Institutt for datateknologi og informatikk

Eksamensoppgåve i TDT4300 Datavarehus og datagruvedrift

Fagleg kontakt under eksamen: Kjetil Nørvåg
Tlf.: 41440433
Eksamensdato: 22. mai 2018
Eksamenstid (fra-til): 1500-1900
Hjelpemiddelkode/Tillatne hjelpemiddel: D: Ingen trykte eller handskrivne hjelpemiddel tilletne. Bestemt, enkel kalkulator tillate.
Annan informasjon:

Merk! Studentane finn sensur i Studentweb. Har du spørsmål om sensuren må du kontakte instituttet ditt. Eksamenskontoret vil ikkje kunne svare på slike spørsmål.

Oppgåve 1 – Diverse – 15 % (alle delar tel likt)

a) Kva er binærisering, og korleis bør ein gjere dette?
b) Silhouett-koeffisienten er gjeve ved følgjande formel: s = (b-a)/*max*(a,b)
   Forklar kva denne kan brukast til, og korleis ein reknar ut *a* og *b* i denne.
c) Forklar viktigaste avgrensingar for bruk av hierarkisk agglomerativ klynging (HAC) på store datasett.

Oppgåve 2 – Modellering – 10 %

Ilsvika Elektrisitetsverk (IE) leverer straum til mange bebuarar i Trøndelag. Alle abonnentar skal no få montert "smarte straummålarar", som ein gang i minuttet sender ei melding til IE om straumforbruk siste minuttet. Med smarte straummålarar det mogleg å tilby dynamisk prising, dvs. prisen kan endre seg frå minutt til minutt, slik at ein f.eks. må betale meir for straumen i periodar med høgt straumforbruk (for eksempel når alle lagar middag på ettermiddagen), og mindre når det er lavt straumforbruk (f.eks. om natta). Ein kunde kan ha straum-abonnement for meir enn ein lokasjon, det er då ein straummålar for kvar lokasjon. IE ønskjer eit datavarehus som kan brukast til å analysere straumforbruk.

Eksempel på analyser ein skal være i stand til å gjere mot datavarehuset:

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

Skildringa er litt upresist formulert og det er ein del av oppgåva å velje ut det som skal vere med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle føresetnader du finn det nødvendig å gjere.

Lag eit stjerne-skjema for denne case-skildringa. Svar på papir.

Oppgåve 3 – OLAP – 10 % (alle delar tel likt)

a) Gjeve ein base-kuboid har ein tre alternative strategiar for datakube-materialisering. Forklar disse, og eventuelle fordelar/ulemper for kvar av dei.
b) Gjeve ein kube med dimensjonar:

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

   Gå utifrå følgjande materialiserte kuboidar:

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

Gjeve følgjande OLAP-spørjing: {*item_name, country*} med vilkår "*year = 2006*"
Kva materialiserte kuboider kan brukast til å prosessere spørjinga? Grunngje svaret.

Oppgåve 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) Gjeve eit *d*-dimensjonalt datasett med 1000 punkt som ein ønskjer å klynge vha. DBSCAN, forklar korleis ein kan finne høvande verdiar for parametrane *MinPts* og *Eps*.
b) Gjeve eit to-dimensjonalt datasett som vist i tabellen ovanfor. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gjeve *MinPts*=4 (inkl. eige punkt) og *Eps*=3 (inkl. punkt som har distanse 3). Bruk Manhattan-distanse som avstandsmål.

Oppgåve 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 teknikkar for å redusere problem med overtilpassing ("overfitting") i avgjerdstre (”decision tree”). Kva for ein av desse er vanlegvis føretrekt?

b) Som en del av ein større applikasjon ønskjer vi å kunne predikere klasse (*J* eller *N*) basert på inndata der kvar post består av et sekvensnummer og attributta A, B, C, D, og E, jfr. tabellen ovanfor.

Gå utifrå at vi skal bruke *avgjerdstre* som klassifiseringsmetode. Vi bruker då data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for ureinheit (”impurity”), og følgjande to formlar kan vere til hjelp for å løyse oppgåva:

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

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

Oppgåve: Målet med klassifiseringa er å kunne predikere ”Klasse”. Rekn ut $GAIN_{split}$ for splitting på (1) ”*A*” og (2) ”*B*”. Kven av disse splittingane ville du valt for å starte opprettinga av avgjerdstreet? Grunngje svaret.

Oppgåve 6 – Assosiasjonsreglar (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) Gå utifrå handlekorg-data som er gjeve ovanfor. 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) Gjeve følgjande lukka frekvente elementsett (closed frequent itemsets): C:3, AC:2, BE:3, BCE:2 (Format: elementsett:støttetall)
Finn alle frekvente elementsett og støttetala deira.

Oppgåve 7 – Assosiasjonsreglar (2) – 10 %

TransaksjonsID	Element
T1	ABG
T2	ABCD
T3	ACJ
T4	BC
T5	ACH
T6	BCL
T7	ABCD

T8	ABCDE
T9	ABK

Gå utifrå handlekorg-data som er gjeve ovanfor. Du skal no bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 22 % (dvs. *minimum support count* er 2).
1) Konstruer eit FP-tre basert på datasettet. Lever dette på papir som oppgåve 8.
2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgjande kolonnar for å vise resultatet:

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

Oppgåve 8 – FP-tre til oppgåve 7 – 5 %

FP-tre til oppgåve 7. Svar på papir.