Dagens gruppetime

- Oblig 2b
- Naive Bayes

Oblig 2b

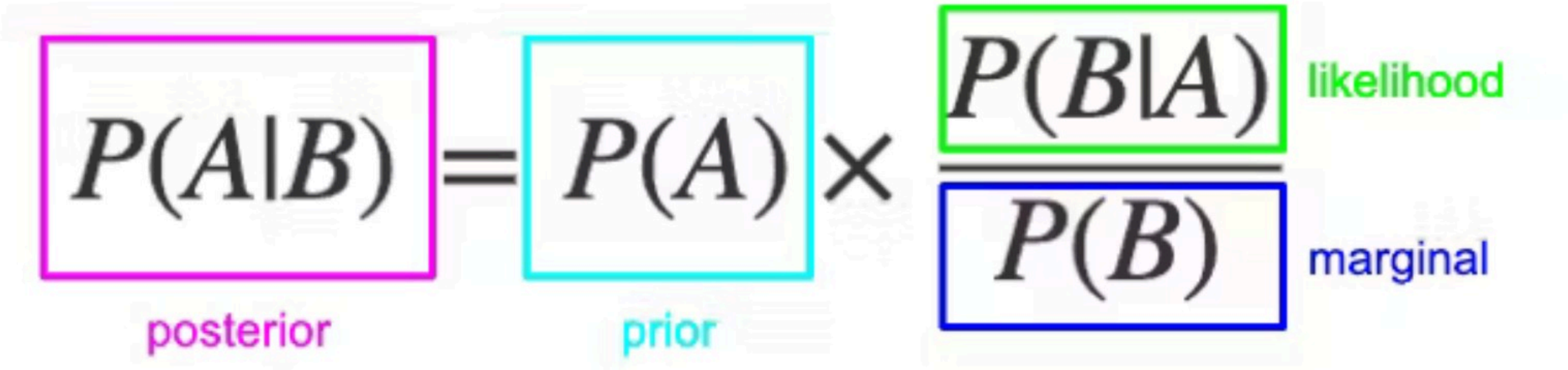
Siste oblig, yay!

Naive Bayes

→ Probilalistisk => basert på sannsynligheter

Bayes' teorem

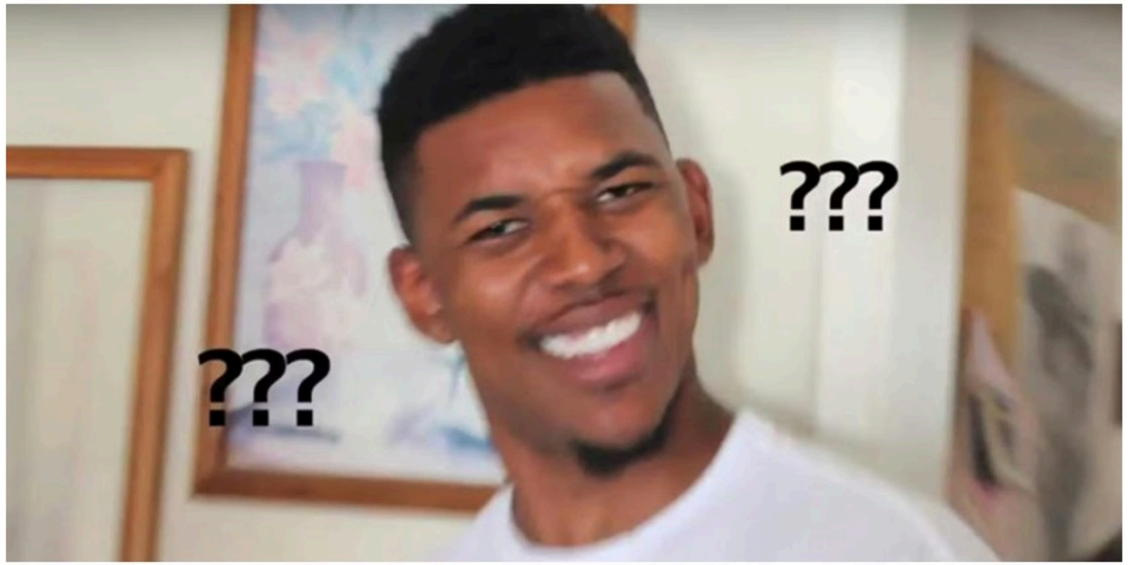
- Sannsynlighetsteori
- Uavhengighet



Naiv?

