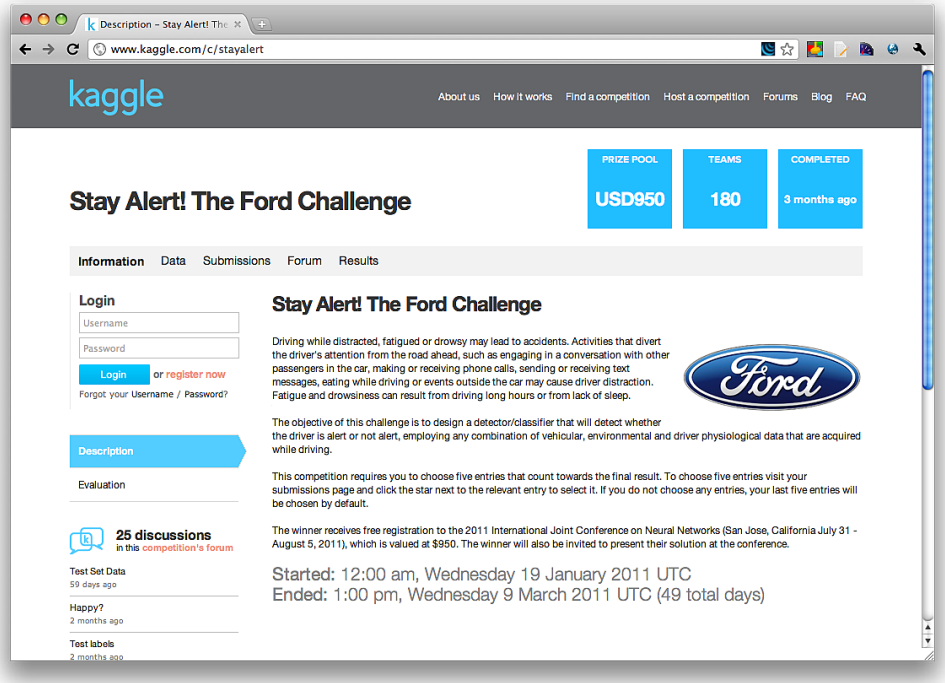
Detection of human alertness using supervised learning

Projektet: **The Ford Challenge (TFC)**

Machine Learning konkurrence på hjemmesiden kaggle.com

Opgaven: **Lav et program der kan forudsige om en bilist er opmærksom eller ej.**

Udgangspunktet er et datasæt med målinger af bilister, som Ford har udarbejdet.



Datasættene til TFC

`fordTrain.csv`

- 500 trials
- 604 329 rækker
- `IsAlert` feature inkluderet

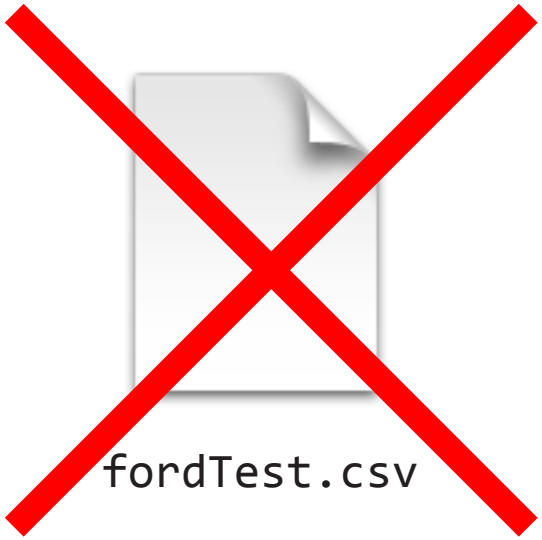
`fordTest.csv`

- 100 trials
- 120 840 rækker
- `IsAlert` feature ikke angivet

Datasættene til TFC

`fordTrain.csv`

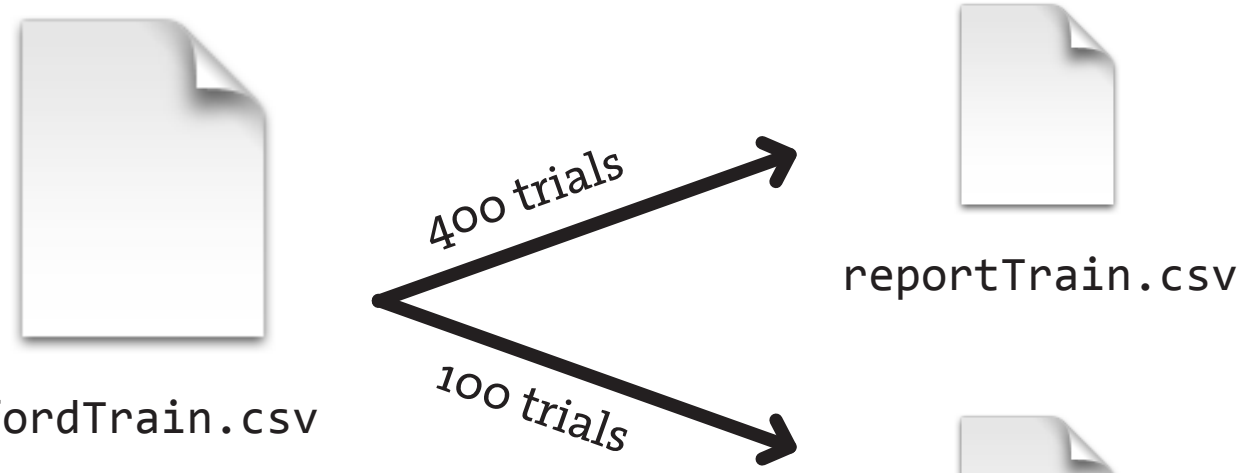
- 500 trials
- 604 329 rækker
- `IsAlert` feature inkluderet



`fordTest.csv`

- 100 trials
- 120 840 rækker
- `IsAlert` feature ikke angivet

Datasættene til TFC



fordTrain.csv

- 500 trials
- 604 329 rækker
- `IsAlert` feature inkluderet

reportTrain.csv

reportTest.csv

Datasættenes opbygning?

