Schriftliche Prüfung zur Vorlesung
Statistische Methoden in der maschinellen Sprachverarbeitung
WS 2017/18
Dozent: Helmut Schmid

Aufgabe 1) Wie lautet das **Theorem von Bayes**? Nennen Sie ein Beispiel, wo wir dieses Theorem benutzt haben. (1 Punkt)

Aufgabe 2) Geben Sie die Formel für die Berechnung des Erwartungswertes der Funktion $-\log p(x)$ für die Zufallsvariable X an. Welchen anderen Namen gibt es für diesen speziellen Erwartungswert? (1 Punkt)

Aufgabe 3) Wie wird die (ungeglättete) bedingte Wahrscheinlichkeit $p(w_3|w_1, w_2)$ aus Häufigkeiten geschätzt? (1 Punkt)

Aufgabe 4) Wie wird die bedingte Wahrscheinlichkeit $\hat{p}(w_3|w_1, w_2)$ mit interpolierter Backoff-Glättung und absolute Discounting geschätzt? Wie wird der Discount berechnet? Wie wird der Backofffaktor berechnet? (3 Punkte)

Aufgabe 5) Geben Sie an, mit welchen Umformungsschritten man die hier gezeigte Formel für das Hidden-Markov-Modell herleiten kann.

$$\arg\max_{t_1^n} p(t_1^n|w_1^n) = ... = \arg\max_{t_1^n} \prod_{i=1}^{n+1} p(t_i|t_{i-1})p(w_i|t_i)$$

(3 Punkte)

Aufgabe 6) Was wird im E-Schritt und im M-Schritt des EM-Algorithmus jeweils berechnet? (Ein Satz genügt jeweils.) (1 Punkt)

Aufgabe 7) Der Forward-Backward-Algorithmus wird verwendet, um erwartete Häufigkeiten von Wort-Tag-Paaren und von Tag-Tag-Paaren zu berechnen.

Geben Sie die Formeln an für die Berechnung (i) der Forward-Wahrscheinlichkeiten $\alpha_t(i)$ (ii) der Backward-Wahrscheinlichkeiten $\beta_t(i)$ und (iii) der erwarteten Häufigkeit $\gamma_t(i)$ (= Aposteriori-Wahrscheinlichkeit) des Wortes w_i mit dem Tag t an. (5 Punkte)

Aufgabe 8) Angenommen, Sie haben einen neuen Wortart-Tagger entwickelt. Von Ihrem Tagger gibt es drei Varianten. Sie trainieren alle drei Taggervarianten auf denselben Trainingsdaten und testen sie auf denselben Testdaten. Zum Vergleich evaluieren Sie einen anderen Tagger (Baseline-Tagger) auch auf denselben Trainings- und Testdaten. Ein statistischer Test ergibt bei Ihrer dritten Taggervariante ein signifikant besseres Ergebnis als beim Baseline-Tagger (mit einer Fehlerwahrscheinlichkeit von 0,04). Bei den beiden anderen Taggervarianten zeigt derselbe statistische Test keine signifikante Verbesserung an. Warum dürfen Sie hier nicht behaupten, dass Ihre dritte Taggervariante signifikant besser ist als der Baseline-Tagger? (1 Punkt)

Aufgabe 9) Sie trainieren einen Conditional-Random-Field-Tagger mit der Gradientenanstiegsmethode. Welche Zielfunktion maximieren Sie hier? Wie ist der Gradient der Zielfunktion definiert? Welchen Algorithmus benötigen Sie zur Berechnung des Gradienten? Wie vermeiden Sie "Overfitting", also dass die Parameter zu stark an die Trainingsdaten angepasst werden?

Ein Satz genügt für jede Frage (ohne Formeln und genaue Beschreibungen). (2 Punkte)

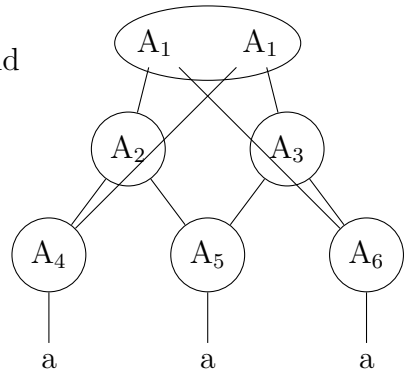
Aufgabe 10) Wie definiert ein loglineares Modell die bedingte Wahrscheinlichkeit der Klasse y für ein gegebenes Objekt x auf Basis der Merkmale $f(x, y)$ (mit Formel)?
(2 Punkte)

Aufgabe 11) Es soll ein Klassifikator trainiert werden, der einem Zeitungsartikel ein Themengebiet (Politik, Wirtschaft, Sport etc.) zuweist. Welches generative statistische Modell eignet sich dafür? Wie trainieren Sie das Modell? Welche Art von Daten benötigen Sie dazu? Wie berechnen Sie die wahrscheinlichste Kategorie eines Artikels (mit Formel und Angabe, was die Variablen bedeuten)?

Anmerkung: Wir haben diese Anwendung nicht in der Vorlesung besprochen. Sie müssen hier Ihr Wissen auf eine neue Anwendung übertragen. (4 Punkte)

Aufgabe 12) Gegeben sei eine PCFG mit zwei Regeln und den Wahrscheinlichkeiten $p(A \rightarrow a) = 1/4$ und $p(A \rightarrow A\ A) = 3/4$, wobei A das Startsymbol und a ein Terminalsymbol ist.

Betrachten Sie den Parsewald



Berechnen Sie die Inside-Wahrscheinlichkeiten der Parsewald-Nichtterminale A_1,...,A_6.

Berechnen Sie dann ihre Outside-Wahrscheinlichkeiten.

Berechnen Sie zum Schluss die Aposteriori-Wahrscheinlichkeit (erwartete Häufigkeit) der Parsewaldregel $A_2 \rightarrow A_4\ A_5$.

Tipp: Überlegen Sie, welche Knoten dieselben Inside-Wahrscheinlichkeiten besitzen.

Tipp 2: Berechnen Sie zur Kontrolle auch noch die Outside-Wahrscheinlichkeiten der Terminalknoten a. Diese müssen mit der Insidewahrscheinlichkeit von A_1 übereinstimmen.

Hilfsmittel: Liste der Zweierpotenzen:

1	2	3	4	5	6	7	8	9	10
2	4	8	16	32	64	128	256	512	1024

(6 Punkte)

(30 Punkte insgesamt)

Viel Erfolg!