- UAVHENGIGHET
- Alle trekk er uavhengige, gitt klasse
- BoW:
- P(<super,dårlig>|c) = P(super|c)*P(dårlig|c)

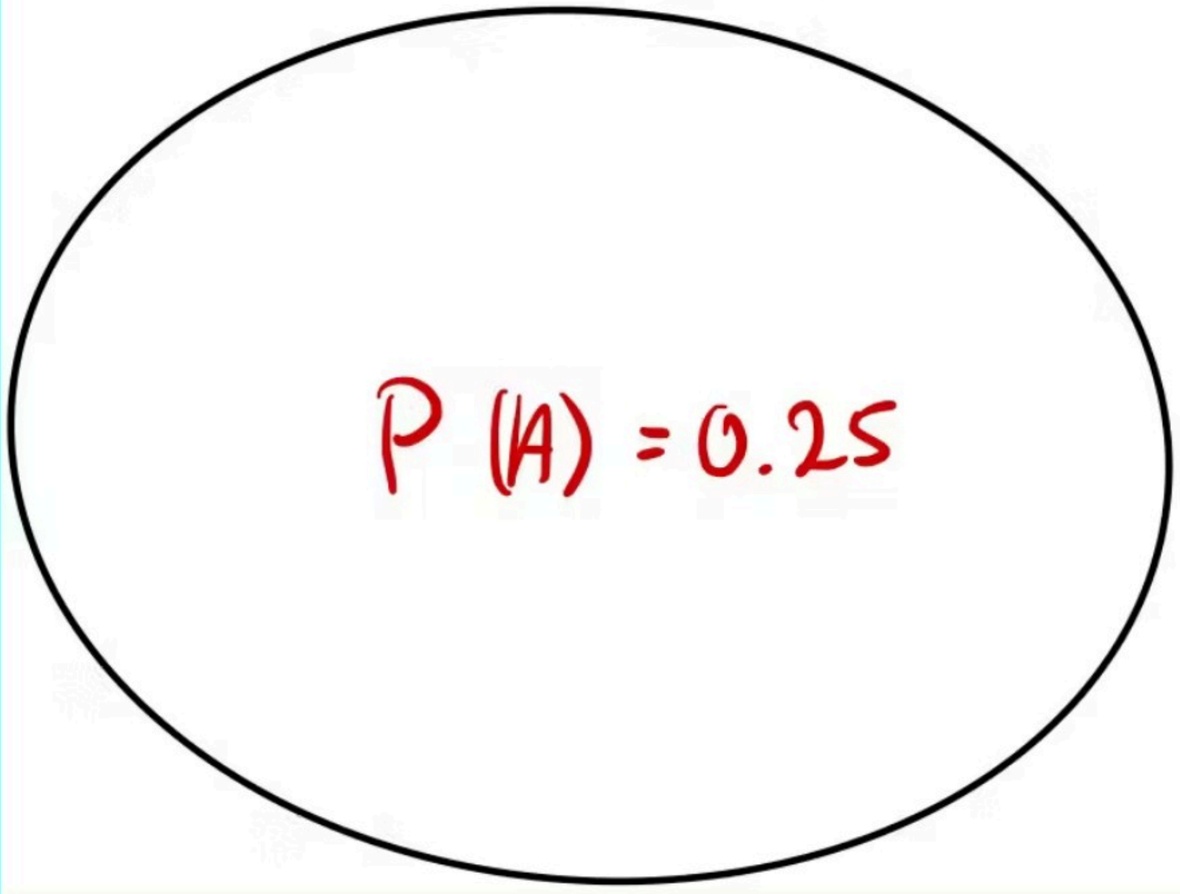
UAVHENGIGHET?

