1. Naive Bayes

Ein Naive Bayes Textklassifikator kann sowohl mit den multinomial Modell als auch mit dem Bernouilli Modell eingerichtet werden.

* 1. Das Multinomial Naive Bayes

Der Naive Bayes Klassifikator ist eine Familie von machine learning Algorithmen, die den Satz von Bayes zur Klassifizierung von Objekten anwendet. Die erste machine learning Methode ist die multinomiale Naive Bayes (Multinomiale NB). Die Berechnung der Wahrscheinlichkeit, dass ein Dokument d ist in eine Klasse enthalten, ist durch folgende Formel:

(13.2)  
 entspricht die Wahrscheinlichkeit des in einem Dokument der Klasse c auftretenden Begriffs . kann auch interpretiert werden als ein Maß für wie hoch der Beitrag von t ist bei Bestimmung der Korrektheit von der Klasse c. P(c) ist die vorherige (Engl.: Prior) Wahrscheinlichkeit eines in Klasse c auftretendes Dokument. Die Menge von bis sind die Tokens in d für die Klassifizierung und entspricht die Anzahl dieser Tokens in d. Für die Textklassifizierung wird die Findung der besten Klasse für das Dokument gezielt. Diese beste Klasse in der Textklassifizierung wird maximale a posteriori (MAP) Klasse

(13.3)

Da die echte Werte der Parametern P(c) und unbekannt sind wird für P geschrieben.

Die Logarithmen Wahrscheinlichkeiten sollen eher addiert werden anstatt eine Multiplikation von Wahrscheinlichkeiten durchzuführen, da diese Multiplikation von Wahrscheinlichkeiten Gleitkommaunterlauf erzeugt. Die Klasse die am wahrscheinlichsten ist, entspricht die Klasse mit dem höchsten Wahrscheinlichkeitswert und Log (xy) = log(x) + log(y). Wegen dem monotoner Aspekt der Logarithmusfunktion, die in den meisten Naive Bayes Implementierung durchgeführte Maximierung ist die folgende:

(13.4)

ist das Prior und entspricht das Gewicht für die relative Häufigkeit c und entspricht eine Gewichtung für wie gut ein Indikator für c ist. Mit der Summe von logarithmischen Prioritäts- und Begriffsgewichte wird die Anzahl an Beweise für das Dokument in der Klasse gemessen. Die Auswahl der Klasse mit den meisten Beweise wird mit der Gleichung (13.4) durchgeführt. Zur Schätzung der Parameter und kann das Maximum -Likelihood-Schätzung (MLS) eingesetzt, eine relative Häufigkeit, die der wahrscheinlichsten Wert jedem Parameter unter Berücksichtigung der Trainingsdaten ist. Diese Schätzung für Prioren ist die folgende:

(13.5)

entspricht die Anzahl an Dokumente in der Klasse c und N die Gesamtzahl von Unterlagen.

Die bedingte Wahrscheinlichkeit kann als relative Häufigkeit von Begriffe t in Dokumente der Klasse c geschätzt werden:

(13.6)

entspricht die Anzahl der Vorfälle von t in Trainingsdokumenten aus Klasse c, sowie mehrfacher Vorfall eines Begriffes in einem Dokument. Falls ein Begriff beispielsweise zwei Vorkommen in einem Dokument hat, jeweils in den Positionen k1 und k2, dann

(tk1 | c) = (tk2 | c). Die Abbildung 1 zeigt multinomial Naive Bayes Trainings- und Testalgorithmen.

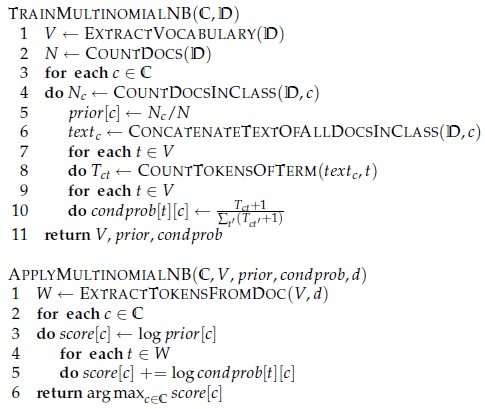


Abbildung 1: Das multinomial Naive Bayes Algorithmus (Training und Test) [MRS08].

Es besteht bei der MLS das Problem, dass das MLS ist gleich Null für die in den Trainingsdaten nicht vorkommende Ausdrucksklassenkombination. Diese Null Schätzung besteht wegen der Seltenheit. Ein Weg zur Beseitigung dieser Null-Werte wird durch Add-One- oder Laplace-Glättung gefunden, wo das Einfügen von eins zu jedem zählen stattfindet:

= (13.7)

B = |V| entspricht die Anzahl an Begriffe in das Vokabular. Die Interpretation der Add-On-Glättung ist, dass diese ein einheitlicher Prior ist, dessen Aktualisierung stattfindet sobald Beweise aus den Trainingsdaten einlaufen.

* 1. Das Bernouilli Modell

Das Bernouilli Modell kann alternativ zu dem multinomialen Modell angewendet werden. Das Bernouilli Modell korrespondiert zu dem binären Unabhängigkeitsmodell, das verantwortlich ist für die Erzeugung eines Indikators für jeden Begriff des Vokabulars (Anzeigen von 0 für Abwesenheit und 1 für die Anwesenheit eines Dokuments) und hat eine äquivalente Komplexität zu dem multinomialen Modell. Im Abbildung 2 werden das Trainings- und Testalgorithmen für das Bernouilli Modell dargestellt.

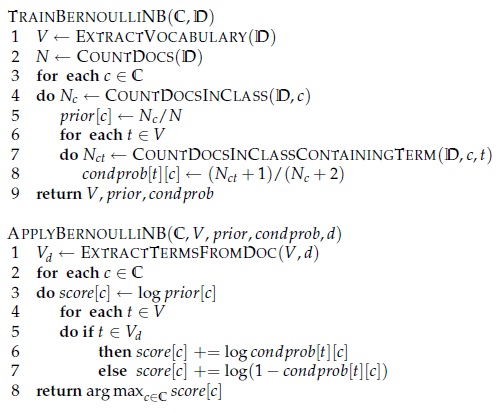


Abbildung : Naive Bayes Algorithmus mit Bernouilli Modell (Training und Test) [MRS08].

Die Erzeugungsmodelle werden mit unterschiedlichen Schätzstrategien und Klassifikationsregeln durchgeführt. Während Bernouilli Modell bewertet P(t | c) als Bruchteil von Begriff enthaltende Dokumenten der Klasse c (Vergleich Abb. 2 Trainbernouilli NB, Zeile 8), bewertet das multinomiale Modell P (t | c) als den Bruchteil von Token oder Bruchteil von Positionen in Begriff enthaltende Dokumenten der Klasse c. Bei der Klassifizierung eines Testdokuments erfolgt durch das Bernouilli Modell die Benutzung von binäre Informationen zum Auftreten. Während dieser Klassifizierung bei dem multinomialen Modell erfolgt die Verfolgung mehrerer Vorkommen. Im Gegensatz zu dem Bernouilli Modell, wo die Anzahl an Vorkommen nicht verfolgt wird. Aus diesem Grund ist das Bernouilli Modell fehlerhaft, was die Klassifizierung langer Dokumenten angeht.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, New York 2008.