NTNU
Kunnskap for en bedre verden

KANDIDAT

10264

PRØVE

TDT4300 1 Datavarehus og datagruvedrift

Emnekode	TDT4300
Vurderingsform	Hjemmeeksamen
Starttid	24.05.2022 13:00
Sluttid	24.05.2022 16:00
Sensurfrist	18.06.2022 21:59
PDF opprettet	27.04.2023 09:40

General Information

Oppgave	Tittel	Oppgavetype
i	General Instructions	Informasjon eller ressurser

Datawarehouses and OLAP Operations

Oppgave	Tittel	Oppgavetype
1	Datawarehouses and OLAP Operations	Filopplasting

Data

Oppgave	Tittel	Oppgavetype
2	Data	Langsvar

Association Rule Mining

Oppgave	Tittel	Oppgavetype
3	FP-Growth Algorithm	Filopplasting

Clustering

Oppgave	Tittel	Oppgavetype
4	Hierarchical Clusterings	Langsvar
5	DB Scan Clustering	Filopplasting

Classification

Oppgave	Tittel	Oppgavetype
6	Decision Trees	Filopplasting

1 Datawarehouses and OLAP Operations

Totalt antall poeng: 20 poeng (=2+6+6+2+4)

De forente nasjoner (FN), i sitt initiativ for å håndtere klimaendringer, samler inn meteorologiske data fra hele verden. Du jobber som dataforsker i FN og har i oppgave å analysere globale værmeldinger for å identifisere uvanlige værmønstre (f.eks. hetebølger). De meteorologiske avdelingene til hvert land gir informasjon om temperatur i grader Celsius for hver større by fire ganger daglig (morgen, ettermiddag, kveld og natt). For å oppdage og forstå uvanlige værmønstre gir de også informasjon om luftkvalitet som en indeksverdi i [0, 201), vindhastighet i km/t, nedbør i mm/t, og UV-nivåer som indeksverdier i [1, 11). Med denne konteksten, svar på følgende spørsmål knyttet til dataanalyse:

1. Du har i oppgave å gi innsikt fra disse store værdataene. Du bestemmer deg for i utgangspunktet å bruke relasjonsdatabaseteknikker. Hva er problemer du vil møte når du prøver å "cross-tabulate" (f.eks. tilby drill-down og roll-up-funksjoner) dataene med tradisjonell SQL?
2. Gitt utfordringene med relasjonsdatabaseteknikker, bestemmer du deg for å lagre dine data i et datavarehus. For dette formålet må du først designe konsepthierarkier for de forskjellige attributtene i datasettet. Skriv ned hierarkiene du foreslår. Husk at for å komme opp med disse hierarkiene må kanskje mappe kontinuerlige attributter til kategoriske nivåer. Oppgi disse mappingene og evt. andre antagelser du gjør.
3. Basert på konsepthierarkiene opprettet ovenfor, utform et stjerneskjema for lagring av dataene i et datavarehus.
4. Gitt konsepthierarkiet du har designet, hvor mange cuboids vil være nødvendig for full materialisering av kuben?
5. På FNs kommende klimakonferanse (COP26) får du oppgaven med å presentere innsikten du har fått fra datasettet. Et av de viktige spørsmålene å svare på er: hva er økningen i gjennomsnittstemperaturen i år som sammenlignet med siste tiår over hele verden? Identifiser sekvensen av OLAP-operasjoner du må utføre på base-cuboiden for å finne økningen i gjennomsnittstemperaturen i år sammenlignet med forrige tiår i Europa da ekstreme forhold eksisterte når det gjaldt nedbør, luftkvalitet, UV-nivåer og vindhastigheter (dvs. "all" i den ene enden av konsepthierarkiene dine).

Hint:

For OLAP-operasjonene forventer vi at operasjoner av typen «Roll up, Drill down , Slice, Dice og Pivot» resulterer i subkuben for å visualisere svaret. For eksempel, for spørringen "hva er det totale datamaskinsalget i Florida for kvartal 1" er svaret:

Roll Up Location: City -> State;
Roll Up Time : Weeks -> Quarter;
Dice: State = ``Florida'' AND Quarter = ``Q1'';

Din fil ble lastet opp og lagret i besvarelsen din.

