

.

SOS2901- Anvendt maskinlæring for samfunnsvitere

Om eksamen

Eksamenssettet består av følgende oppgaver:

10 flervalgsoppgaver.

2 kortsvarsoppgaver.

1 sett oppgaver knyttet til analyse av datasett.

Poeng

For alle flervalgsoppgaver får du 1 poeng hvis du svarer riktig, og minus 0.25 poeng hvis du svarer feil. Du får 0 poeng for

oppgaver som er ubesvart. (Andre typer oppgaver gis det ikke trekk ved feil).

Tekstoppgaver har gjerne flere elementer og du bør passe på å besvare hver del da det påvirker poenggivningen. Svar presist

og utfyllende, men ikke lengre tekst enn nødvendig.

Det gis totalt 6 poeng (3 poeng per oppgave).

For oppgavene knyttet til datasettet er det viktig at du legger ved script med all R-kode. Her kopierer du fra RStudio og limer

inn i eget felt for kode. Du får ikke poeng for selve scriptet, men det er dokumentasjon for jobben du har gjort. Hvis script

mangler, får du ikke full uttelling på noen av oppgavene selv om svaret er riktig. Ganske riktig kode kan i noen tilfeller gi delvis

poeng selv om oppgitt svar er feil. Det gis totalt 28 poeng for riktig svar.

Språk

Eksamensoppgaven er på norsk. Du kan besvare på norsk, svensk, dansk eller engelsk.

Hjelpemidler

Alle R-script som er blitt brukt i kurset.

1

Del 1: flervalgsoppgaver (publiseres ikke)

For alle flervalgsoppgaver får du 1 poeng hvis du svarer riktig, og minus 0.25 poeng hvis du svarer feil. Du får 0 poeng for oppgaver som er ubesvart. (Andre typer oppgaver gis det ikke trekk ved feil).

Del 2: kortsvarsoppgaver

Tekstoppgaver har gjerne flere elementer og du bør passe på å besvare hver del da det påvirker

poenggivningen. Svar presist og utfyllende, men ikke lengre tekst enn nødvendig.

Det gis totalt 6 poeng (3 poeng per oppgave).

2.1 Forklar hva «partial dependence plot» viser, og hvordan det skiller seg fra «variable importance».

2.2 Forklar hva OOB-feilraten (out-of-bag error) er i random forest, og hvorfor det ikke alltid er nødvendig å splitte dataene på forhånd.

Del 3: datasett oppgaver

For oppgavene knyttet til datasettet er det viktig at du legger ved script med all R-kode. Her kopierer du fra

RStudio og limer inn i eget felt for kode. Du får ikke poeng for selve scriptet, men det er dokumentasjon for

jobben du har gjort. Hvis script mangler, får du ikke full uttelling på noen av oppgavene selv om svaret er riktig.

Ganske riktig kode kan i noen tilfeller gi delvis poeng selv om oppgitt svar er feil.

Det gis totalt 28 poeng for riktig svar.

Bakgrunn:

Se for deg at en høyere utdanningsinstitusjon ønsker å identifisere studenter som er i risiko for å avbryte

studiet, med mål om å tilby dem målrettet oppfølging tidlig i studieløpet. En forutsetning for slike tiltak er at

man kan predikere hvem som er i risikozonen allerede ved studiestart, basert på opplysninger som foreligger

ved innmelding.

Kommunen og institusjonen har ikke angitt hva slags tiltak som skal settes inn. Man kan tenke seg to svært ulike tilnærminger:

1. Frafall henger sammen med faglige, sosiale eller personlige utfordringer. Tiltak vil ha som siktemål å

styrke studentenes mestringsevne, redusere risikofaktorer og generelt gi disse studentene bedre

forutsetninger for å fullføre.

2. Institusjonen må beskytte sitt omdømme og sine ressurser. Tiltak kan derfor ha som formål å kanalisere

disse studentene mot kortere og mindre ressurskrevende programmer, eller å håndtere dem særskilt for

å redusere kostnader.

Begge fremgangsmåter baserer seg på målrettet inngripen mot enkeltpersoner og forutsetter at man kan

predikere hvem som er i risikozonen. Tiltakene rettes mot de predikerte fremtidige frafallerne.