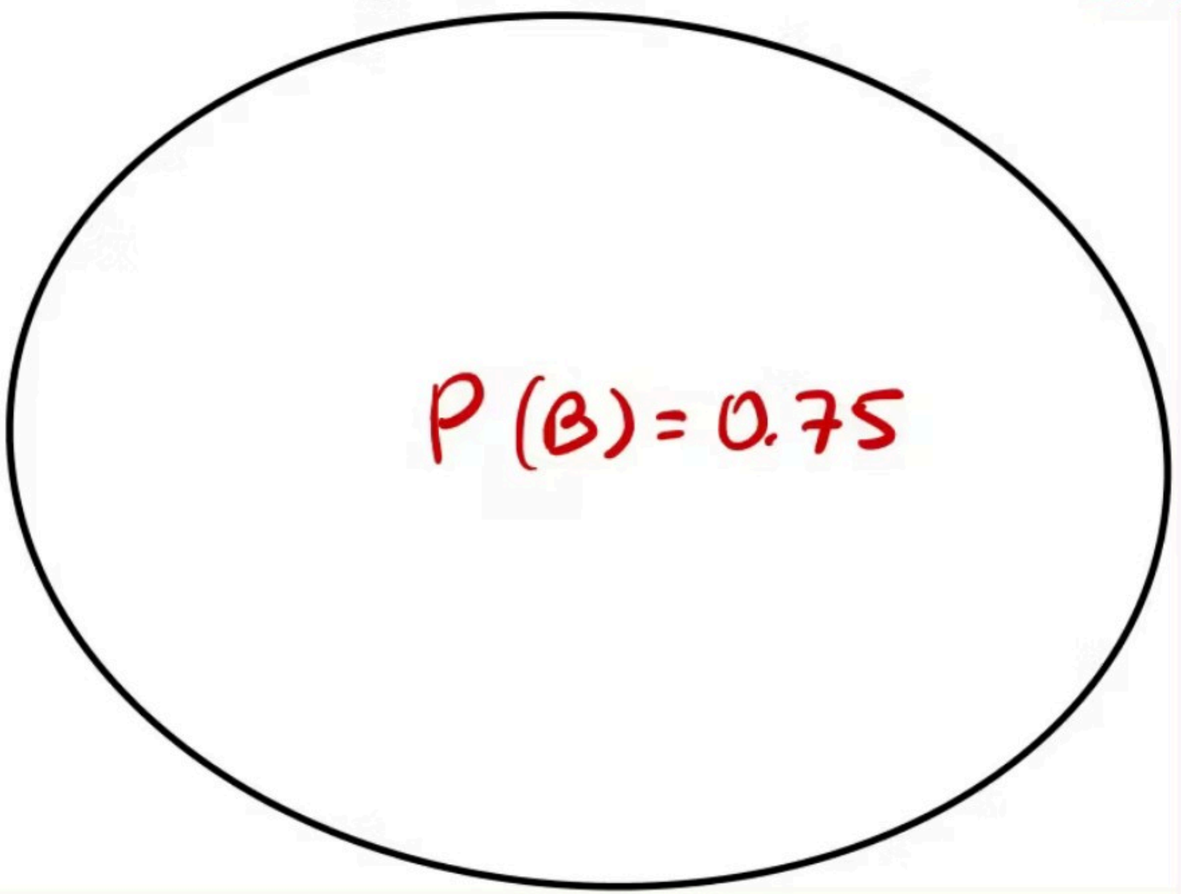


Utfallsrommet

Hendelse A

Hendelse B

