Lösungen für Aufgabenblatt 6 — 28.11.2019

Einführung in die Computerlinguistik (WS 19/20) — Abgabe bis 7.12.2019

Aufgabe 1) Naive Bayes Classification **Punkte: 8**

Gegeben ist das folgende Trainingsset und Testdokument:

	ID	Dokument	c=Veganismus
Traininsset	3	Tofu Margarine	+
	4	Hefe Seitan Seitan Tofu	+
	1	Milch Käse Joghurt	-
	2	Käse Butter Käse	-
Testset	5	Käse Hefe Seitan Hefe	?

a) Multinomial Naive Bayes *Punkte: 4*

Schätzen Sie einen **multinomialen** Naive Bayes Klassifizier auf Basis der Trainingsdaten. Welche Klasse weist Ihr Klassifizier dem Dokument aus dem Testset zu? Geben Sie alle Berechnungen als Teil der Lösung mit ab. Benutzen Sie Laplace-Smoothing.

b) Binomial Naive Bayes *Punkte: 4*

Schätzen Sie einen **binomialen** Naive Bayes Klassifizier auf Basis der Trainingsdaten. Welche Klasse weist Ihr Klassifizier dem Dokument aus dem Testset zu? Geben Sie alle Berechnungen als Teil der Lösung mit ab. Benutzen Sie Laplace-Smoothing. Woraus erklären sich die Unterschiede zum multinomialem Modell?

Lösung zu 1b)

Vokabular = {Milch, Käse, Joghurt, Butter, Tofu, Hefe, Seitan, Margarine}

$|Vokabular| = 8$

Prior probabilities:

$$P(c) = \frac{2}{4} = 0.5$$
$$P(\bar{c}) = \frac{2}{4} = 0.5$$

MULTINOMIAL:

Training:

$$P(\text{Tofu}|c) = P(\text{Seitan}|c) = \frac{2+1}{6+8} = 0.214$$

$$P(\text{Margarine}|c) = P(\text{Hefe}|c) = \frac{1+1}{6+8} = 0.143$$

$$P(\text{Milch}|c) = P(\text{Joghurt}|c) = P(\text{Käse}|c) = P(\text{Butter}|c) = \frac{0+1}{6+8}$$

$$P(\text{Milch}|\bar{c}) = P(\text{Butter}|\bar{c}) = P(\text{Joghurt}|\bar{c}) = \frac{1+1}{6+8} = 0.143$$

$$P(\text{Käse}|\bar{c}) = \frac{3+1}{6+8} = 0.286$$

$$P(\text{Tofu}|\bar{c}) = P(\text{Seitan}|\bar{c}) = P(\text{Margarine}|\bar{c}) = P(\text{Hefe}|\bar{c}) = \frac{0+1}{6+8}$$

Testdokument:

$$P(c|\text{Käse Hefe Seitan Hefe}) \propto \frac{2}{4} \cdot \frac{1}{14} \cdot \frac{2}{14} \cdot \frac{3}{14} \cdot \frac{2}{14}$$

$$P(\bar{c}|\text{Käse Hefe Seitan Hefe}) \propto \frac{2}{4} \cdot \frac{4}{14} \cdot \frac{1}{14} \cdot \frac{1}{14} \cdot \frac{1}{14}$$

Man entscheidet sich für *Veganismus*.

BINOMIAL:

Training:

$P(\text{Hefe=1}|c) = P(\text{Seitan=1}|c) = P(\text{Margarine=1}|c) = \frac{1+1}{2+2}$ $P(\text{Hefe=0}|c) = P(\text{Seitan=0}|c) = P(\text{Margarine=0}|c) = \frac{1+1}{2+2}$

$P(\text{Tofu=1}|c) = \frac{2+1}{2+2}$ $P(\text{Tofu=0}|c) = \frac{0+1}{2+2}$

$P(\text{Käse =1}|c) = P(\text{Butter=1}|c) = P(\text{Milch=1}|c) = P(\text{Joghurt=1}|c) = \frac{0+1}{2+2}$

$P(\text{Käse =0}|c) = P(\text{Butter=0}|c) = P(\text{Milch=0}|c) = P(\text{Joghurt=0}|c) = \frac{3+1}{2+2}$