TrialID	ObsNum	IsAlert	P1	…	P8	E1	…	E11	V1	…	V11
0	0	0	12.2	…	1.2	4.3	…	33	12	…	7.34
⋮											⋮
0	1200	1	11.1	…	10.7	1.3	…	21	8	…	8.82
⋮											⋮
510	1198	0	11.1	…	10.7	1.3	…	21	8	…	8.82

Hver trial er ca. 2 minutter.
Målinger taget med 0.1 sekunds intervaller.
Giver ca. 1200 målinger pr. trial.

Består af 3 feature-kategorier:
- Fysiologi (P)
- Miljø (E)
- Bil (V).

Intet andet er oplyst om features!

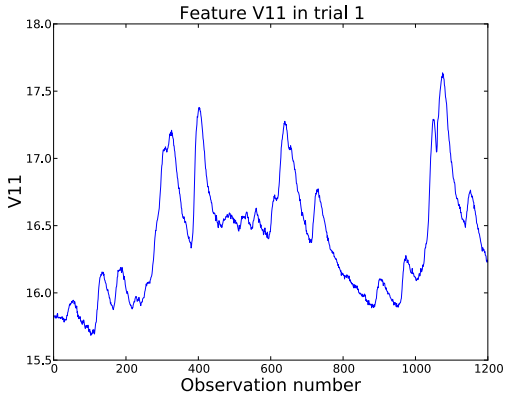
Next step: **Dataundersøgelse**

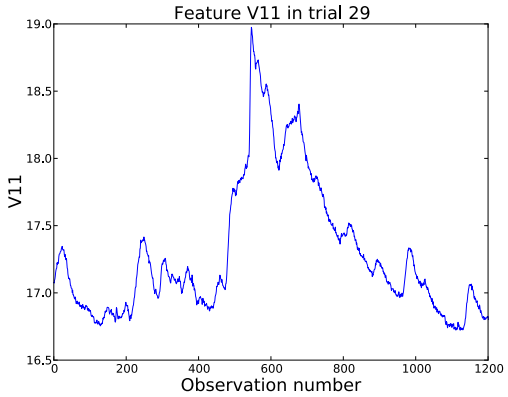
Ekstra vigtigt da intet vides om de enkelte features!

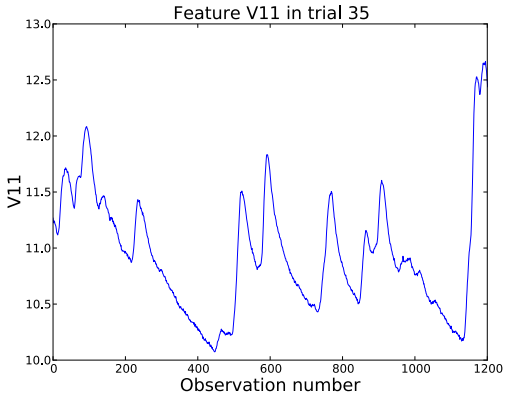
Køreplan:

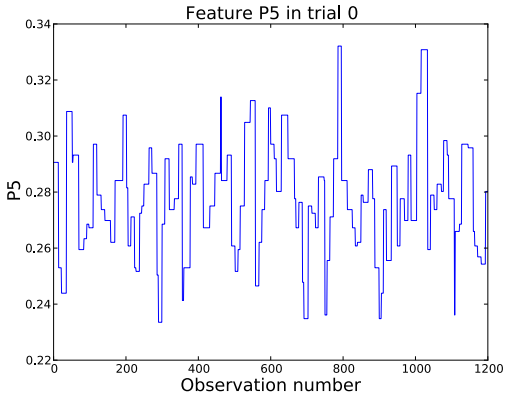
- Udregn middelværdi, standard afvigelse, min og max for features
- Udregn antal unikke værdier for features i de enkelte trials.
- **Plot features for de enkelte trials**
- Scatterplots
- **Outlierhåndtering**

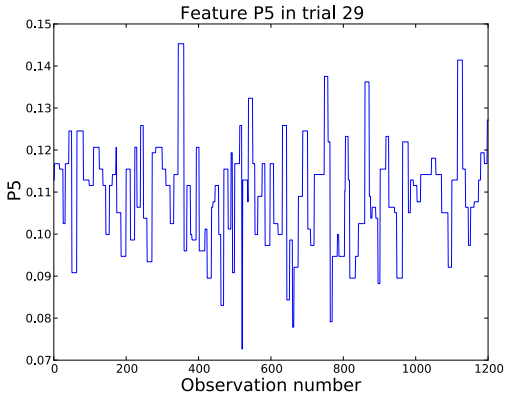
Plot af de enkelte features: **V11, P5**

