NTNU
Kunnskap for en bedre verden

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 1. juni 2017

Eksamenstid (fra-til): 09.00-13.00

Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpemiddel tillatt.
Bestemt, enkel kalkulator tillatt.

Annen informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 3

Antall sider vedlegg: 0

Kontrollert av:

_____

Dato Sign

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig X **2-sidig ☐**
sort/hvit X **farger ☐**

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 – 10 % (alle deler teller likt)

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

b) En form for preprosessering i web-bruk-gruvedrift er s*ti-fullføring* (path completion). *Hvorfor* må man gjøre dette, og *hvordan*?

Oppgave 2 – Modellering – 15 %

Miljøbomringen AS vil om kort tid få ansvaret for bomstasjonene i alle de store byene i Norge, og ønsker et datavarehus som kan brukes til å analysere trafikk, dvs. passering av bomstasjoner. Som en del av denne reorganiseringen, skal alle biler ha AutoPass (brikke) for automatisk registrering av passering. En kunde kan ha flere biler, og må da ha en brikke for hver bil. Prisen for hver passering endres dynamisk/kontinuerlig for hver stasjon uavhengig av andre, basert på tid på døgnet, forurensing, kø-dannelse, etc.

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

- Antall bom-passeringer for hvert kvartal for hver stasjon.
- Antall bom-passeringer for hvert kvartal for hver bil.
- Gjennomsnittlig antall passeringer per måned.
- Gjennomsnittspris per bil for en bestemt stasjon.

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.

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

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

b) Gitt en dimensjonstabell *Book* i et datavarehus, der vi ønsker å bruke *bitmap-indekser* på attributtene Language og Binding for å kunne utføre spørringer mer effektivt. Vis struktur og innhold for bitmap-indeksene med utgangspunkt i innholdet 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

Oppgave 4 – Klynging – 10 %

Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. 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

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

a) Anta et datasett med samplene 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). Samplene hører til de tre klyngene C1 = {P1,P2,P3,P4}, C2 = {P5,P6,P7,P8} og C3 ={P9,P10,P11,P12,P13}. Anta at klyngen de tilhører er klasse-merkelapp (class label). Klassifiser samplene A = (6,6), B = (4,6), C = (4,5), og D = (2,6) ved å bruke k-nærmeste-nabo-metoden (k-nearest neighbor, k-NN). Bruk Manhattan-distanse og k = 3. Forklar hvordan du kommer fram til klassifiseringen av de fire punktene.

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, 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*

Anta at vi skal bruke *beslutningstre* (”decision tree”) 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 – 25 % (10 % på a og 15 % på b)

a) Anta handlekorg-data som er gitt 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) Anta handlekorg-data som er gitt under. Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 40 % (dvs. *minimum support count* er 2).
1) Konstruer et FP-tre basert på datasettet.
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

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