Und genauso: für die andere Klasse:

$$P(\text{Seitan} = 1|\bar{c}) = P(\text{Margarine} = 1|\bar{c}) = P(\text{Tofu} = 1|\bar{c}) = P(\text{Hefe=1}|\bar{c}) = \frac{0+1}{2+2} = \frac{1}{4}$$

$$P(\text{Seitan} = 0|\bar{c}) = P(\text{Margarine} = 0|\bar{c}) = P(\text{Tofu} = 0|\bar{c}) = P(\text{Hefe=0}|\bar{c}) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$P(\text{Butter} = 1|\bar{c}) = P(\text{Milch} = 1|\bar{c}) = P(\text{Joghurt} = 1|\bar{c}) = \frac{1+1}{2+2}$$

$$P(\text{Butter} = 0|\bar{c}) = P(\text{Milch} = 0|\bar{c}) = P(\text{Joghurt} = 0|\bar{c}) = \frac{1+1}{2+2}$$

$$P(\text{Käse} = 1|\bar{c}) = \frac{2+1}{2+2}$$

$$P(\text{Käse} = 0|\bar{c}) = \frac{0+1}{2+2}$$

Testing:

$$P(c|\text{Käse Hefe Seitan Hefe}) = P(c)\cdot P(\text{Milch} = 0|c)\cdot P(\text{Käse} = 1|c)\cdot P(\text{Joghurt} = 0|c)\cdot P(\text{Butter} = 0|c)\cdot P(\text{Tofu} = 0|c)\cdot P(\text{Hefe} = 1|c)\cdot P(\text{Seitan} = 1|c)\cdot P(\text{Margarine} = 0|c) = \frac{2}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{2}{4}$$

$$P(\bar{c}|\text{Käse Hefe Seitan Hefe}) \propto P(\bar{c})\cdot P(\text{Milch} = 0|\bar{c})\cdot P(\text{Käse} = 1|\bar{c})\cdot P(\text{Joghurt} = 0|\bar{c})\cdot P(\text{Butter} = 0|\bar{c})\cdot P(\text{Tofu} = 0|\bar{c})\cdot P(\text{Hefe} = 1|\bar{c})\cdot P(\text{Seitan} = 1|\bar{c})\cdot P(\text{Margarine} = 0|\bar{c}) = = \frac{2}{4}\cdot\frac{2}{4}\cdot\frac{3}{4}\cdot\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}$$

Hier sind die beiden Klassen gleichwahrscheinlich geworden, da wir auch negative Evidenz einbeziehen und das doppelte Vorkommen von *Hefe* nicht berücksichtigen.

Aufgabe 2) Evaluation **Punkte: 4**

Sie richten ein geselliges Abendessen aus und möchten den Ernährungsgewohnheiten all Ihrer Gäste entgegenkommen. Dazu konsultieren Sie einen Multiclass-Klassifizierer, um zwischen vegetarischen (aber nicht veganen), veganen und fleischhaltigen Rezepten zu unterscheiden. Nach manueller Auswertung der Ergebnisse erhalten Sie diese Konfusionsmatrix (hierbei steht z.B. *meat* für die wirklich richtige Lösung und *meat'* für das Resultat des Klassifizierers).

system labels \ gold labels	**meat**	**veggy**	**vegan**	
meat′	109	5	0	114
veggy′	20	65	7	92
vegan′	6	2	7	15
	135	72	14	

Berechnen Sie nun die Precision und Recall-Werte für jede einzelne Klasse, sowie die Micro- & Macroaverage Precision.

Lösung zu 2)

$$\text{Recall}_{meat} = \frac{109}{109 + 20 + 6} = 0.81$$

$$\text{Recall}_{veggy} = \frac{65}{5 + 65 + 2} = 0.90$$

$$\text{Recall}_{vegan} = \frac{7}{0 + 7 + 7} = 0.5$$

$$\text{Prec}_{meat} = \frac{109}{109 + 5 + 0} = 0.96$$

$$\text{Prec}_{veggy} = \frac{65}{20 + 65 + 7} = 0.71$$

$$\text{Prec}_{vegan} = \frac{7}{6 + 2 + 7} = 0.47$$

$$\text{Prec}_{macro} = \frac{\text{Prec}_{meat} + \text{Prec}_{veggy} + \text{Prec}_{vegan}}{3} = 0.713$$

$$\text{Prec}_{micro} = \frac{109 + 65 + 7}{(109 + 65 + 7) + (5 + 27 + 8)} = 0.819$$

Letzteres kann man mit der Tabularmethode wie in Jurafsky und Martin sowie der Vorlesung ausrechnen. Sinnvoll wäre aber auch sich zu überlegen, dass bei one-of Klassifizierern micro-averaged Precision = micro-averaged recall = overall accuracy ist. Bitte letzteres im Tutorium durchsprechen. (Beweise finden sich im Zweifelsfall auch online). Bei any-of Klassifizierern gilt dies nicht!