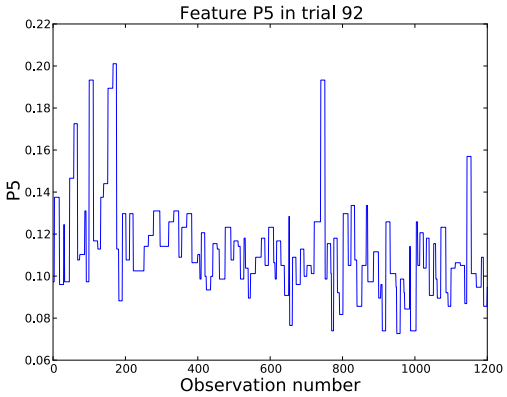




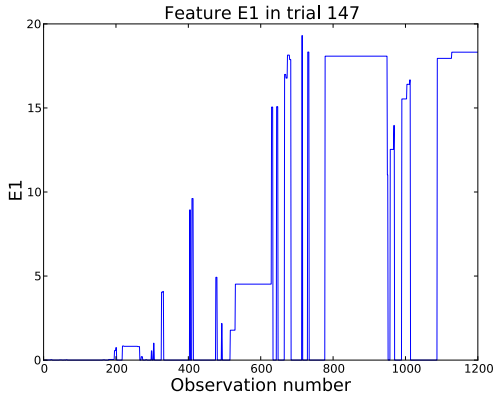


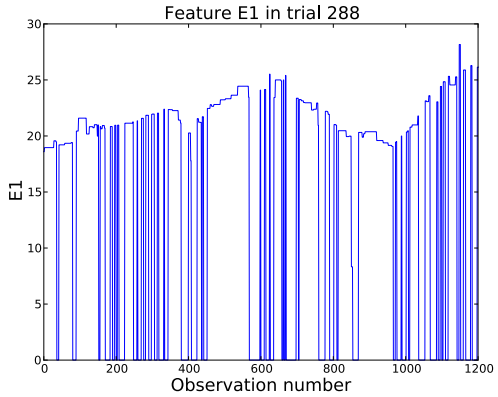


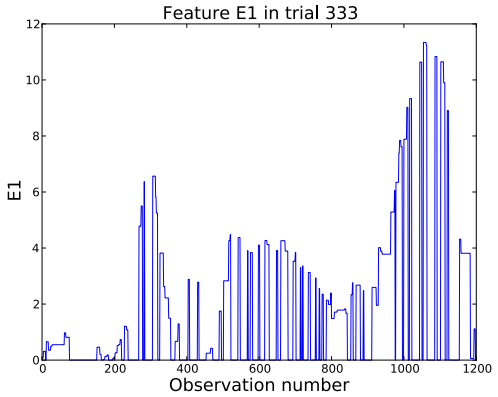


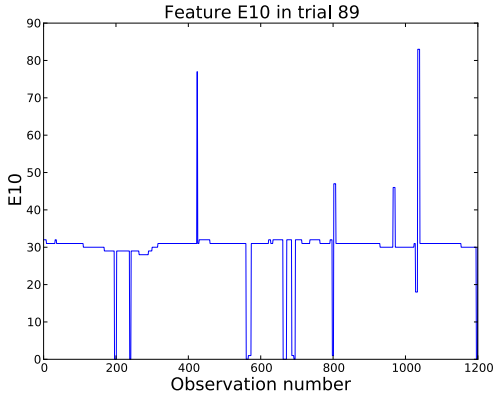


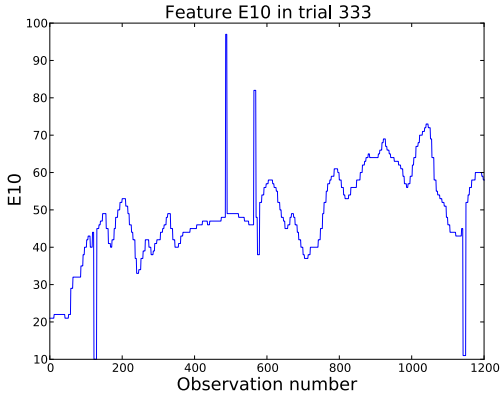
Plot af de enkelte features: **E1, E10**

