Tutorium zu „Einführung in die Computerlinguistik“

Naive Bayes + Evaluation

Naive Bayes

Klassifikator:

$$c_{MAP} = \text{argmax}_{c \in C} P(c|d) = \text{argmax}_{c \in C} [P(d|c)\, P(c)] = \text{argmax}_{c \in C} P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

Bayes Rule

>> auswählen welche Klasse den höchsten Wert hat

Wahrscheinlichkeit für Klasse c gegeben Dokument d:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) = P(c) \times P(t_1|c) \times P(t_2|c) \times \ldots \times P(t_n|c)$$

>> Apriori-Wahrscheinlichkeit Klasse mal bedingte Wahrscheinlichkeiten Token in Klasse

Vorgerechnetes Beispiel:
https://www.youtube.com/watch?v=OWGVQfuvNMk&list=PLQiyVNMpDLKnZYBTUOlSI9mi9wAErFtFm&index=29

Was brauchen wir?

Apriori-Wahrscheinlichkeiten (Priors)

P(Klasse) = Anzahl Docs in Klasse / Anzahl Docs gesamt

Bedignte Wahrscheinlichkeiten (Conditional Probabilities)

P(Token|Klasse) = Anzahl dieses Token in Klasse / Anzahl alle Token in Klasse

0 vermeiden >> Smoothing z.B. LaPlace- bzw. Add-One-Smoothing

P(Token|Klasse) = (Anzahl dieses Token in Klasse + 1) / (Anzahl alle Token in Klasse + Vokabular × 1)

Vokabular = Anzahl Types im ganzen Trainingsset >> Vokabular ist für alle Klassen gleich

Wörter, die nicht in unserem Trainingsset vorkommen? >> <UNK> verwenden

P(<UNK>|Klasse) = (0 (Anzahl <UNK> in Klasse) + 1) / (Anzahl alle Token in Klasse + Vokabular + 1)

<UNK> zum Vokabular hinzufügen

Übung

Kategorie	Wörter im Dokument
Essen	Honig Brot Käse
Essen	Tomaten Käse
Essen	Käse Nudeln Mensa
Uni	Mensa Student Kaffee
Uni	Professor Vorlesung
Uni	Klausur Mensa Lernen
Uni	Mensa Professor
Testdokument	Wörter im Testdok.
d1	Student Kaffee Tee
d2	Mensa Käse Professor

To Do:

Apriori Wahrscheinlichkeiten

Bedingte Wahrscheinlichkeiten (mit UNK)

Scores (auf die 6. Nachkommast. runden)

P(Klasse) = Anzahl Docs in Klasse / Anzahl alle Docs

P(Token|Klasse) = Anzahl Token in Klasse + 1
/ Anzahl Token gesamt in Klasse + Vok. + 1
(+1 für UNK)

Evaluation

Accuracy = TP(aller Klassen) / Dokumente gesamt >> alle richtig kassifizierten / alle Dokumente
Precision = TP / (TP + FP) >> richtig als K klassifiziert / alles als K klassifiziertes
Recall = TP / (TP + FN) >> richtig als K klassifiziert / alle tatsächlichen Ks
F1 = 2PR / (P + R)

Beispiel:

Anz. Dokumente: 100

Acc = (50 + 15) / 100 = 0,65
P(A) = 50 / (50 + 25) = 0,67
R(A) = 50 / (50 + 10) = 0,83
F1(A) = 2(50/75)(50/60) / (50/75) + (50/60)= 0,74

Gold \ Klass.	A	B
A	50 TP A	10 FN A
B	25 FP A	15 TN A

Übung

Gold \ Klass.	A	B	C
A	15	10	5
B	9	17	4
C	3	12	25

To Do:

Accuracy

Precision für alle Kategorien

Recall für alle Kategorien

F1 für alle Kategorien

(auf die 2. Nachkommast. runden)

Lösungen

Naive Bayes:

d1 -> Uni (Score: 0,000188)
d2 -> Uni (Score: 0,000564)

Evaluation:

Acc = 0,57

$P_A = 0{,}56$ $R_A = 0{,}50$ $F1_A = 0{,}53$
$P_B = 0{,}44$ $R_B = 0{,}57$ $F1_B = 0{,}49$
$P_C = 0{,}74$ $R_C = 0{,}63$ $F1_C = 0{,}68$