Kodutöö 7, Naiivne Bayes

Urmas Pitsi, 16.nov.2019

Käesolevas töös rakendame Naiivse Bayesi meetodit rämpsposti kindlaksmääramisel. Sisendandmeteks on Enroni e-kirjavahetus (http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html), mis tuleks klassifitseerida kahte klassi: mitte-rämpspost ja rämps-post (ham ja spam).

Vastused ülesannetele: Kiri 1 : ei ole rämpspost ja Kiri 2 on rämpspost.

Allpool on tabelid milles vasakul on klassifikatsiooni raport ja parema 'confusion matrix'. Need aitavad meil hinnata, kui hästi meie mudel klassifitseerib. Võib järeldada, et mudel on suhteliselt üle-kohandunud treeningandmetele, mille korral on tulemused suhteliselt head võrreldes ennustustega andmetel, mida mudel varem näinud ei ole. Mitterämpsposti suhteliselt madal precision tähendab, et ei filtreeri rämpsu eriti efektiivselt, näiteks Enron4 puhul on 53% hinnangutest korrektne. Samas on ehk hea uudis see, et suhteliselt vähe on vale-positiivseid ehk me ei saada normaalseid kirju rämpsu alla.

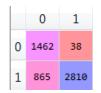
Test treeningandmetel:

| | precision | recall | f1-score | support |
|---------------|--------------|--------------|--------------|--------------|
| False True | 0.84 1.00 | 0.99 0.94 | 0.91 0.97 | 1500 4500 |
| accuracy | | | 0.95 | 6000 |

| | 0 | 1 |
|---|------|------|
| 0 | 1492 | 8 |
| 1 | 281 | 4219 |

Test Enron5 andmetel:

| | precision | recall | f1-score | support |
|---------------|--------------|--------------|--------------|--------------|
| False True | 0.63 0.99 | 0.97 0.76 | 0.76 0.86 | 1500 3675 |
| accuracy | | | 0.83 | 5175 |



Test Enron4 andmetel:

| | precision | recall | f1-score | support |
|---------------|--------------|--------------|--------------|--------------|
| False True | 0.53 1.00 | 1.00 0.70 | 0.69 0.82 | 1500 4500 |
| accuracy | | | 0.78 | 6000 |

| | 0 | 1 |
|---|------|------|
| 0 | 1495 | 5 |
| 1 | 1339 | 3161 |