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: 26. mai 2015

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): 4

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 – 15 % (alle deler teller likt)

a) Beskriv kort fire formål med klyngevalidering/evaluering.

b) Forklar fire teknikker for data-vasking i kontekst av web-bruk-data.

c) Anta to bit-vektorer *p* og *q*:

p = 1 0 1 0 0 0 0 1 1 1
q = 1 0 0 0 0 0 1 1 0 1

Regn ut Jaccard-koeffisienten for bitvektorene *p* og *q*.

Oppgave 2 – Modellering – 15 %

I denne oppgaven skal du modellere et datavarehus for bilskader i forsikringsselskapet Lillebrand. Lillebrand ønsker et datavarehus for å kunne analysere hendelser som har medført forsikringsutbetalinger.

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

- Antall skader i 2015.
- Gjennomsnittlig antall skader per måned.
- Antall skader for hvert kvartal i 2015.
- Totalt beløp utbetalt for hver biltype.
- Antall skader av type ”kollisjon” per by.

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) 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, item_name, city*}
2) {*year, brand, country*}
3) {*year, brand, province_or_state*}
4) {*item_name, province_or_state*} where *year = 2004*

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

b) Gitt et datavarehus med tre tabeller Location/Item/Sales, der Sales er fakta-tabellen og de to andre er dimensjonstabeller. Vi ønsker å bruke *join-indekser* for å kunne utføre spørringer mer effektivt. Vis struktur og innhold for join-indeksene Location/Sales og Item/Sales med utgangspunkt i innholdet i de tre tabellene under.

Location	
LocKey	**CityName**
L1	Oslo
L2	Athen
L3	Trondheim

Item	
ItemKey	**ItemName**
I1	Sony-TV
I2	Rolex
I3	Lexus

Sales			
TransID	**LocKey**	**ItemKey**	**Price**
T1	L1	I1	5
T2	L2	I2	8
T3	L1	I1	6
T4	L3	I1	5
T5	L3	I3	9
T6	L1	I2	8
T7	L1	I1	4

Oppgave 4 – Klynging – 10 %

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

X	Y
4	8
4	9
4	10
4	13
4	14
5	3
5	7
5	14
6	15
6	16
6	19
7	11
7	16
7	17
7	18
7	19

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

a) Forklar *kryssvalidering* (”cross validation”) og hva denne teknikken brukes til.

b) Et bilforsikringsselskap har for eksisterende kunder lagret informasjon som inkluderer kundenr, alder (L/M/H, dvs. 18-25/26-70/71-100), biltype, kjørelengde per år (4000/8000/20000/Ubegrenset), bonus (Lav/Middels/Høy) og om de har hatt skade på bilen som ble dekket av forsikringen. Når nye kunder ber om tilbud på forsikring, ønsker selskapet å sette prisen til normal eller høy basert på om de tror kunden kommer til å få skade på bilen eller ikke, dvs. de ønsker å predikere attributtet ”Skade”.

Kundenr	**Alder**	**Biltype**	**Kjørelengde per år**	**Bonus**	***Skade***
1	L	Ferrari	8000	Lav	*Ja*
2	M	BMW	8000	Høy	*Nei*
3	H	Lexus	Ubegrenset	Høy	*Ja*
4	L	Audi	8000	Høy	*Nei*
5	H	Opel	8000	Lav	*Ja*
6	M	Toyota	8000	Lav	*Nei*
7	M	Honda	8000	Høy	*Nei*
8	M	Nissan	8000	Høy	*Nei*
9	M	Audi	Ubegrenset	Høy	*Nei*
10	M	BMW	8000	Lav	*Ja*
11	H	Toyota	Ubegrenset	Høy	*Nei*
12	L	Nissan	4000	Lav	*Ja*
13	L	Opel	Ubegrenset	Høy	*Ja*
14	M	Audi	8000	Høy	*Nei*
15	M	Opel	8000	Høy	*Nei*
16	M	Toyota	4000	Lav	*Nei*

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 ”Skade”. Regn ut $GAIN_{split}$ for splitting på (1) ”*Alder*” og (2) ”*Bonus*”. Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 6 – Assosiasjonsregler – 25 % (10 % på a, 5 % på b, og 10 % på c)

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	ABCDFGH
T2	DKM
T3	FK
T4	ACGH
T5	ACDDGH
T6	BM
T7	DFKM
T8	ABCDGH

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 60 % (dvs. *minimum support count* er 3).
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	f, a, c, d, g, i, m, p
T2	a, b, c, f, l, m, o
T3	b, f, h, j, o
T4	b, c, k, s, p
T5	a, f, c, e, l, p, m, n