Institutt for datateknologi og informatikk

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg
Tlf.: 41440433
Eksamensdato: 22. mai 2018
Eksamenstid (fra-til): 1500-1900
Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.
Annen informasjon:

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Oppgave 1 – Diverse – 15 % (alle deler teller likt)

a) Hva er binærisering, og hvordan bør man gjøre dette?
b) Silhouett-koeffisienten er gitt ved følgende formel: s = (b-a)/*max*(a,b)
   Forklar hva denne kan brukes til, og hvordan man regner ut *a* og *b* i denne.
c) Forklar viktigste begrensninger for bruk av hierarkisk agglomerativ klynging (HAC) på store datasett.

Oppgave 2 – Modellering – 10 %

Ilsvika Elektrisitetsverk (IE) leverer strøm til et stort antall beboere i Trøndelag. Alle abonnenter skal nå få montert "smarte strømmålere", som en gang i minuttet sender en melding til IE om strømforbruk siste minuttet. Med smarte strømmålere er det mulig å tilby dynamisk prising, dvs. prisen kan endre seg fra minutt til minutt, slik at man f.eks. må betale mer for strømmen i perioder med høyt strømforbruk (for eksempel når alle lager middag på ettermiddagen), og mindre når det er lavt strømforbruk (f.eks. om natten). En kunde kan ha strøm-abonnement for mer enn en lokasjon, det er da en strømmåler for hver lokasjon. IE ønsker et datavarehus som kan brukes til å analysere strømforbruk.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Total-forbruk for hver time.
- Total-forbruk for hver time for hver kunde.
- Total-forbruk for hver time for hver lokasjon.
- Totalt-forbruk per døgn per kommune.

Beskrivelsen er litt upresis og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

Lag et stjerne-skjema for denne case-beskrivelsen. Svar på papir.

Oppgave 3 – OLAP – 10 % (alle deler teller likt)

a) Gitt en base-kuboid har man tre alternative strategier for datakube-materialisering. Forklar disse, og eventuelle fordeler/ulemper for hver av dem.
b) Gitt en kube med dimensjoner:

   Time(day-month-quarter-year)
   Item(item_name-brand-type)
   Location(street-city-province_or_state-country)

   Anta følgende materialiserte kuboider:

   1) {*year, brand, city*}
   2) {*year, brand, street*}
   3) {month, *brand, province_or_state*}
   4) {*item_name, province_or_state*} where *year = 2006*

Gitt følgende OLAP-spørring: {*item_name*, *country*} med vilkår "*year = 2006*"
Hvilke(n) materialiserte kuboider kan brukes til å prosessere spørringen? Begrunn svaret.

Oppgave 4 – Klynging – 15 % (5 % på a og 10 % på b)

X	**Y**
2	4
2	5
2	6
2	10
2	11
3	3
3	11
4	12
4	13
4	16
7	2
7	2

a) Gitt et *d*-dimensjonalt datasett med 1000 punkt som man ønsker å klynge vha. DBSCAN, forklar hvordan man kan finne passende verdier for parameterne *MinPts* og *Eps*.
b) Gitt et to-dimensjonalt datasett som vist i tabellen ovenfor. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt *MinPts*=4 (inkl. eget punkt) og *Eps*=3 (inkl. punkt som har distanse 3). Bruk Manhattan-distanse som avstandsmål.

Oppgave 5 – Klassifisering – 20 % (5 % på a og 15 % på b)

Nr	**A**	**B**	**C**	**D**	**E**	***Klasse***
1	L	K	R	J	2	*J*
2	H	F	S	N	4	*J*
3	H	T	S	N	4	*J*
4	L	F	S	J	2	*N*
5	L	F	G	N	5	*N*
6	H	T	G	N	2	*N*
7	L	F	S	N	6	*N*
8	L	K	G	N	4	*N*
9	H	T	H	N	2	*J*
10	L	F	S	J	5	*N*
11	L	K	B	N	7	*N*
12	H	F	B	N	9	*J*
13	L	K	R	J	2	*N*
14	L	F	H	J	1	*N*
15	L	F	H	N	7	*N*

a) Forklar to teknikker for å redusere problem med overtilpasning ("overfitting") i beslutningstre (”decision tree”). Hvilken av disse er vanligvis foretrukket?

b) Som en del av en større applikasjon ønsker vi å kunne predikere klasse (*J* eller *N*) basert på innda-ta der hver post består av et sekvensnummer og attributtene A, B, C, D, og E, jfr tabellen ovenfor.

Anta at vi skal bruke *beslutningstre* som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet (”impurity”), og følgende to formler kan være til hjelp for å løse oppgaven:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GAIN_{split} = GINI(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \right)$$

Oppgave: Målet med klassifiseringen er å kunne predikere ”Klasse”. Regn ut $GAIN_{split}$ for splitting på (1) ”*A*” og (2) ”*B*”. Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 6 – Assosiasjonsregler (1) – 15 % (10 % på a og 5% på b)

TransaksjonsID	Element
T1	ACDK
T2	ADK
T3	CBDJK
T4	CEF
T5	BDEJK
T6	ADK
T7	ABDEJK
T8	BDFJK

a) Anta handlekorg-data som er gitt ovenfor. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$-metoden for kandidat-generering.
b) Gitt følgende lukkede frekvente elementsett (closed frequent itemsets): C:3, AC:2, BE:3, BCE:2 (Format: elementsett:støttetall)
Finn alle frekvente elementsett og deres støttetall.

Oppgave 7 – Assosiasjonsregler (2) – 10 %

TransaksjonsID	Element
T1	ABG
T2	ABCD
T3	ACJ
T4	BC
T5	ACH
T6	BCL
T7	ABCD
T8	ABCDE
T9	ABK

Anta handlekorg-data som er gitt ovenfor. Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 22 % (dvs. *minimum support count* er 2).
1) Konstruer et FP-tre basert på datasettet. Lever dette på papir som oppgave 8.

2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgende kolonner for å vise resultatet:

- Element
- ”Conditional pattern base”
- ”Conditional FP-tree”
- Frekvente elementsett

Oppgave 8 – FP-tre til oppgave 7 – 5 %

FP-tre til oppgave 7. Svar på papir.