NTNU
Det skapande universitetet

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Heri Ramampiaro
Tlf.: 7359 3440

Eksamensdato: 7. august
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: 3
Antall sider vedlegg: 0

Kontrollert av:

Dato Sign

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

a) Forklar hensikt og teknikker for diskretisering og dimensjonalitetsreduksjon.

b) Forklar prinsippene bak join-indekser.

c) Gitt settet med verdier under, finn 10 %, 25 %, 50 %, 75 % og 100 % persentiler (*10th,25th,50th,75th and 100th percentiles*)
Verdier: 1,2,3,5,5,7,7,10,12,15,18,18,19,20,22,23,24,27,28,32

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

I denne oppgaven ser vi på en kjede av butikker som selger aviser og blader. Kjeden selger mange forskjellige typer publikasjoner (for eksempel knyttet til mote, barn, biler, sport) fra mange forskjellige utgivere. Butikkene spenner fra små nærbutikker til større butikker med samlokaliserte kafeer. Imidlertid er kjeden litt gammeldags, der hver butikksjef på slutten av hver dag legger inn informasjon i et regneark om hvor mange eksemplarer som ble solgt av hver publikasjon. Dette regnearket blir så sendt til hovedkvarteret, og for tiden er dette den eneste måten hovedkvarter kan samle inn og analysere salgsdata fra butikkene. Ledelsen ønsker nå å lage et datavarehus for å få mer innsikt i salget av de ulike publikasjonene (og typer publikasjoner) fra hver butikk.

Beskrivelsen er litt upresist formulert og det er en del av oppgaven å velge ut hva 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 snøflak-skjema for denne case-beskrivelsen.

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

Oppgave 3 – Klassifisering – 20 % (8 % på a og b, 4 % på c)

a) Forklar hvordan man kan bestemme hvilke attributter man skal splitte på når man lager et beslutningstre.

b) Forklar hvordan man kan splitte på kontinuerlige attributter.

c) Forklar metrikkene *Accuracy* og *Error rate* i forbindelse med klassifisering.

Oppgave 4 – Klynging – 25 % (5 % på a, 10 % på b og c)

a) Silhuett-koeffisienten kan beskrives som s = (b-a)/*max*(a,b).
Forklar innholdet i formelen, og hva Silhuett-koeffisienten kan brukes til.

b) Forklar algoritmen for *bisecting k-means.* Hva er fordelen med denne framfor vanlig *k*-means?

c) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av hierarkisk klynging på dette datasettet. Bruk MIN (single link) som inter-klynge distanse.

X	Y
2	2
3	2
4	8
5	4
5	7
7	4
9	17
13	4
17	4

Oppgave 5 – 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). Finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 %.

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