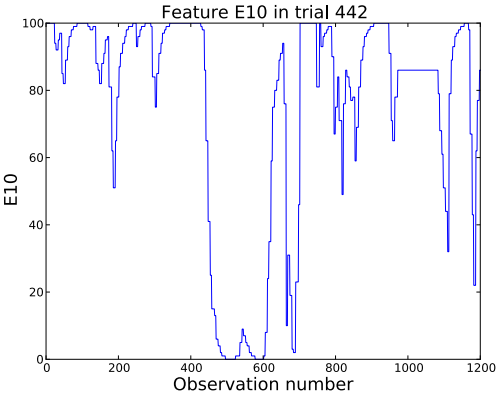




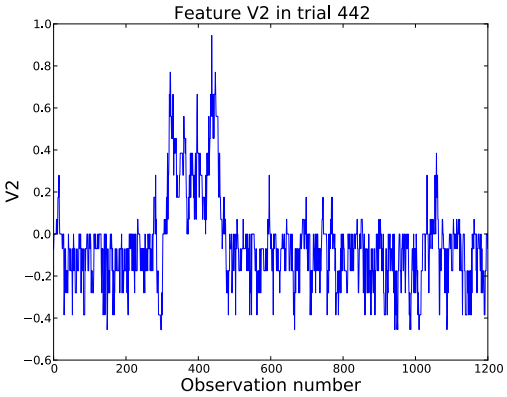


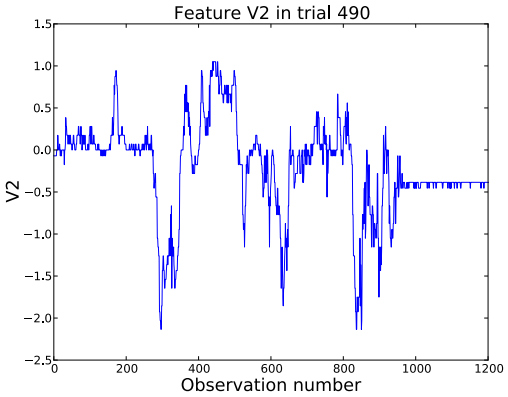


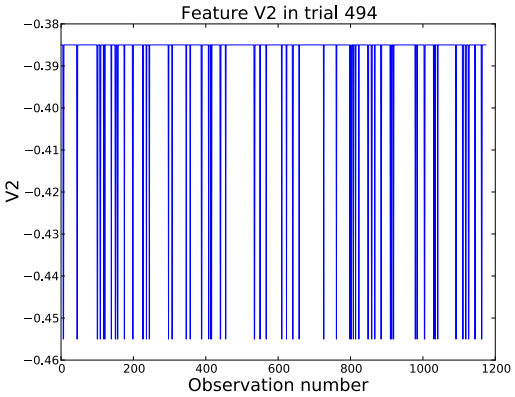




Plot af de enkelte features: **V2**





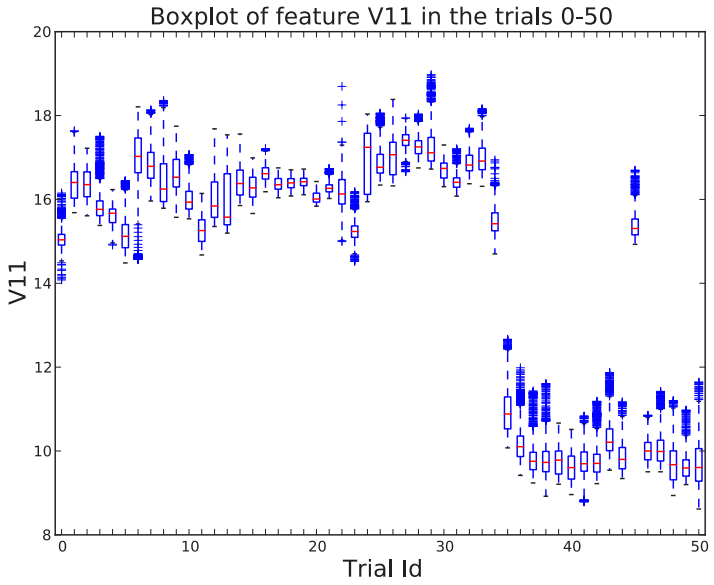


Konklusion: **Det er data fra den virkelige verden.**

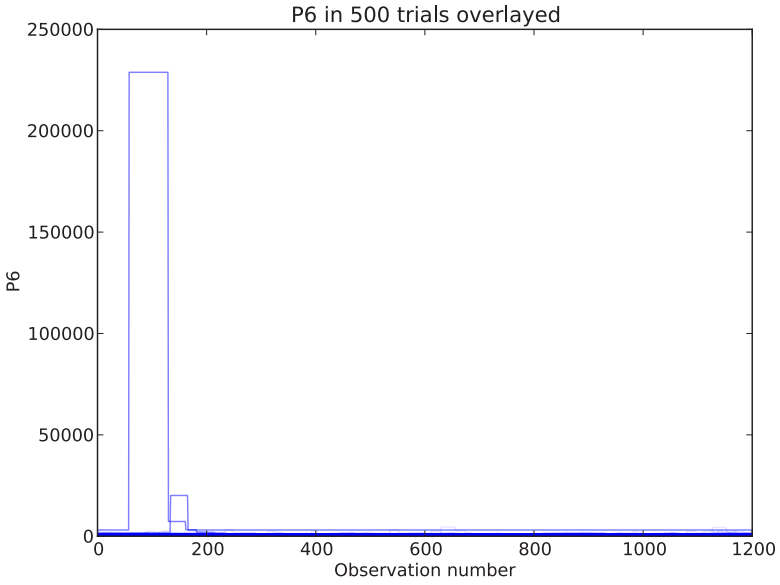
Spørgsmål: **Hvordan skal outliers håndteres?**

Outliers: **Definition?**

På dataraekke-niveau?

