Tehisintellekti ja masinõppe alused: "Naiivne" Bayesi klassifikaator

Liine Kasak

April 10, 2019

1 Lahendus

Rämpsposti klassifitseerimisel rakendati naiivset Bayesi teooriat. Teooria on naiivne, sest eeldatakse, et sündmused (antud juhul sõnad) on sõltumatud.

Üldine tõenäosus arvutatakse valemiga

$$H_c = P(c) \cdot P(w_1|c) \cdot P(w_2|c) \cdot ... \cdot P(w_n|c),$$

kus c on klassifikaator (rämps või mitterämps) ja w_x on sõna kohal x.

Sõnade esinemise individuaalsed tõenäosused arvutatakse:

$$P(w|c) = \frac{N_{w,c} + 1}{N_c + |V|},$$

kus $N_{w,c}$ on sõna w esinemiste kord c tüüpi kirjades, N_c on sõnade arv c tüüpi kirjades ja $|V|$ on unikaalsete sõnade arv kokku.

Kuna tõenäosus läheneb nullile nii rämps- kui ka mitterämpsposti puhul, siis logaritmime tõnäosusi. Tõenäosuse suurusi saab jätkuvalt võrrelda ning sõnade individuaalsete tõenäosuste asemel liidetakse tõenäosuste logaritme.

2 Tulemus

Näitekirjadest tuvastati esimene kui mitterämpspost ja teine kui rämpspost.