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: 26. mai 2016

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): 4

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 – 15 % (alle delar tel likt)

a) Nemn fire formål med klyngevalidering/evaluering.

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

c) Gjeve to bit-vektorar *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

Rekn ut Jaccard-koeffisienten for bitvektorane *p* og *q*.

Oppgåve 2 – Modellering – 15 %

I denne oppgåva skal de modellere eit datavarehus for bilskadar i forsikringsselskapet Lillebrand. Lillebrand ønskjer eit datavarehus for å kunne analysere hendingar som har medført forsikringsutbetalingar.

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

- Tal på skadar i 2015.
- Gjennomsnittleg tal på skadar per månad.
- Tal på skadar for kvart kvartal i 2015.
- Totalt beløp utbetalt for kvar biltype.
- Tal på skadar av type ”kollisjon” per by.

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

Gjeve følgjande OLAP-spørjing: {*item_name, province_or_state*} med vilkår "*year = 2006*" Kven av dei materialiserte kuboidane kan brukast til å prosessere spørjinga? Grunngje svaret.

b) Gjeve eit datavarehus med tre tabellar Location/Item/Sales, der Sales er fakta-tabellen og dei to andre er dimensjonstabellar. Vi ønskjer å bruke *join-indekser* for å kunne utføre spørjingar meir effektivt. Vis struktur og innhald for join-indeksane Location/Sales og Item/Sales med utgangspunkt i innhaldet i dei tre tabellane 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

Oppgåve 4 – Klynging – 10 %

Gjeve et to-dimensjonalt datasett som vist i tabellen til høgre. 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.

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

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

a) Forklar *kryssvalidering* (”cross validation”) og kva denne teknikken vert brukt til.

b) Eit bilforsikringsselskap har for eksisterande kundar lagra informasjon som inkluderer kundenr, alder (L/M/H, dvs. 18-25/26-70/71-100), biltype, køyrelengde per år (4000/8000/20000/Uavgrensa), bonus (Lav/Middels/Høg) og om dei har hatt skade på bilen som vart dekka av forsikringa. Når nye kundar bed om tilbod på forsikring, ønskjer selskapet å sette prisen til normal eller høg basert på om dei trur kunden kjem til å få skade på bilen eller ikkje, dvs. dei ønskjer å predikere attributtet ”Skade”.

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

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 ”Skade”. Rekn ut $GAIN_{split}$ for splitting på (1) ”*Alder*” og (2) ”*Bonus*”. Kven av disse splittingane ville du valt for å starte opprettinga av avgjerdstreet? Grunngje svaret.

Oppgåve 6 – Assosiasjonsreglar – 25 % (10 % på a, 5 % på b, og 10 % på c)

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

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 60 % (dvs. *minimum support count* er 3).
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	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