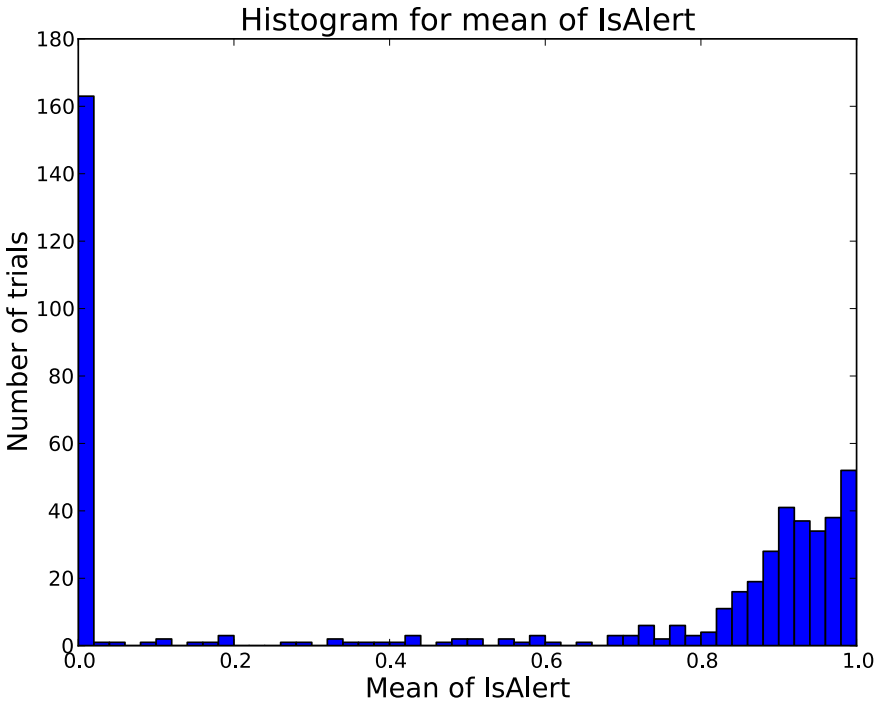


På trial-niveau?



Outliers: **IsAlert=0?**

Hvad med trials hvor IsAlert er konstant lig med 0?



Outliers: **Konklusion**

Ingen data fjernes

Next step: **Modellering**

Tid til det, det hele handler om.

Køreplan:

- Forstå teorien.
- Efterligne vindermetoden
- **Forbedre vindermetoden**
- Neurale Netværk

Teori: **Binær klassifikation**

Bestem en klassifikationsregel $f : \mathbb{R}^d \to \{0, 1\}$ der givet en d-dimensional inputvektor $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T$ kan prediktere en binær outputvariabel $t \in \{0, 1\}$ så fremtidige fejl bliver minimeret.

To spørgsmål

- Usikkerhed i prediktering?
- Hvad menes der præcist med minimering af fremtidige fejl?

Teori: **Modellering af usikkerhed**

Det antages at sandsynligheden $P(t = 1|\mathbf{x})$, har en bestemt parametrisk form. Dvs. $P(t = 1|\mathbf{x})$, som funktion af **x**, kan bestemmes ved at finde en k-dimensionel parameter **w**.

For et givet træningsdatasæt $\{(\mathbf{x}_i, t_i)\}_{i=1}^{n}$, bestemmes **w**, ved at maksimere sandsynligheden for at få træningsdatasættet. Denne metode kaldes maksimum likelihood.

For binær klassifikation fås likelihood funktionen

$$\mathcal{L}(\mathbf{w}) = \prod_{i=1}^{n} p_i^{t_i} (1 - p_i)^{1-t_i}$$

hvor $p_i = P(t_i = 1|\mathbf{x_i})$.

Teori: **Modellering af usikkerhed**

Når først sandsynligheden $P(t = 1|\mathbf{x})$, er fundet kan en klassifikationsregel laves ved

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad P(t = 1|\mathbf{x}) > P(t = 0|\mathbf{x}) \\ 0 & \text{else} \end{cases}$$

Eksempel: **Logistisk regression**

I logistisk regression antages følgende parametriske form

$$P(t = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x} + w_0)}}$$

Virker måske lidt umotiveret, men det er ensbetydende med

$$\log \frac{P(t = 1|\mathbf{x})}{P(t = 0|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + w_0$$

Bemærk at for inputvektorer hvor $P(t = 1|\mathbf{x}) = P(t = 0|\mathbf{x})$, gælder at

$$\mathbf{w}^T\mathbf{x} + w_0 = 0$$

Beregning af fejl på datasæt

I denne opgave benyttes AUC-score. Tager udgangspunkt i størrelserne

- *True positives* (TP) – Antal positive inputs, korrekt klassificeret
- *False negatives* (FN) – Antal positive inputs, forkert klassificeret
- *False positives* (FP) – Antal negative inputs, forkert klassificeret
- *True negatives* (TN) – Antal negative inputs, korrekt klassificeret

