NTNU
Kunnskap for ei betre verd

Institutt for datateknikk og informasjonsvitskap

# Eksamensoppgåve i TDT4300 Datavarehus og datagruvedrift

**Fagleg kontakt under eksamen: Kjetil Nørvåg**

**Tlf.: 41440433**

**Eksamensdato: 5. juni 2015**

**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): 3**

**Sidetal vedlegg: 0**

**Kontrollert av:**

_____

Dato Sign

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) Forklar *asymmetriske attributt.* Gje eit eksempel på eit slikt attributt.

b) Gjeve to bit-vektorar *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

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

c) I mange datasett kan verdiar mangle for attributt i nokre av objekta, ofte fordi nokre attributt ikkje er relevante for alle (f.eks. barn har typisk ikkje inntekt). Gje tre metodar/strategiar ein kan bruke for å handsame manglande verdiar.

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

I denne oppgåva skal de modellere eit datavarehus for Netflix. Netflix tilbyr strauming av TV-seriar og filmar, og ønskjer eit datavarehus for å kunne analysere visningar av TV-seriar (for enkelheits skuld kan de sjå bort frå filmar i denne oppgåva). Ei *visning* er i denne samanheng definert som hendinga at ein brukar ser på ein TV-episode eller delar av en TV-episode.

For å forenkle modelleringa kan de gå utifrå at tidspunktet for ei visning er tidspunktet den startar, og at lavast granularitet for visning er *kapittel* (dvs. de treng ikkje modellere start- og slutt-tidspunkt), der ein går utifrå at ein episode består av eit eller fleire kapittel.

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

- Gjennomsnittleg lengde (tid) på kvar visning.
- Visningsmetode (t.d. Android-app, nettleser, etc.) per kvartal.
- Tal på visningar for kvart kapittel av ein bestemt TV-serie for kvart land.

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-skjema for denne case-skildringa.

b) Konsepthierarki for tid kan t.d. vere *år-kvartal-månad-dag*. Kan *veke* vere ein del av dette hierarkiet? Grunngje svaret.

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

a) Forklar fordelar og ulemper med k-means.

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

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

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

a) Forklar *forvekslingsmatrise* (”confusion matrix”), innhaldet i denne, og korleis ein reknar ut *nøyaktigheit* (”accuracy”) basert på denne.

b) Rosenborg og Vålerenga skal i morgon (laurdag) spele tippeliga-kamp på Ullevaal (som er heimestadion for Vålerenga). Desse har spela mot kvarandre mange gongar tidlegare, og vi ønskjer å bruke resultat og informasjon frå tidlegare kampar til å predikere morgondagens resultat. Denne informasjonen er vist i tabellen under (kampar som har enda uavgjort er ikkje med i datasettet, H/B betyr Rosenborg heime/borte).

| **Dag** | **Turnering** | **Stad** | **Tidspunkt** | **Resultat** |
|---|---|---|---|---|
| Fredag | Tippeligaen | H | Ettermiddag | R |
| Søndag | NM | H | Kveld | R |
| Søndag | Tippeligaen | B | Ettermiddag | R |
| Søndag | Tippeligaen | H | Kveld | R |
| Laurdag | Tippeligaen | B | Ettermiddag | V |
| Søndag | Tippeligaen | H | Ettermiddag | R |
| Søndag | Tippeligaen | H | Kveld | R |
| Laurdag | Tippeligaen | B | Ettermiddag | R |
| Søndag | Tippeligaen | H | Kveld | R |
| Søndag | Tippeligaen | H | Ettermiddag | R |
| Fredag | Tippeligaen | H | Kveld | R |
| Søndag | Tippeligaen | B | Kveld | V |
| Laurdag | Tippeligaen | H | Ettermiddag | R |
| Søndag | Tippeligaen | B | Ettermiddag | R |
| Søndag | Tippeligaen | B | Kveld | V |
| Laurdag | Tippeligaen | H | Ettermiddag | V |

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 utfallet av morgondagens kamp mellom Rosenborg og Vålerenga. Rekn ut $GAIN_{split}$ for splitting på (1) ”*Stad*” og (2) ”*Dag*” Kven av desse splittingane ville du valt for å starte opprettinga av avgjerdstreet? Grunngje svaret ditt.

## Oppgåve 5 – Assosiasjonsreglar – 20 %

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). Vis korleis kandidatsetta vert generert.

Eit av dei frekvente elementsetta er BDE. Finn alle assosiasjonsreglar basert på dette settet, gjeve konfidens på 75 % (det er ikkje nødvendig å bruke apriori til å finne assosiasjonsreglane, men vis korleis konfidens vert rekna ut for kvar av kandidatreglane 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 |