NTNU
Det skapande universitetet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

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

Eksamensdato: 29. mai
Eksamenstid (fra-til): 09.00-13.00
Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrivne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.

Annen informasjon:

Målform/språk: Bokmål
Antall sider: 4
Antall sider vedlegg: 0

Kontrollert av:

_____

Dato Sign

Oppgave 1 – Diverse – 15% (alle deler teller likt)

a) Gitt to bit-vektorer *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

Regn ut Jaccard-koeffisienten for bitvektorene *p* og *q*. Hva er fordelen med Jaccard i forhold til "simple matching"?

b) Forklar 3 teknikker som kan brukes til pre-prosessering av numeriske data.

c) Forklar prinsippene bak bitmap-indekser. For hva slags data er denne type indeks egnet, og når er den ikke egnet?

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

I denne oppgaven skal dere modellere et datavarehus for en regional værvarslingstjeneste. Denne har ca. 1000 målestasjoner, som er spredt over ulike land- og hav-områder i regionen for å samle inn grunnleggende værdata, herunder lufttrykk, temperatur og nedbør for hver time. Alle data blir sendt til hovedsentralen, som har samlet inn slike data i over 10 år. Ditt design bør legge til rette for effektive spørringer og on-line analytisk behandling, og utlede generelle værmønstre.

Beskrivelsen er litt upresist formulert 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.

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

b) Lag to forskjellige konsepthierarkier (fritt valgte dimensjoner).

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

a) Forklar potensielle ulemper med hierarkisk klynging.

b) 1) Forklar DBSCAN-algoritmen.
2) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt 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

Oppgave 4 – Assosiasjonsregler – 20 %

Anta handlekurv-data til høyre. Bruk apriori-algoritmen for å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Velg deretter et av de frekvente 3-elementsettene og finn alle assosiasjonsregler basert på dette settet, gitt 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

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

a) Forklar hva som er hensiktene med *klassifisering*. Gi tre eksempler på typiske oppgaver som kan løses ved hjelp av klassifisering.

b) Forklar kort prinsippene bak *nærmeste-nabo-klassifisering (nearest neighbour classification).*

c) Magnus Carlsen og Vishy Anand skal spille VM-finale mot hverandre senere i år. Det er bestemt at finalen skal gå i India. Se for deg at disse to har truffet hverandre flere ganger tidligere, og at vi har fått tak i data og informasjon om møtene. Vi får også vite at resultatene til Carlsen tidligere har vært avhengig av hvor mye innsats han har lagt i å forberede møtene. I følgende tabell er et datasett som viser verdien 1 hvis Carlsen har ytt full innsats mens verdien 0 betyr innsatsen hans ikke har vært helt topp. Typen kamp og kamptidspunkt, samt sted for kampene tas også hensyn 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

Kamper som har endt uavgjort (remis) er ikke med i datasettet.

Anta at vi skal bruke *beslutningstre* som klassifiseringsmetode. Vi bruker da dataene i tabellen over som treningsdata. For å avgjøre den beste splitten trenger vi å bruke **Entropy** for en node *t* som følger: $Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$, hvor *p(j|t)* er sannsynligheten for klasse *j* gitt node *t (*dvs. andelen av klasse *j* i node *t)* . For hver splitting er "information gain" angitt som $GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$, hvor n_i er antall elementer i node *i* og *n* total elementer i forelder-noden *p*.

Oppgave: Målet med klassifiseringen er å kunne predikere utfallet av fremtidige kamper mellom Carlsen og Anand. Beregn GAIN for splitting på (1) "**Tid**" og (2) "**Kamptype**". Hvilken av disse splittingene ville du valgt for å starte opprettelsen av beslutningstreet? Begrunn svaret ditt.