Som giver de to størrelser

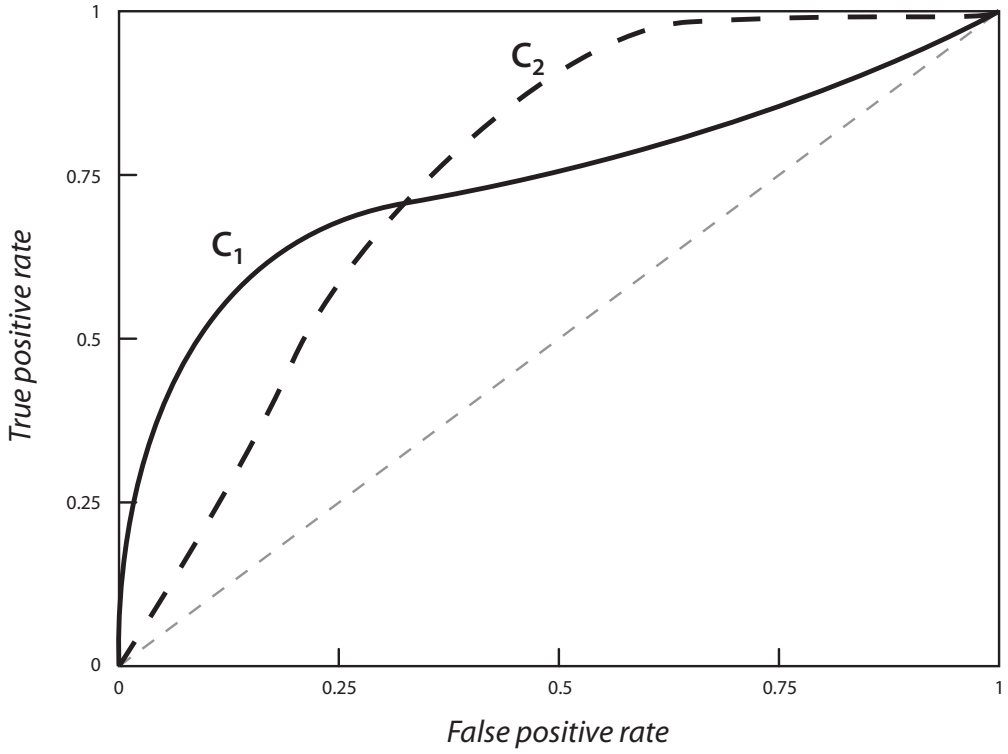
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

AUC-score

Variér grænse for hvad der predikteres som positiv klasse.

For hver grænse udregnes TPR og FPR, og punktet indsættes på grafen. Den fremkomne graf kaldes ROC-kurven.

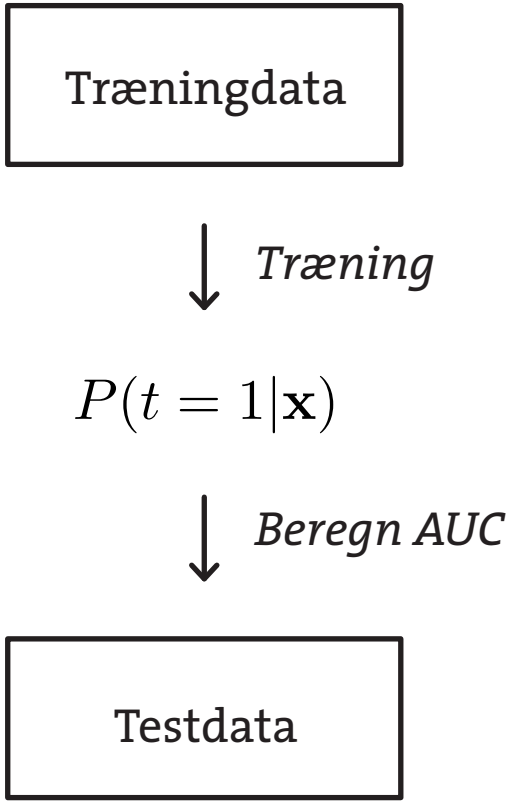
AUC er defineret som arealet under ROC-kurven



Estimér den forventede AUC

Sandsynlighedsfordelingen $P(t = 1|\mathbf{x})$, er fundet ved maksimum likelihood på træningsdatasættet. Hvordan estimeres den forventede AUC-score på fremtidige datasæt?

Beregn AUC på separat testdatasæt. Hvis testdatasæt er stort, opdeles det i mindre dele, og AUC beregnes på hver del. Derved kan konfidensinterval beregnes.



Vindermodellen

Vinderen benyttede en logistisk regression på de tre features sdE5, V11 og E9

$$\log \frac{P(t=1|\mathbf{x})}{P(t=0|\mathbf{x})} = -392.4317 \cdot \mathrm{sdE5} + 0.2209 \cdot \mathrm{V11} + 3.6544 \cdot \mathrm{E9}$$

Featuren sdE5, er den løbende standardafvigelse af E5, indenfor hver enkel trial. Dvs. for den *j*'te observation i trial *k* er sdE5, lig med standardafvigelsen for E5 over observationerne 1,2,...,*j* i trial *k*

AUC-score på ***Ford testsættet*** = 0.8492

AUC-score på ***Ford træningssættet*** = [0.81, 0.83]

Efterlign vindermodel

Ved maksimum likelihood på rapport træningssættet findes modellen

$$\log \frac{P(t=1|\mathbf{x})}{P(t=0|\mathbf{x})} = -87.2753 \cdot \mathrm{sdE5} + 0.2244 \cdot \mathrm{V11} + 3.6545 \cdot \mathrm{E9} - 5.3645$$

Bemærk at intercept er med. Kan forklare forskellen i sdE5-parametren.

95%-KI for AUC-score på ***rapport træningssættet*** = [0.8058, 0.8107]

Forbedring af vindermodel

Benytter *forward feature selection*. Tilføj features til modellen så længe AUC-score forbedres.