Wie man hier sehen kann, ist die Microaveraged-Precision höher, da man auf den großen Klassen gut abgeschnitten hat. Dies ist meistens der Fall: will man daher beurteilen, ob man auch die kleineren Klassen gut erkennen kannen, sollte man macro-average benutzen.

Aufgabe 3) Textklassifikation Fragen **Punkte: 8**

1. Wir haben im multinomialen Modell und im binomialen Modell, die a-priori Wahrscheinlichkeit der Klasse c mit

   $P(c) = (Anzahl\ der\ Dokumente\ der\ Klasse\ c)/(Gesamtanzahl\ Dokumente)$

   geschätzt. Im multinomialem Modell hätten wir ja auch die folgende Formel wählen können

   $(Anzahl\ der\ Worte\ in\ Dokumenten\ der\ Klasse\ c)/(Gesamtanzahl\ Worte\ in\ allen\ Dokumenten)$

   Dies wäre analog zu den Featureschätzungen. Warum wird dies wohl nicht getan? (2 Punkte)

2. Die Verwendung von Mutual Information als Merkmalsselektion ist (theoretisch) am geeignetsten für das Bernouillimodell. Warum ist dies so? Wie könnte man die Formel und Berechnungen modifizieren, so dass sie auch für das Multinomialmodell geeignet ist? (2 Punkte)

3. Was ist der Wert von Mutual Information, wenn Klasse und Wort/Term komplett abhängig sind? Was ist der Wert von MI, wenn Klasse und Wort/Term komplett unabhängig sind? (4 Punkte)

Lösung zu 3)

1. Führt zu einer Überbewertung langer Dokumente. Man klassifiziert auch am Ende ganze Dokumente, egal welcher Länge.

2. Da man in MI Worte als binäre Events (Wort kommt vor/nicht vor) sieht sowie für die Dokumente ebenfalls binäre Zufallsvariablen benutzt, ist dies parallel zu den Abschätzungen für binomial NB. Zum Beispiel ist bei der MI-Tabelle (unten) die Interpretation von N Anzahl der Dokumente (nicht Anzahl aller Tokens in allen Dokumenten), die Interpretation von $N11$ ist die Anzahl der Dokumente der Klasse, in denen das Wort vorkommt etc.

	Klasse =1	Klasse =0	
Wort =1	N11	N10	N1.
Wort =0	N01	N00	N0.
	N.1	N.0	N

   Entsprechend ist die Interpretation von $p(Wort = 1, Klasse = 1) = \frac{N11}{N}$, also die Zahl aller Dokumente der Klasse, in denen das Wort vorkommt, durch alle Dokumente.

   Man könnte nun alles stattdessen auf Tokenbasis machen. Dann würde man die entsprechenden Tabellenitems folgendermaßen auswerten (und wie beim multinomail NB alle Dokumente der Klasse bzw Nichtklasse als aneinandergehängt betrachten).

   - N11: Anzahl der Tokens der Klasse, die das Wort sind

- N10: Anzahl der Tokens der nicht-Klasse, die das Wort sind
- N1.: Anzahl der Tokens, die das Wort sind, in allen Dokumenten
- N01: Anzahl der Tokens der Klasse, die nicht das Wort sind
- N00: Anzahl der Tokens der nicht-Klasse, die nicht das Wort sind
- N0.: Anzahl der Tokens im Gesamtkorpus, die nicht das Wort sind
- N.1: Anzahl der Tokens in der Klasse
- N.0: Anzahl der Tokens in der nicht-Klasse
- N= Anzahl der Tokens im gesamten Korpus

Die MI-Formel kann dann wieder über die vier Felder aufsummieren, aber die Interpretation der Wahrscheinlichkeiten ist nun tokenbasiert.

3. Sind Wort W und Klasse C vollständig abhängig, heisst dies, dass W in jedem Dokument der Klasse C vorkommt und nie außerhalb (das Wort ist also der perfekte Indikator). Damit ist die Tabelle

	$C=1$	$C=0$	
$W=1$	$\#c$	0	$\#c$
$W=0$	0	$N-\#c$	$N-\#c$
	$\#$c	$N-\#c$	N

Hier soll #c für die Anzahl der Dokumente in Klasse C stehen.

Dies kann man nun in die Formel einsetzen. Man beachte hierbei $0 \log_2 0 := 0$ (siehe Entropievorlesungen).

