Centrum f"ur Informations- und Sprachverarbeitung (CIS)
"Ubungen zum Masterkurs P2 ,,Vertiefung Grundlagen der Computerlinguistik"
WS 2019/20 Klaus U. Schulz

Kontrollfragen zur Statistik

Grundlagen

1. Deskriptive Statistik
Begriffe Mittelwert, Median, Modus, empirische Varianz, Standardabweichung, Spannweite bei einer Messung auf einer Stichprobe.

2. Wahrscheinlichkeitsbegriff
Was ist ein Wahrscheinlichkeitsraum (Bedingungen), was ist mit einem Ereignis genannt?
Wie kann man Ereignisse "Boolesch" kombinieren, was sind disjunkte Ereignisse?
Welche Grundrechenregeln für die Wahrscheinlichkeit von kombinierten Ereignissen gibt es?
Wann nennt man n Ereignisse unabhängig bzw. abhängig? Laesst sich das auf die Unabhängigkeit zweier Ereignisse zurückführen?
Was versteht man unter der bedingten Wahrscheinlichkeit (Intuition, Formel?)
Was ist das Bayessche Gesetz, was kann man damit anstellen, warum ist es so wichtig?

3. Diskrete Zufallsvariablen und Wahrscheinlichkeitsverteilungen
Was versteht man unter einer (binären, diskreten) Zufallsvariablen? Wo treten diese auf?
Was ist eine Wahrscheinlichkeitsverteilung?
Welche Rolle spielen hier "Parameter" und "Familien von Wahrscheinlichkeitsverteilungen"?
Was versteht man unter dem Erwartungswert, der Varianz, der Standardabweichung einer Zufallsvariablen (Intuition, Formel)? Welche einfachen Rechengesetze gibt es für Erwartungswerte, Varianzen?
Beispiele diskreter ZV: Was ist die Binomialverteilung, welche Parameter charakterisieren diese? Was ist die Poissonverteilung (Parameter?)? Wo treten binomialverteilte ZV auf? In welchen Zusammenhang stehen Binomial- und Poissonverteilung? Was ist ein Beispiel für einen "Poissonprozess"?

4. Stetige Zufallsvariable und Wahrscheinlichkeitsverteilungen
Was eine stetige Zufallsvariable?
Was ist eine Dichtefunktion, in welchen Zusammenhang stehen stetige Zufallsvariable, diskrete Approximationen und Dichtefunktionen?
Wie sehen Erwartungswert, Varianz und Standardabweichung bei stetigen Zufallsvariablen aus (Formel)?
Was versteht man unter der/einer Normalverteilung, was unter der Standardnormalverteilung?
Was versteht man unter der Chi-Quadrat-Verteilung?

5. Zwei- und mehrdimensionale Wahrscheinlichkeitsverteilungen
Was sind natürliche Praxis-Beispiele für mehrdimensionale Zufallsvariablen? (Mietspiegel: Nettomiete, Wohnfläche, Zimmerzahl, Roulette: Farbe als Zahl und gerade ungerade)
Wahrscheinlichkeitsfunktion (gemeinsame Verteilung) zweier diskreter Zufallsvariablen.
Was sind Randverteilungen oder marginale Verteilungen (2 ZV)? Was ist eine Kontingenztabelle?
Wann nennt man zwei Zufallsvariablen unabhängig bzw. abhängig?
Was versteht man unter der Kovarianz, dem Korrelationskoeffizienten zweier Zufallsvariable?

6. Parameterschätzung
Was ist Idee, Sinn und Zweck einer Parameterschätzung?
Was versteht man unter "Stichprobenvariablen", was ist eine Punktschätzung, eine Schätzfunktion?
Was versteht man unter dem Begriff der "Erwartungstreue" bzw, der Verzerrung (Bias)?
Was ist eine Likelihoodfunktion, was ist eine "Maximum likelihood-Schätzung"? Welche Rolle hat die "log-Likelihood", warum wird sie betrachtet?
Beispielverfahren zu Ermittlung der "Maximum likelihood"?
Was ist eine Intervallschätzung? In welchen Zusammenhang stehen "Irrtumswahrscheinlichkeiten" und "Konfidenzintervalle"?

7. Testen von Hypothesen
Was sind die allgemeinen Prinzipien des Hypothesentestens?
Welche Rolle spielen Nullhypothese und Gegenhypothese,Signifikanzniveau, Prüfgröe und Prüfverteilung, Ablehnungsbereich?
Welche Fehlerarten können beim Testen auftreten?

8. Statistik und Informationstheorie
Was versteht man unter der Entropie einer (diskreten) ZV, Formel und Interpretation?
Bei n möglichen Werten der ZV, welche Verteilungen haben hohe (niedrige) Entropie?
Was versteht man unter der bedingten Entropie zweier ZV und unter der "Mutual Information" zweier ZV bzw. zweier Werte?

Anwendungen

Überblick
Was sind wichtige oder typische Anwendungen der Statistik in den Bereichen Syntax, Semantik, Morphologie, Texklassifikation, Lexikographie und Computerlinguistik?

Zipfsches Gesetz
Was besagt das Zipfsche Gesetz (einfache Form), als Formel, intuitiv?
Zu welcher Art Vorhersagen über das Vokabular eines Textes kann man verwenden?
Was folgt aus dem Zipfschen Gesetz für Erweiterungen eines gegebenen Korpus?

Kollokationen
Was sind "Kollokationen"?
Welche Test- und korpusbasierten Methoden gibt es, um zu prüfen, ob ein Wortpaar eine Kollokation darstellt?
Auf welchen Annahmen beruhen diese meist? Sind diese realistisch?
Wie kann man Score-Rangfolgen einsetzen, um TEILautomatisch ECHTE Kollokationen zu finden?

Statistische Sprachmodelle
(Vgl. Fink-Buch) Wie kann man die Wahrscheinlichkeit einer Wortfolge ausdrücken?
Was ist ein Markov-Modell n-ter Ordnung?
Wie stellt sich unter dieser Vereinfachungssnnahme die W einer Wortfolge dar?
Was ist ein n-Gramm Modell?
Welche Probleme stellen sich bei der Erstellung von n-Gramm-Modellen (für unterschiedliches n) aus einem Korpus?
Was versteht man unter "Smoothing", was unter Laplace (adding one) Smoothing, was unter Discounting, unter "Backing off"?

Was versteht man unter der Perplexität eines Sprachmodells auf einem Text, und was bedeutet dies für die Qualität des Sprachmodells?

HMMs
Was ist ein stochastischer Prozess, was eine Markov-Kette?
Was ist formal gesehen ein HMM? Wie kann man es interpretieren, warum "Hidden"?
Was versteht man unter dem Viterbi-Verfahren, was wird dabei ermittelt?
Was versteht man unter der Forward- bzw- Backward-Variablen?
Wie kann man die Gesamtwahrscheinlichkeit einer Beobachtungsfolge bei gegebenem Modell ermitteln?
Was sind die Ideen beim Baum-Welch bzw. Viterbi-Training?
Anwendungen: wie kann man HMMs zum POS-Tagging einsetzen?