Last ned | Fjern | Erstatt

Filnavn:	Task_1.pdf
Filtype:	application/pdf
Filstørrelse:	44.31 KB
Opplastingstidspunkt:	24.05.2022 16:19
Status:	**Lagret**

2 Data

Totalt antall poeng: 10 poeng (= 4 x 2.5 poeng)

Du jobber som dataingeniør på Piazza, som er vert for en plattform hvor studenter som tilhører et emne kan stille og diskutere spørsmål. Du har i oppgave med å bygge et system for å analysere studentenes innlegg på Piazza for TDT4300 . Følgende spørsmål dukker opp i ditt forsøk på å bygge dette systemet:

1. For å analysere innlegg må du først modellere postattributtene. Hva kan være potensielle attributter som brukes til å modellere en post beregningsmessig?
2. Hva vil være den tilknyttede typen (f.eks. diskret, kontinuerlig, intervall osv.) til attributtene du har kommet frem til?
3. En viktig funksjonalitet på Piazza er å søke etter lignende innlegg basert på tekst. Basert på hva du vet om ulike typer datasett. Hvordan ville modellere teksten i innleggene slik at søkefunksjonalitet kan bli implementert? Hvordan vil det representere følgende spørsmålsutdrag: "How do we solve the question on Apriori algorithm?".
4. For å hjelpe både instruktørene og studentene tilbyr Piazza en funksjonalitet for å oppdage innlegg som stiller det samme spørsmålet. Basert på tekstmodell valgt ovenfor, hvilken likhetsmetrikk vil du velge for å implementere denne duplikatdeteksjonsfunksjonen?

Skriv ditt svar her

1. Some potential attributed that can be used are time, subject, topic (if assignment or lecture question), title, likes.

2. time: interval, subject: Continuous, topic: Discrete, title: Continuous, likes: Discrete

3. Would use Information Retrieval consepts and clean text and make keywords. And ofcourse make document with keywords to then be able to compare to other texts. "solve" "question" "apriori" "algorithm"

4. Would use cosine similarity to find similar documents using the text model above

Ord: 75

3 FP-Growth Algorithm

Totalt antall poeng: 20 poeng

Finn de frekvente elementsettene for transaksjonsdatabasen gitt i tabellen nedenfor ved hjelp av FPGrowth-algoritmen, med minimum støttetall lik 3.

tid	itemset
t_1	ACDEF
t_2	ABCDE
t_3	BCF
t_4	ACDEF
t_5	DB

Vis FPGrowth-prosedyren trinn for trinn, inkludert byggingen av FP-treet og de projiserte FP-trærne.

Din fil ble lastet opp og lagret i besvarelsen din.

Last ned | Fjern | Erstatt

Filnavn:	Task-3.pdf
Filtype:	application/pdf
Filstørrelse:	1.55 MB
Opplastingstidspunkt:	24.05.2022 13:46
Status:	**Lagret**

4 Hierarchical Clusterings

Totalt antall poeng: 5 poeng

Dendogrammer fanger kort og godt resultatet av en hierarkisk agglomerativ klynging. I bunn og grunn er et dendogram et binært tre. For et gitt sett med n punkter, hvor mange mulige dendogrammer kan opprettes? Hint: Et tre med m noder inneholder m − 1 kanter. Videre har et binært tre med n bladnoder n − 1 interne noder.

Skriv ditt svar her

Given that dendogram is a binary tree there would be (2n)!/ (n! *(n+1)!) different trees

Ord: 15

5 DB Scan Clustering

Totalt antall poeng: 20 poeng (=10+5+5)

Gitt punktene i figuren nedenfor og følgende parametere for DB Scan-algoritmen: $\mathbf{eps} = 2$ og $\mathbf{minpts} = 2$. Gitt også følgende avstandsfunksjon L_{min} :