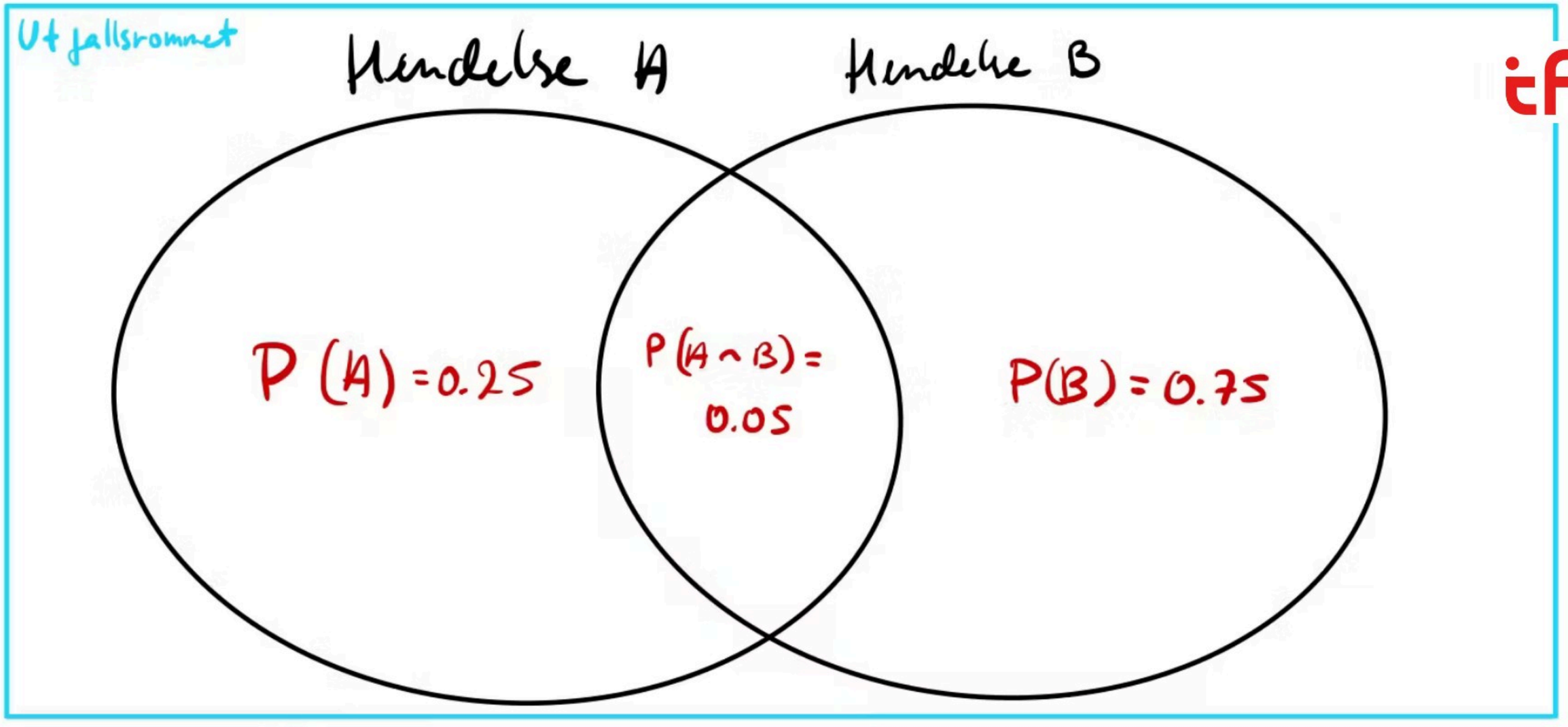




Disse sannsynlighetene er uavhengige.