De tre bedste features fundet: sdE1, V11 og E9.
Dvs. vindermodellens features hvor sdE5 udskiftes med sdE1

Gav modellen

$$\log \frac{P(t=1|\mathbf{x})}{P(t=0|\mathbf{x})} = -0.0988 \cdot \mathrm{sdE1} + 0.2019 \cdot \mathrm{V11} + 3.6418 \cdot \mathrm{E9} - 4.2076$$

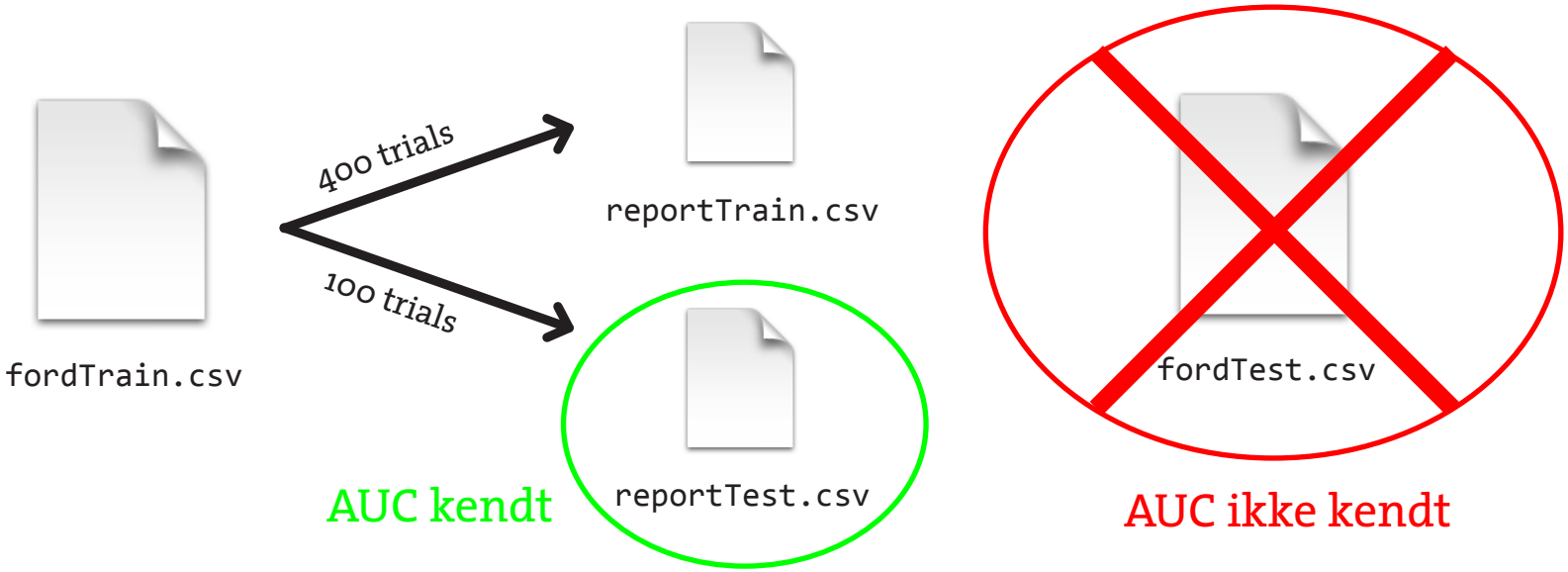
95%-KI for AUC-score på ***rapport træningssættet*** = [0.8278, 0.8358]

Forbedring?

Forbedring opnået med et lille forbehold.

Forbedring opnået på Ford træningssættet. AUC-scoren på Ford testsættet kendes ikke.

Der blev rapporteret om store forskelle mellem AUC-score på Ford træningssættet og Ford testsættet.