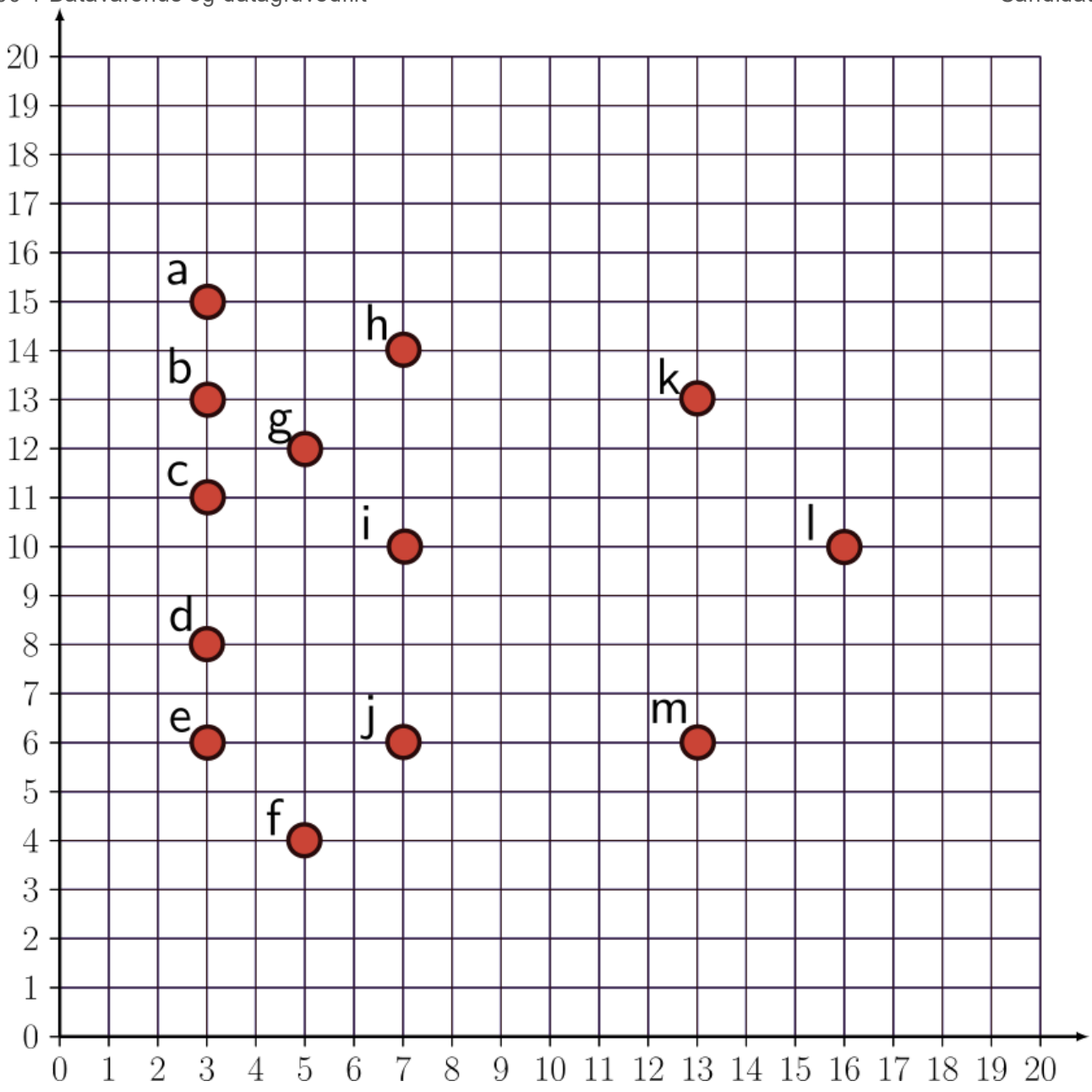
$$L_{min}(x, y) = \min_{i=1}^{d} \{|x_i - y_i|\}$$

Som et eksempel på beregning av avstander anta $x = \langle 1, 2 \rangle$ og $y = \langle 2, 4 \rangle$. Deretter beregnes L_{min} som:

$$\begin{aligned} L_{min}(x, y) &= \min_{i=1}^{2} \{|x_i - y_i|\} \\ &= \min\{|1 - 2|, |2 - 4|\} \\ &= \min\{1, 2\} \\ &= 1 \end{aligned}$$

Svar på følgende spørsmål:

1. Finn alle kjernepunktene for settet med punkter vist i figuren nedenfor.
2. To punkter sies å være tetthetsnåbare hvis det er en sekvens av kjernepunkter frem til destinasjonen. For eksempel er v tetthetnåbar fra u om det er en sekvens av kjernepunkter fra u til v. Er punktet j tetthetsnåbart fra punkt b i figuren under? Forklar resonnementet ditt.
3. Er tetthetsnåbarhet et symmetrisk forhold? Forklar resonnementet ditt.


a
b
c
d
e
f
g
h
i
j
k
l
m

Din fil ble lastet opp og lagret i besvarelsen din.

Last ned | Fjern | Erstatt

Filnavn:	Task_5.pdf
Filtype:	application/pdf
Filstørrelse:	79.06 KB
Opplastingstidspunkt:	24.05.2022 15:11
Status:	**Lagret**

6 Decision Trees

Totalt antall poeng: 25 poeng

Konstruer et beslutningstre ved å bruke Hunt's Algorithm for datasettet gitt i tabellen nedenfor der attributtet **Class** er klassifiseringsetiketten for hver post. Bruk Gini-indeksen for å bestemme de beste splittingene.

Instance	a_1	a_2	a_3	Class
1	M	X	A	YES
2	F	Y	B	YES
3	M	Y	C	YES
4	F	Y	C	YES
5	M	X	C	YES
6	F	Y	D	NO
7	M	Y	A	NO
8	F	X	A	NO
9	M	Y	A	NO
10	F	X	C	NO

Din fil ble lastet opp og lagret i besvarelsen din.

Last ned | Fjern | Erstatt

Filnavn:	Task_6.pdf
Filtype:	application/pdf
Filstørrelse:	163.84 KB
Opplastingstidspunkt:	24.05.2022 16:03
Status:	**Lagret**