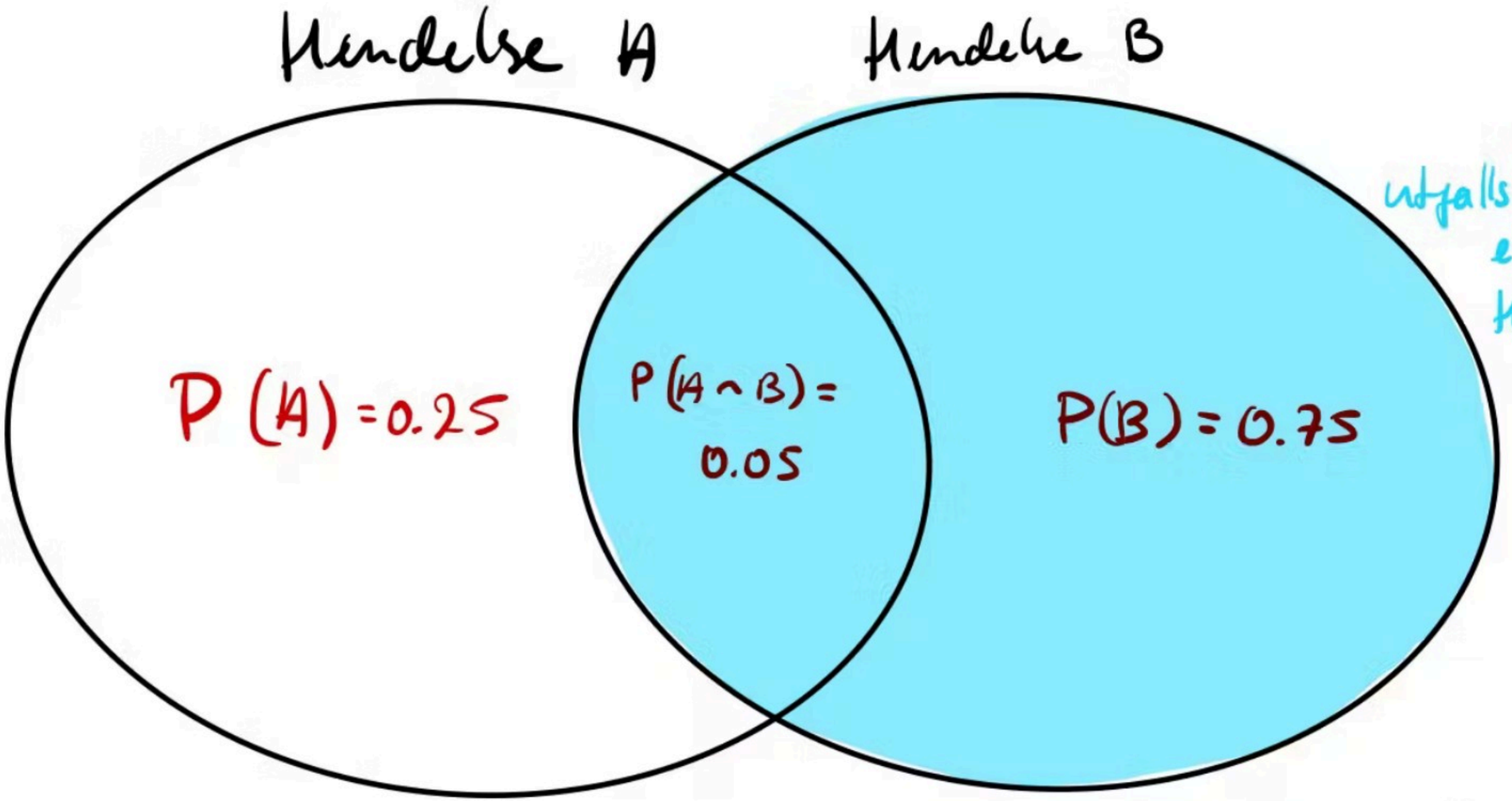
Dersom Hendelse B har inntruffet er sannsynligheten for at Hendelse A inntreffer fremdeles 0.25.

→ $P(A|B) = P(A) = 0.25$



Disse sannsynlighetene er ikke uavhengige.

Dersom Hendelse B inntreffer vil utfallsrommet bli mindre...



Sannsynligheten for Hendelse A blir dermed:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.75} = 0.066\dot{} \approx 0.067$$

Hva vil dette si for oss?

Når vi snakker om uavhengige "trekk" (som med BoW i handler om ulike ord), vil dette si at vi ser på ordene separat uten kontekst.

OBS! Dette var bare en kort inføring/introduksjon til konseptet om uavhengighet, gjerne se litt mer på det selv.

