NTNU
Kunnskap for ei betre verd

Institutt for datateknikk og informasjonsvitskap

Eksamensoppgåve i TDT4300 Datavarehus og datagruvedrift

Fagleg kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 1. juni 2017

Eksamenstid (frå-til): 09.00-13.00

Hjelpemiddelkode/Tillatne hjelpemiddel: D: Ingen trykte eller handskrivne hjelpemiddel tilletne. Bestemt, enkel kalkulator tillate.

Annan informasjon:

Målform/språk: Nynorsk

Sidetal (utan framside): 3

Sidetal vedlegg: 0

Kontrollert av:

_____

Dato Sign

Informasjon om trykking av eksamensoppgåve

Originalen er:

1-sidig X 2-sidig ☐

svart/kvit X fargar ☐

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 – 10 % (alle delar tel likt)

a) Forklar *sidevisning* (pageview) i kontekst av web-bruk-gruvedrift.

b) Ei form for preprosessering i web-bruk-gruvedrift er s*ti-fullføring* (path completion). *Kvifor* må ein gjere dette, og *korleis*?

Oppgåve 2 – Modellering – 15 %

Miljøbomringen AS vil om kort tid få ansvaret for bomstasjonane i alle dei store byene i Noreg, og ønskjer eit datavarehus som kan brukast til å analysere trafikk, dvs. passering av bomstasjonar. Som en del av denne reorganiseringa, skal alle biler ha AutoPass (brikke) for automatisk registrering av passering. Ein kunde kan ha fleire bilar, og må då ha ein brikke for kvar bil. Prisen for kvar passering vert endra dynamisk/kontinuerleg for kvar stasjon uavhengig av andre, basert på tid på døgnet, forureining, kø-danning, etc.

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

- Tal på bom-passeringar for kvart kvartal for kvar stasjon.
- Tal på bom-passeringar for kvart kvartal for kvar bil.
- Gjennomsnittleg tal på passeringar per månad.
- Gjennomsnittspris per bil for ein bestemt stasjon.

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.

Oppgåve 3 – OLAP – 15 % (5 % på a og 10 % på b)

a) Forklar *roll-up* og *drill-down.*

b) Gitt ein dimensjonstabell *Book* i eit datavarehus, der vi ønskjer å bruke *bitmap-indeksar* på attributta Language og Binding for å kunne utføre spørjingar meir effektivt. Vis struktur og innhald for bitmap-indeksane med utgangspunkt i innhaldet i tabellen under.

Book				
RowID	**BookID**	**Title**	**Language**	**Binding**
1	45	The Hobbit	English	Hardcover
2	63	À la recherche du temps perdu	French	Hardcover
3	88	For Whom the Bell Tolls	English	Paperback
4	143	Madame Bovary	French	Paperback
5	236	La Peste	French	Hardcover
6	463	The Grapes of Wrath	English	Hardcover
7	768	The Great Gatsby	English	Paperback

Oppgåve 4 – Klynging – 10 %

Anta eit to-dimensjonalt datasett som vist i tabellen til høgre. Utfør klynging ved hjelp av K-means, med k=3 og initialsentroider S1=(4,4), S2=(5,8) og S3=(5,11). Bruk Manhattan-distanse som avstandsmål.

	X	Y
P1	4	8
P2	4	10
P3	4	13
P4	5	3
P5	5	7
P6	7	11

Oppgåve 5 – Klassifisering – 25 % (10 % på a og 15 % på b)

a) Anta eit datasett med sampla P1 = (4,8), P2 = (8,8), P3 = (8,4), P4 = (6,7), P5 = (1,10), P6 = (3,6), P7 = (2,4), P8 = (1,7), P9 = (6,4), P10 = (6,2), P11 = (6,3), P12 = (4,3), og P13=(4,4). Sampla høyrer til dei tre klyngene C1 = {P1,P2,P3,P4}, C2 = {P5,P6,P7,P8} og C3 ={P9,P10,P11,P12,P13}. Anta at klyngene dei tilhøyrer er klasse-merkelapp (class label). Klassifiser sampla A = (6,6), B = (4,6), C = (4,5), og D = (2,6) ved å bruke k-næraste-nabo-metoden (k-nearest neighbor, k-NN). Bruk Manhattan-distanse og k = 3. Forklar korleis du kjem fram til klassifiseringa av dei fire punkta.

b) Som ein del av ein større applikasjon ønskjer vi å kunne predikere klasse (*J* eller *N*) basert på inn-data der kvar post består av eit sekvensnummer og attributta A, B, C, og D:

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

Gå utifrå at vi skal bruke *avgjerdstre* (”decision tree”) 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 – 25 % (10 % på a og 15 % på b)

a) Gå utifrå handlekorg-data som er gjeve under. 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.

TransaksjonsID	Element
T1	ABCDEG
T2	CDFH
T3	AFG
T4	DF
T5	BDEG
T6	BDEG
T7	BCDEGH
T8	ACF

b) Gå utifrå handlekorg-data som er gjeve under. Du skal no bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 40 % (dvs. *minimum support count* er 2).
1) Konstruer eit FP-tre basert på datasettet.
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

TransaksjonsID	Element
T1	ACE
T2	BCE
T3	BCDE
T4	CDE
T5	DE