Damit ergibt sich

$$MI(W,C) = \frac{\#c}{N} \cdot \log\left(\frac{\#c/N}{(\#c/N)\cdot(\#c/N)}\right) + 0 + 0 + \frac{N-\#c}{N} \cdot \log\left(\frac{(N-\#c)/N}{(N-\#c)/N \cdot (N-\#c)/N}\right) =$$
$$\frac{\#c}{N} \cdot \log\frac{N}{\#c} + \frac{N-\#c}{N} \cdot \log\frac{N}{N-\#c} = -\frac{\#c}{N} \cdot \log\frac{\#c}{N} - \frac{N-\#c}{N} \cdot \log\frac{N-\#c}{N}$$

Das heisst, man bekommt nun die Entropie der Klassenverteilung.

Als Beispiel kann man sich noch anschauen, was passiert, wenn die Klasse genau die Hälfte der Dokumente umfasst, dann gälte, bei perfekter Abhängigkeit:

$MI(W,C) = 0.5 \cdot \log 2 + 0.5 \log 2 = \log 2 = 1$

d.h. die MI ist nur noch von der Klassenverteilung abhängig und in diesem Fall maximal.

Sind das Wort und die Klasse vollständig unabhängig, dann ist die Mutual Information Null, da für alle Werte der beiden Variablen C und W gilt $p(C = x, W = y) = p(C =$

$x) \cdot p(W = y)$ sowie $\log 1 = 0$.

Aufgabe 4) Bonusaufgabe: Textklassifikation mit non-word features **Punkte: 4**

Sie möchten einen Spamklassifizierer schreiben. Ihnen stehen folgende Merkmale und Trainingsdaten zur Verfügung.

1. Anzahl der Ausrufezeichen
2. Menge an "Bild" im Vergleich zu "Text"
3. Ist die *subject line* in *all caps*?
4. Vorkommen des Wortes *Gewinn*
5. Ist die subject line all caps?

	ID	Ausrufezeichen	Bild/Text	all caps	"Gewinn"	Spam
Trainingsset	1	viel	niedrig	ja	ja	+
	2	wenig	niedrig	ja	ja	+
	3	wenig	mittel	ja	nein	+
	4	viel	mittel	nein	nein	+
	5	mittel	hoch	nein	ja	+
	6	viel	hoch	nein	ja	+
	7	viel	mittel	nein	nein	+
	8	mittel	niedrig	nein	nein	-
	9	wenig	niedrig	ja	nein	-
	10	wenig	niedrig	nein	ja	-

Beachten Sie, dass es Sie hier ihr Modell leicht abwandeln müssen, denn Sie müssen nun ja Wahrscheinlichkeiten wie die untigen schätzen:

- $P(Ausrufezeichen = wenig|+)$
- $P(Ausrufezeichen = mittel|+)$

1. Wie würden Sie diese Wahrscheinlichkeiten schätzen? (Smoothing können Sie hier einmal ignorieren.). (1 Punkt)
2. Berechnen Sie nun die Bewertung für eine email mit den Features **[mittel, mittel, ja, ja]** mit einem NB-Äquivalent (ohne Smoothing). Hierzu müssen Sie nur die Wahrscheinlichkeiten schätzen, die für genau diese Test-Email notwendig sind. (2 Punkte)
3. Was wäre die Wahrscheinlichkeit für $p(Ausrufezeichen = viel|ja)$ mit Laplace-Smoothing (1 Punkt)

Lösung zu 4)

1. Man sollte dokumentweise, wie im binomialen Modell abschätzen, also zum Beispiel:

$$P(Ausrufezeichen = wenig|+) = \frac{2}{7}$$

Tokenweise macht bei Features wie Bild/Text keinen Sinn.

2. Man ignoriert negative Evidenz und berechnet:

   $$P(+|mittel, mittel, ja, ja) \propto \frac{7}{10} \cdot \frac{1}{7} \cdot \frac{3}{7} \cdot \frac{3}{7} \cdot \frac{4}{7}$$

   sowie

   $$P(-|mittel, mittel, ja, ja) \propto \frac{3}{10} \cdot \frac{1}{3} \cdot \frac{0}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}$$

3. $P_{LP}(Ausrufezeichen = viel|ja) = \frac{4+1}{7+3}$

   Hierbei kommt die "3" im Nenner von der Anzahl der Werte, die die Variable *Ausrufezeichen* annehmen kann (*viel, wenig,mittel*).