Informasjon om hjemmeeksamen

Hjemmeeksamen i SOS2900 - Algoritmer, store data og samfunnsendring

2018 VÅR

Varighet: 25. april kl. 10:00 til 27. april kl. 14:00
Husk at besvarelsen skal være anonym, du skal ikke oppgi navnet ditt. Oppgavene skal merkes med ditt kandidatnummer.

Besvarelsen skal inneholde kildehenvisninger. Dersom du leverer en besvarelse med mangelfulle eller misvisende kildehenvisninger kan du bli mistenkt for fusk. Du kan lese mer om kildebruk og referanser her.

Informasjon om eksamensoppgaven
Eksamensoppgaven består av 2 oppgaver.
Begge oppgavene må besvares.

Emneansvarlig Torbjørn Skardhamar har treffetid **13:00-14:00 onsdag 25. april** for klargjørende spørsmål rundt eksamen.

Forside
Filene du laster opp må ha følgende informasjon på forsiden:
1) Oppgavedel
2) Emnets og faglærers navn
3) Dato for eksamen
4) Kandidatnummer
5) Kun for oppgave 2: Antall ord i oppgaven (forside og litteraturliste regnes ikke med)

Oppgave 1)
Tilhørende oppgaven finner du 2 oppgavesett. Disse kan lastes ned til din maskin.

Oppgaven skrives i et eget dokument som du laster opp i Inspera som en egen pdf-fil. I forbindelse med besvarelsen av oppgave 1 må du lime inn grafikk og tabeller i tillegg til tekstlig besvarelse der det trengs.
Koden du skriver i *R* legges ved til slutt i samme dokument. Pass på at koden kommer i den rekkefølgen du kjørte det.

R-koden i seg selv er primært dokumentasjon på hva du har gjort, mens det er den tekstlige besvarelsen som primært blir vurdert.

Oppgave 2)
Oppgaven skrives i et eget dokument som du laster opp i Inspera som en egen pdf-fil. Oppgaven skal ikke overstige 1500 ord.

Hvis du ønsker begrunnelse:
Fristen for å be om begrunnelse er en uke etter at karakteren er kunngjort. For muntlige og praktiske eksamener er fristen straks etter at du har fått vite karakteren din.

Du skal normalt få begrunnelsen innen to uker etter at du har bedt om den. Sensor avgjør om begrunnelsen blir gitt skriftlig eller muntlig.

Se informasjon på emnesiden for å se hvordan du kan be om begrunnelse på karakter i dette emnet.

Lykke til!

1 Oppgave 1

OBS! Husk å laste opp i pdf. Usikker på hvordan du konverterer til pdf? Se denne siden.

Datasett

- spam_data.csv
- nyepost.csv

Oppgavetekst

De fleste av oss er mer eller mindre plaget av spam i eposten. Disse epostene er laget for å få oss til å kjøpe noe, donere penger, oppgi personlig informasjon etc. Spamfilter er et automatisk system for å skille ut spam fra de epostene vi ønsker å motta. På tross av den store variasjonen i spammail er det likevel noen fellestrekk ved disse som mye sjeldnere brukes i de epostene vi vil ha. Dette er et klassifikasjonsproblem egnet for maskinlæring.

Datasettet `spam_data.csv` er offentliggjort av Hewlett-Packard i 1999 og inneholder informasjon om 4601 eposter. For hver epost er det summert opp visse egenskaper som er samlet i variable i datasettet. Følgende beskrivelse av datasettet er angitt fra HP, der variabelnavnene er uthevet:

The last column denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. For the statistical measures of each attribute, see the end of this file. Here are the definitions of the attributes:
48 continuous real [0,100] attributes of type ***word_freq_WORD*** *= percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.*
6 continuous real [0,100] attributes of type ***char_freq_CHAR*** *= percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail*
capital_run_length_average *= average length of uninterrupted sequences of capital letters*
capital_run_length_longest *= length of longest uninterrupted sequence of capital letters*
capital_run_length_total *= total number of capital letters in the e-mail*
spam *= denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.*

For eksempel inneholder variabelen `word_freq_make` hvor ofte ordet "make" er brukt angitt som andelen av totalt antall ord i eposten, og variabelen `word_freq_address` angir hvor ofte "address" er brukt som andel av totalt antall ord i eposten. Det er 48 tilsvarende variable for ulike typer ord.

Variabelen `char_freq_ordParenthesis` angir tilsvarende hvor ofte det er brukt ordinære parenteser, dvs slike: `( )` , i mailen som andel av antall tegn i eposten, mens `char_freq_bracketParenthesis` angir andel slike parentes: `[ ]`. Det er 6 tilsvarende variable for andre spesialtegn.

Til denne oppgaven trenger du følgende pakker som altså lastes ved følgende kode:

```
library(tidyverse)

library(rpart)
```

```
library(rpart.plot)

library(randomForest)

library(caret)
```

Hvis du ikke har disse installert på maskinen din fra før kan du gjøre det med følgende kode:

```
pkgs <- c("tidyverse", "rpart", "rpart.plot", "randomForest", "caret")

install.packages(pkgs, repos ="http://cran.uib.no/", dependencies =
c("Depends","Imports"))
```

Oppgave

a. Split datasettet i et training og test datasett. Lag et klassifikasjonstre med rpart( ). Dette er ditt spamfilter og nye eposter som klassifiseres som spam kan f.eks. flyttes over i en egen mappe eller slettes.
   Plot resultatet og kommenter hvordan en ny epost vil bli behandlet i dette treet.

b. Gi en vurdering av hvor presist spamfilteret vil virke ved å lage en confusion matrix med test-datasettet.

c. Datasettet `nyepost.csv` inneholder én ny epost. Bruk predict( ) på dette datasettet. Vil denne eposten klassifiseres som spam?

d. Lag et nytt spamfilter med bruk av random forest. Rapporter confusion matrix og kommenter hvor godt dette spamfilteret forventes å virke.

Maks poeng: 0

2 Oppgave 2

OBS! Maks 1500 ord

I disse dager er det stor oppmerksomhet i USA omkring bruk av data fra sosiale medier til politisk påvirkning.

Diskutér noen muligheter for bruk av «big data» og maskinlæring for å lage målrettede budskap med formål om å påvirke individers politiske valg. Ta gjerne utgangspunkt i pensum.

Vurdér noen etiske aspekter ved slik bruk av personlige data, og drøft hvorvidt de etiske aspektene er vesensforskjellig fra andre typer målretting av budskap.

Last opp filen her. Maks én fil.

Alle filtyper er tillatt. Maksimal filstørrelse er **1 GB**.

Velg fil for opplasting

OBS! Husk å laste opp i pdf. Usikker på hvordan du konverterer til pdf? Se denne siden.

Maks poeng: 0