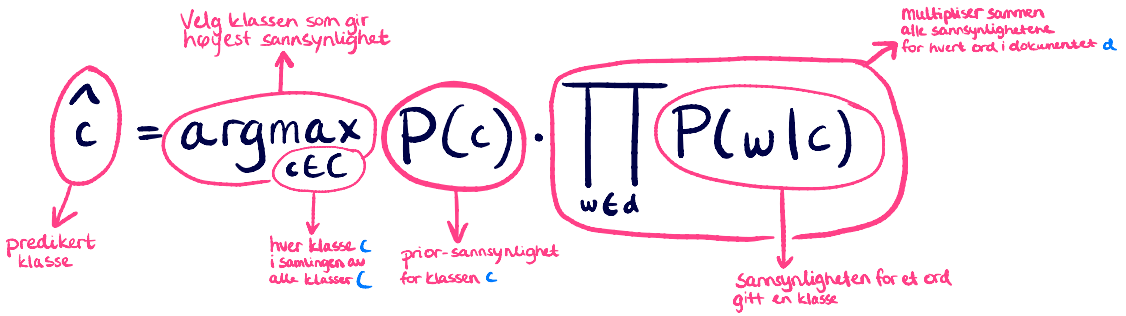
Anika Willoch Olstad
annkaol@uio.no

Naive Bayes



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

$$P(c) = \frac{N_c}{N_{dok}}$$ → antall dokumenter med gitt klasse; → totalt antall dokumenter i treningsdataen

$$P(w_i|c) = \frac{count(w_i, c)}{\left(\sum_{w \in V} count(w, c)\right)}$$ → antall ganger ordet forekommer i klassen c; → antall tokens i klassen

Hovedantagelse

Formelen over er basert på hovedantagelsen:

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

——— det vil si ⟶

Vi kan predikere korrekt klasse ved å velge den som har høyest sannsynlighet for et gitt dokument

Utledning

- fra hovedantagelsen til Naive Bayes

1) Hovedantagelsen

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

2) Bayes teorem
- betinget sannsynlighet + produktsetningen

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

3) Fjern $P(d)$

Siden $P(d)$ er uavhengig av k og vil være konstant for hver d, kan den fjernes

$$P(c|d) = \frac{P(d|c)P(c)}{\cancel{P(d)}} = P(d|c)P(c)$$

4) (Naiv) uavhengighetsantagelse

Antar (naivt) at alle trekk er uavhengige av andre trekk, gitt klassen.
→ da kan vi bruke multiplikasjonsregelen for uavhengige hendelser.

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

erstatt

$$P(c|d) = P(d|c)P(c)$$

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

5) Fullfør

Setter sammen til ferdig formel

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