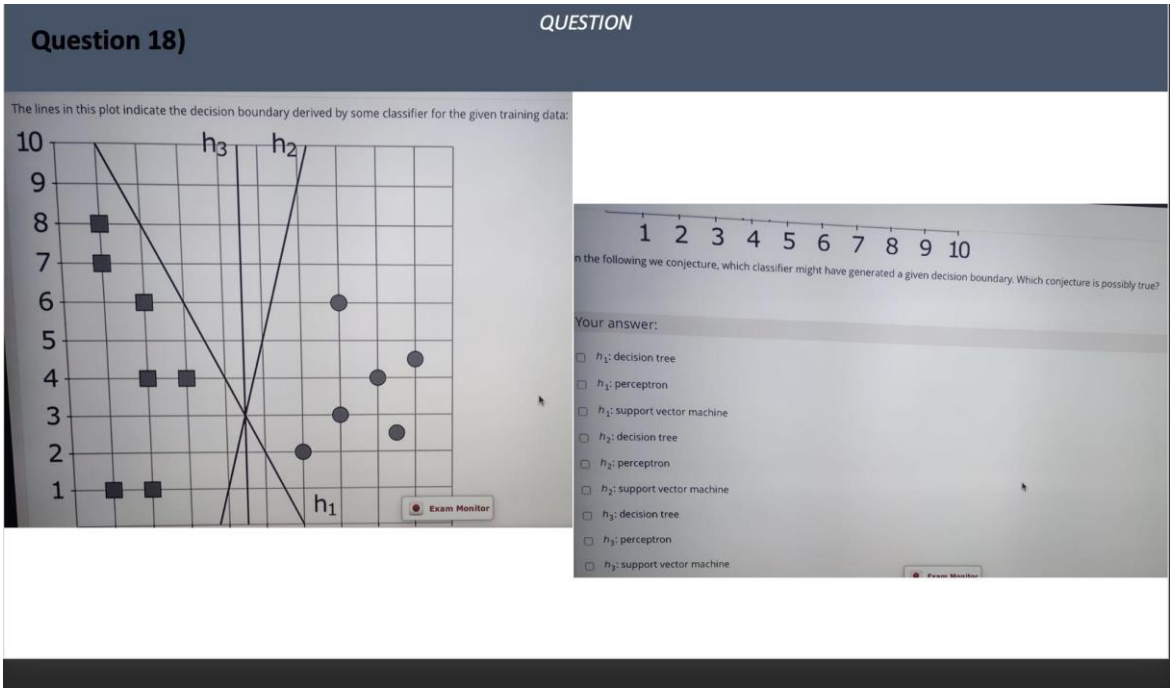
Svar til opgaven

Materiale: I forhold til ROC og AUC opgaver, så kig i følgende dokument:

- Exercise 10-4 Evaluation of Outlier Scores

Direkte svar:

Question 18: Support vector machinces



Svar til opgaven

Materiale: I forhold til given decision boundary, så kig i følgende dokument:

- Exercise 13-2 Support vectors and margin

Direkte svar

H1 = support vector machines og perceptions (decision tree cannot have a slope, therefore it cannot be a decision tree)

H2 = support vector machines og perceptions (decision tree cannot have a slope, therefore it cannot be a decision tree)

H3 = Decision tree og Support vector machines

- **This can be a decision tree, based as the line do not have a slope, but it can also be a support vector machine, which have made these decision boundaries.**