I dag – veiledning 2/4

TIME 1
Samle noen tråder fra obligen
Gå igjennom litt teori til prosjektene.
Svare på spørsmål

TIME 2
Prosjektveiledning

Et par bokanbefalinger

- Introduction to statistical learning (2014), ny versjon kommer til sommeren. Gratis på springerlink
- Interpretable machine learning (Nettbok)

Obligen

Binære prediktorer

Gjøre til 0 v 1 heller enn til dummy-variable

XGBoost i oblig 1

Eksempel på state of the art!

Å ikke jukse

Overfitting

Hyperparameter overfitting

Train/Test/Validate-split

Særlig aktuelt i Studentkarakterer-prosjektet

Kodeceller

```
def predict(x):
    return x*x

x_1 = np.linspace(0,1,100)
my_prediction = predict(x_1)
```

Jupyter notebooks

Tekstceller (markdown). Tekstcellene kjører ingen kode, så der kan du trygt skrive hva som helst!

```
# Sigmoid-funksjonen

Sigmoiden er definert som

$$ e^{x}/(e^{x}+1) $$
```

→

Sigmoid-funksjonen

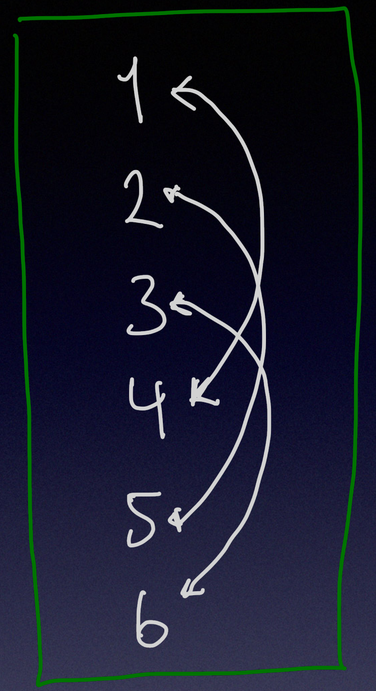
Sigmoiden er definert som

$$e^x/(e^x+1)$$

Prosjektene

Dere har jobbet med å se på datasettene

Underveisoppgave



1. Hvis din gruppe er N, finn gruppe (N+3)%7+1. Dette er gruppen dere skal samarbeide med. *1 minutt*

2. Én person fra hver gruppe skal snakke med én person fra hver av de andre gruppene. Kontakte hverandre enten i levende live, eller over én-til-én digital plattform. *1 minutt. Noen grupper går ikke opp. Forklare da 2–1.*

3. Forklare hvordan datasettet ditt ser ut. Hva det inneholder, og hva det har blitt brukt til tidligere. Først forklarer den som jobber med COMPAS, så forklarer den som jobber med studentkarakterer. *3 + 3 = 6 minutter*

COMPAS

En øvelse i betingede sannsynligheter

Betinget sannsynlighet

Matchende fingeravtrykk i drapssak. La oss sette beviskravet til 95 % sjanse for at tiltalte er skyldig for å kunne dømme.

Anta avisen fokuserer på de følgende (sanne) påstandene:
P(match | uskyldig person) = 1/1 000 000
P(match | skyldig person) = 999/1000

Kan de være fornøyde med jobben, og er fingeravtrykket fellende bevis?

Betinget sannsynlighet forts.

P(match | tilfeldig uskyldig person) = 1/1 000 000
P(match | skyldig person) = 99.9 %

Anta Oslo: 5 millioner innbyggere i en alternativ virkelighet der politiet har alles fingeravtrykk.

Hva er sannsynligheten for at tiltalte er skyldig?

Betinget sannsynlighet forts.

P(skyldig | match) ?

Vel, det er 4.9999990 uskyldige som kommer til å matche, og dessuten kommer den skyldige til å matche med 99.9% sannsynlighet. Dermed har vi 5.998999 stykker som matcher og blant 0.999 skyldig (Det er en liten sjanse for at den skyldige gjemmer seg blant de andre).

P(skyldig | match) = 0.999/5.998999 = 0.16653..

Bayes' setning

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

$$P(\text{skyldig} \mid \text{match}) = \frac{P(\text{match}|\text{skyldig}) \cdot P(\text{skyldig})}{P(\text{match})}$$

For vårt eksempeltrenger vi
P(match) = 5.998999/5 000 000, siden 4.999999 falske positiv og 0.999 ekte positiv beregnet på befolkningen.
P(skyldig) = 1 / 5 000 000 siden én er skyldig
P(match|skyldig) = 999/1000 som avisen skriver

Bayes' setning

$$P(\text{skyldig} \mid \text{match}) = \frac{P(\text{match}|\text{skyldig}) \cdot P(\text{skyldig})}{P(\text{match})}$$

= 0.999*(1/5000000)/(5.998999/5000000)

= 0.16653..

Studentkarakterer

Modellsammenligning

XGBoost i studentkarakterer

Mer avansert bruk enn i oblig 1: Modellsammenligning

(Vi kommer tilbake til mer om dette neste uke)

Gradient boosted decision trees