NTNU
Kunnskap for en bedre verden

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 41440433

Eksamensdato: 5. juni 2015

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 (uten forside): 3

Antall sider vedlegg: 0

Kontrollert av:

_____

Dato Sign

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

a) Forklar *asymmetriske attributt.* Gi et eksempel på et slikt attributt.

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

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

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

c) I mange datasett kan verdier mangle for attributt i noen av objektene, ofte fordi noen attributt ikke er relevante for alle (f.eks. barn har typisk ikke inntekt). Gi tre metoder/strategier man kan bruke for å håndtere manglende verdier.

Oppgave 2 – Modellering – 20 % (17 % på a, 3 % på b)

I denne oppgaven skal dere modellere et datavarehus for Netflix. Netflix tilbyr strømming av TV-serier og filmer, og ønsker et datavarehus for å kunne analysere visninger av TV-serier (for enkelhets skyld kan dere se bort fra filmer i denne oppgaven). En *visning* er i denne sammenheng definert som hendelsen at en bruker ser på en TV-episode eller deler av en TV-episode.

For å forenkle modelleringen kan dere anta at tidspunktet for en visning er tidspunktet den starter, og at laveste granularitet for visning er *kapittel* (dvs. dere trenger ikke modellere start- og slutt-tidspunkt), der man antar at en episode består av ett eller flere kapitler.

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

- Gjennomsnittlig lengde (tid) på hver visning.
- Visningsmetode (f.eks. Android-app, nettleser, etc.) per kvartal.
- Antall visninger for hvert kapittel av en bestemt TV-serie for hvert land.

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-skjema for denne case-beskrivelsen.

b) Konsepthierarki for tid kan f.eks. være *år-kvartal-måned-dag*. Kan *uke* være en del av dette hierarkiet? Begrunn svaret.

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

a) Forklar fordeler og ulemper med k-means.

b) 1) Forklar hierarkisk agglomerativ klynging.
2) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør hierarkisk agglomerativ klynging på dette datasettet ved å bruke MIN (single link) og Manhattan-distanse. Vis det resulterende dendrogrammet.

X	Y
2	3
4	5
6	4
6	5
7	5
7	12
8	2
8	10

Oppgave 4 – Klassifisering – 25 % (5 % på a og 20 % på b)

a) Forklar *forvekslingsmatrise* (”confusion matrix”), innholdet i denne, og hvordan man regner ut *nøyaktighet* (”accuracy”) basert på denne.

b) Rosenborg og Vålerenga skal i morgen (lørdag) spille tippeliga-kamp på Ullevaal (som er hjemmestadion for Vålerenga). Disse har spilt mot hverandre mange ganger tidligere, og vi ønsker å bruke resultat og informasjon fra tidligere kamper til å predikere morgendagens resultat. Denne informasjonen er vist i tabellen under (kamper som har endt uavgjort er ikke med i datasettet, H/B betyr Rosenborg hjemme/borte).

Dag	**Turnering**	**Sted**	**Tidspunkt**	**Resultat**
Fredag	Tippeligaen	H	Ettermiddag	R
Søndag	NM	H	Kveld	R
Søndag	Tippeligaen	B	Ettermiddag	R
Søndag	Tippeligaen	H	Kveld	R
Lørdag	Tippeligaen	B	Ettermiddag	V
Søndag	Tippeligaen	H	Ettermiddag	R
Søndag	Tippeligaen	H	Kveld	R
Lørdag	Tippeligaen	B	Ettermiddag	R
Søndag	Tippeligaen	H	Kveld	R
Søndag	Tippeligaen	H	Ettermiddag	R
Fredag	Tippeligaen	H	Kveld	R
Søndag	Tippeligaen	B	Kveld	V
Lørdag	Tippeligaen	H	Ettermiddag	R
Søndag	Tippeligaen	B	Ettermiddag	R
Søndag	Tippeligaen	B	Kveld	V
Lørdag	Tippeligaen	H	Ettermiddag	V

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 utfallet av morgendagens kamp mellom Rosenborg og Vålerenga. Regn ut $GAIN_{split}$ for splitting på (1) ”*Sted*” og (2) ”*Dag*” Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 5 – Assosiasjonsregler – 20 %

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). Vis hvordan kandidatsettene blir generert.

Et av de frekvente elementsettene er BDE. Finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 % (det er ikke nødvendig å bruke apriori til å finne assosiasjonsreglene, men vis hvordan konfidens blir regnet ut for hver av kandidatreglene som er basert på BDE).

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