Datasettet:

Datasettet leser du inn med følgende kode i R:

```
readRDS("C:/UiOExam/Files/SOS2901_280426.rds")
```

```
SOS2901_280426
```

Variabeloversikt:

Variabel Beskrivelse

Dropout (utfall) Dropout / Enrolled

Marital status Sivilstatus (kategorier)

Application mode Søknadsmåte/-vei (kategorier)

Course Studieprogram (kategorier)

Daytime/evening attendance Dag- eller kveldsundervisning (binær)

Previous qualification (grade) Karakter fra tidligere utdanning (kontinuerlig)

Admission grade Opptakskarakter (kontinuerlig)

Displaced Studiested utenfor hjemkommune (binær)

Educational special needs Tilretteleggingsbehov (binær)

Debtor Har utestående studiegjeld (binær)

Tuition fees up to date Semesteravgift betalt (binær)

Gender Kjønn (binær)

Scholarship holder Stipendmottaker (binær)

Age at enrollment Alder ved innmelding (kontinuerlig)

Curricular units 1st sem

(approved)

Beståtte emner 1. semester

Curricular units 1st sem (grade) Snittskarakter 1. semester

Curricular units 2nd sem

(approved)

Beståtte emner 2. semester

Unemployment rate Arbeidsledighetsrate (makroøkonomisk)

GDP BNP-vekst (makroøkonomisk)

SOS2901 V26

7/13

3.1 Siden du har tatt et kurs i maskinlæring, blir du bedt om å lage et slikt verktøy for å predikere frafall. På hvilken måte vil en

strategi basert på 1 eller 2 ha betydning for hvordan du estimerer algoritmen? Ta stilling til forhold mellom feilrater som du vil

bruke til resten av oppgavene om dette datasettet. Begrunn hvert punkt.

Du behøver ikke ta stilling til type algoritme, men du kan gjerne eksemplifisere med random forest.

Words: 0/300

Maks poeng: 2

3.2 Lag et klassifikasjonstre der du gjør treet mer komplekst med mål om å øke andel riktig predikert så mye som mulig.

Lag confusion matrix med både trenings- og testdata. Kommenter resultatene.

Maks poeng: 4

3.3 Estimer en random forest med standardinnstillinger der du bruker samtlige forklaringsvariabler. Bruk kun treningsdataene i estimeringen.

1. Lag et plot av feilrater etter antall trær og kommenter resultatet.
2. Prediker og lag en confusion matrix på treningsdataene. Kommenter resultatene.

Maks poeng: 4

3.4 Lag et variable importance plot.

1. Gi en kort forklaring på hva variable importance er og hvordan det tolkes.
2. Kommenter resultatet for dette datasettet - er det noe som overrasker deg?

Maks poeng: 2

3.5 Estimer en ny random forest-modell, men bruk tuning med mål om å få tilnærmet ønsket forhold mellom feilratene.

Beskriv kort hva du har gjort, og noen av verdiene av tuningparametrene du har forsøkt. Kommenter kort i hvilken grad du

klarte å oppnå ønsket resultat. (Bruk bare treningsdata.)

Maks poeng: 4

3.6 Du skal vurdere hvor rettferdig modellen er for kjønn. Bruk bare treningsdata.

1. Bruk pakken fairmodels til å beregne fairness-mål for kjønn. Velg en rimelig referansegruppe og begrunn valget. Lag et fairness-plott og gi en vurdering av hvorvidt dette er en rettferdig modell.
2. Juster modellen med stratifisering og sett verdiene for sample size til å være lik for de to kjønnskategoriene. Hva begrenser hvilke verdier du kan sette for denne parameteren?
3. Sammenlign det nye fairness-plottet med resultatet i punkt: Ble modellen mer rettferdig?

Maks poeng: 6

3.7 Gjør en ny vurdering av confusion matrix mot **testdata** og sjekk også rettferdighet med `fairness_check()` på **testdataene**.

Kommenter på om det er vesentlige forskjeller fra resultatene på treningsdata.

Maks poeng: 2

3.8 Legg inn fullstendig script her. Scriptet skal være slik at sensor skal kunne kjøre koden og få samme resultater som deg.

Husk derfor å bruke `set.seed()` før hver prosedyre som bruker tilfeldige tall på noe vis.

Skriv inn kommentarer i scriptet på hvilken oppgave hver del er til.

Merk: Det gis 4 poeng for denne oppgaven hvis scriptet fungerer som intendert og viser resultater som i besvarelsene i deloppgavene.

Maks poeng: 4