NTNU
Det skapande universitetet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgåve i TDT4300 Datavarehus og datagruvedrift

Fagleg kontakt under eksamen: Kjetil Nørvåg
Tlf.: 735 96755

Eksamensdato: 29. mai
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: 4
Sidetal vedlegg: 0

Kontrollert av:

_____

Dato Sign

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

a) Gjeve to bit-vektorar *p* og *q*:

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

Rekn ut Jaccard-koeffisienten for bitvektorane *p* og *q*. Kva er fordelen med Jaccard i høve til ”simple matching”?

b) Forklar 3 teknikkar som kan brukast til pre-prosessering av numeriske data.

c) Forklar prinsippa bak bitmap-indeksar. For kva slags data er denne type indeks eigna, og når er den ikkje eigna?

Oppgåve 2 – Modellering – 20% (15% på a, 5% på b)

I denne oppgava skal de modellere eit datavarehus for ein regional vervarslingsteneste. Denne har ca. 1000 målestasjonar, som er spreidd over ulike land- og hav-område i regionen for å samle inn grunnleggjande verdata, som lufttrykk, temperatur og nedbør for kvar time. Alle data vert sende til hovedsentralen, som har samla inn slike data i over 10 år. Ditt design bør leggje til rette for effektive spørjingar og on-line analytisk handsaming, og utleie generelle vermønster.

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.

a) Lag eit stjerne- eller snøflak-skjema for denne case-beskrivelsen.

b) Lag to forskjellige konsepthierarki (fritt valde dimensjoner).

Oppgåve 3 – Klynging – 20 % (5% på a, 15% på b)

a) Forklar potensielle ulemper med hierarkisk klynging.

b) 1) Forklar DBSCAN-algoritmen.
2) Gjeve eit to-dimensjonalt datasett som vist i tabellen til høgre. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gjeve MinPts=3 og Eps=3.

X	Y
2	3
4	5
6	4
6	5
7	5
7	12
8	2
8	10
8	14
9	12
9	13
10	12
11	16
13	16
13	18
16	16
16	19

Oppgåve 4 – Assosiasjonsreglar – 20 %

Gå utifrå handlekorg-data til høyre. Bruk apriori-algoritmen for å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Vel deretter eit av dei frekvente 3-elementsetta og finn alle assosiasjonsreglar basert på dette settet, gjeve konfidens på 75 %.

TransaksjonsID	Element
T1	A,B,C,D
T2	A,G
T3	A,C,E,F
T4	B,C,G
T5	A,C,E,F
T6	C,D
T7	A,B,C,E,F
T8	A,B,C,E,F,G

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

a) Forklar kva som er føremålet med *klassifisering*. Gje tre eksempel på typiske oppgåver som ein kan løyse ved hjelp av klassifisering.

b) Forklar kort prinsippa bak *næraste-nabo-klassifisering (nearest neighbour classification).*

c) Magnus Carlsen og Vishy Anand skal spele VM-finale mot kvarandre seinare i år. Det er bestemt at finalen skal gå i India. Sjå for deg at disse to har møtt kvarandre fleire gongar tidligare, og at vi har fått tak i data og informasjon om møta. Vi får også vite at resultata til Carlsen tidlegare har vært avhengig av kor mye innsats han har lagt i å forberede møta. I fylgjande tabell er eit datasett som viser verdien 1 dersom Carlsen har ytt full innsats mens verdien 0 betyr innsatsen hans ikkje har vore heilt topp. Typen kamp og kamptidspunkt, samt stad for kampane vert også teke omsyn til her.

Tid	**Kamptype**	**Sted**	**Innsats**	**Resultat**
Morgen	Master	Indoor	1	C
Ettermiddag	Grand Tour	Indoor Crowded	1	C
Kveld	Show	Mall	0	C
Ettermiddag	Show	Mixed	0	A
Ettermiddag	Master	Indoor Crowded	1	A
Ettermiddag	Grand Tour	Indoor	1	C
Ettermiddag	Grand Tour	Mall	1	C
Ettermiddag	Grand Tour	Mall	1	C
Morgen	Master	Indoor	1	C
Ettermiddag	Grand Tour	Indoor Crowded	1	A
Kveld	Show	Mall	0	C
Kveld	Master	Mixed	1	A
Ettermiddag	Master	Indoor Crowded	1	A
Ettermiddag	Master	Indoor	1	C
Ettermiddag	Grand Tour	Mall	1	C
Ettermiddag	Grand Tour	Indoor Crowded	1	C

Kampar som har enda uavgjort (remis) er ikkje med i datasettet.

Gå utifrå at vi skal bruke *avgjerdstre* som klassifiseringsmetode. Vi bruker då data i tabellen over som treningsdata. For å avgjere den beste splitten treng vi å bruke **Entropy** for ein node *t* som følgjer: $Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$, kor *p(j|t)* er sannsyn for klasse *j* gitt node *t (*dvs. andelen av klasse *j* i node *t)* . For kvar splitting er "information gain" gjeve som $GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$, der n_i er tal på element i node *i* og *n* total element i forelder-noden *p*.

Oppgåve: Målet med klassifiseringa er å kunne predikere utfallet av framtidige kamper mellom Carlsen og Anand. Rekn ut GAIN for splitting på (1) "**Tid**" og (2) "**Kamptype**". Kven av desse splittingane ville du valt for å starte opprettinga av avgjerdstreet? Grunngje svaret ditt.