Naive Bayes klassifisering

- Finne mest sannsynlige klasse $\hat{c}$, gitt et dokument d

Naive Bayes klassifisering

- Finne mest sannsynlige klasse $\hat{c}$, gitt et dokument d

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

Naive Bayes klassifisering

- Finne mest sannsynlige klasse $\hat{c}$, gitt et dokument d

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

↓

Regner ut $P(c|d)$ for alle $c \in C$

Naive Bayes klassifisering

$\in$ er et symbol som sier at et element er i en mengde.
Eks. $x \in A$, elementet x finnes i mengden A.

- Finne mest sannsynlige klasse $\hat{c}$, gitt et dokument d

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

↓

Regner ut $P(c|d)$ for alle $c \in C$

Eks. ved semantisk analyse (Oblig 2b) vil $C = \{NEG, POS\}$

Altså må vi regne ut $P(c|d)$ hvor $c = NEG$ og $P(c|d)$ hvor $c = POS$

Naive Bayes klassifisering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

Hvordan finner vi denne?

Naive Bayes klassificering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

Bayes' regel:

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; \frac{P(d|c) \cdot P(c)}{P(d)}$$

Naive Bayes klassifisering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$$

Bayes' regel:

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; \frac{P(d|c) \cdot P(c)}{\cancel{P(d)}}$$

Konstant for alle $c \in C$!
Det er c som varierer, ikke d...

Naive Bayes klassifisering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c) \cdot P(c)$$

hvordan er d definert?

Naive Bayes klassifisering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c) \cdot P(c)$$

$\rightarrow d = \langle w_1, w_2, \ldots, w_k \rangle$

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(w_1, w_2, \ldots, w_k|c) \cdot P(c)$$

Naive Bayes klassifisering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c) \cdot P(c)$$

$\rightarrow d = \langle w_1, w_2, \ldots, w_k \rangle$

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; \boxed{P(w_1, w_2, \ldots, w_k|c)} \cdot P(c)$$

Naive Bayes klassificering

"Naiv" antagelse:

$$P(d|c) = P(w_1, w_2, \ldots, w_k | c)$$

$$\approx P(w_1|c) \cdot P(w_2|c) \cdot \ldots \cdot P(w_k|c) = \prod_{i=1}^{k} P(w_i|c)$$

Naive Bayes klassifisering

"Naiv" antagelse:

$$P(d|c) = P(w_1, w_2, \ldots, w_k | c)$$

$$\approx P(w_1|c) \cdot P(w_2|c) \cdot \ldots \cdot P(w_k|c) = \prod_{i=1}^{k} P(w_i|c)$$

hvordan regner vi dette?

Naive Bayes klassifisering

$$P(w_i \mid c) = \frac{\text{frekvens av } w_i \text{ i } c}{\text{frekvens av alle } V \text{ i } c}$$

MLE!

(Maximum likelihood Estimation, som vi er godt kjent med 🙂)

Naive Bayes klassifisering

$$P(w_i \mid c) = \frac{\text{frekvens av } w_i \text{ i } c}{\text{frekvens av alle } V \text{ i } c}$$

→ Hva om w_i ikke eksisterer i c ???

Naive Bayes klassifisering

$$P(w_i \mid c) = \frac{\text{frekvens av } w_i \text{ i } c + 1}{\text{frekvens av alle } V \text{ i } c + |V|}$$

→ Hva om w_i ikke eksisterer i c ???

Glatting :)

Hovedantagelse → $\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(c|d)$

↓↓ Bayes' regel

$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(d|c) \cdot P(c)$

↓↓ Definisjon av d

$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \; P(w_1, w_2, \ldots, w_k|c) \cdot P(c)$

↓↓ "Naiv" antagelse

$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \prod_{i=1}^{k} P(w_i|c) \cdot P(c)$ → Naive Bayes

Naive Bayes klassifisering

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \prod_{i=1}^{k} P(w_i \mid c) \cdot P(c)$$

$$P(c) = \frac{\text{antall dokumenter i treningsdataen som er i } c}{\text{totalt antall dokumenter}}$$

Denne vil være forskjellig for alle $c \in C$, så derfor må den være med.