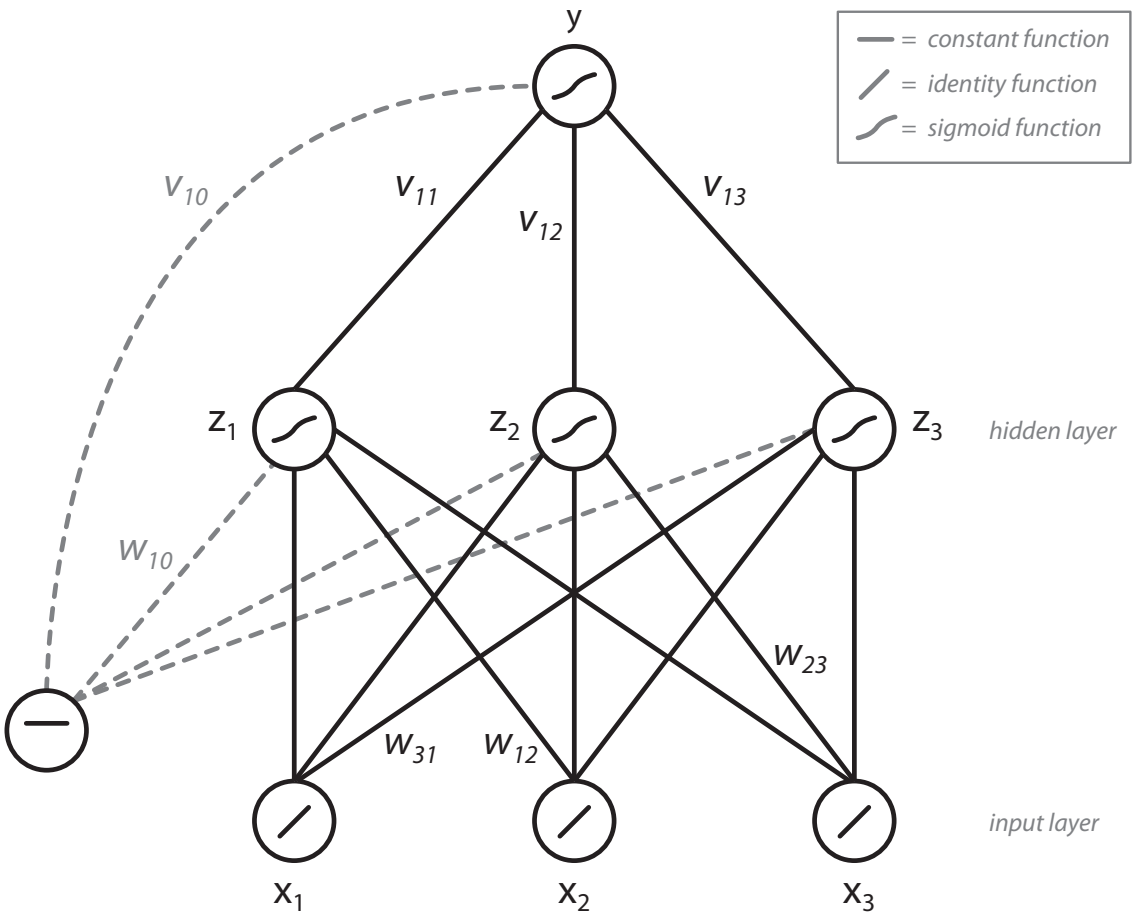


Neuralt Netværk

Forsøgte at forbedre den logistiske model med neuralt netværk.

Umiddelbart ingen forbedringer.

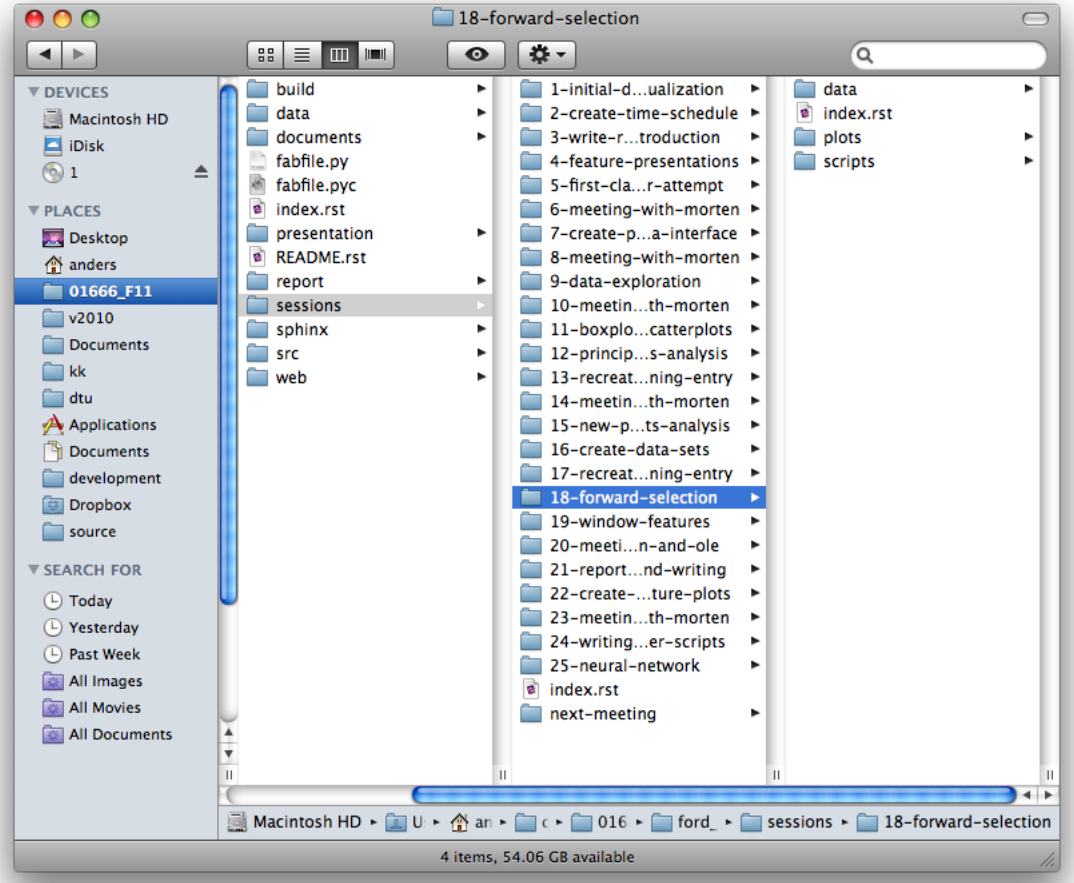
Skyldes evt. manglende computerkraft.



En lille idé: **Strukturering af filer**

Problem: Hvor skal plots, data og dokumentation placeres, så det kan findes når rapporten skal skrives.

Indførte konceptet "session". En blok arbejde på projektet er en session.



En lille idé: **Strukturering af filer**