Gruppetime #9

IN1140, gruppe 1

Komponentene i Naive Bayes-formelen

Naive Bayes-formelen:
(slik den står i oblig 2b)

$$\hat{k} = \underset{k \in K}{\operatorname{argmax}} \; P(k) \prod_{j=1}^{n} P(v_j|k)$$

$\hat{k}$ *Predikert klasse*

$\underset{k \in K}{\operatorname{argmax}}$ *Returnerer klassen til den høyeste sannsynligheten*

$P(k)$ *Prior-sannsynligheten for en klasse*

$P(v_j|k)$ *Sannsynligheten for et ord gitt en klasse*

Komponentene i Naive Bayes-formelen

Hvordan finne priorsannsynlighet for klasse og sannsynlighet for ord gitt klasse?

(Videre bruker vi c i stedet for k, og w i stedet for v.)

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

Predikert prior-sannsynlighet for gitt klasse

Teller: antall dokumenter med gitt klasse
Nevner: antall dokumenter totalt i treningsdataen

$$\hat{P}(w_i|c) = \frac{count(w_i, c)}{\left(\sum_{w \in V} count(w, c)\right)}$$

Predikert sannsynlighet for et ord gitt en klasse

Teller: antall ganger ordet forekommer i klassen
*Nevner: antall **token** i klassen*

Glatting

Laplace, "legg til én-glatting"

Brukes for å unngå at resulterende sannsynlighet blir 0 fordi et ord ikke forekommer i en gitt klasse i treningsdataen.

Vi endrer formelen for predikert sannsynlighet for et ord gitt en klasse:

$$\hat{P}(w_i|c) = \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c)\right) + |V|}$$

Teller: *legg til én*

Nevner: *legg til størrelsen på ordforrådet, det vil si antall* ***ordtyper*** *i all treningsdataen*

Dersom et ord i testdataen ikke forekommer i treningsdataen i det hele tatt, ignoreres det helt. *(Jurafsky & Martin, side 60 om "unknown words")*