Time	Group	Submission in Moodle; Mails with subject: [SMD2023]
Th. 12:00–13:00	A	lukas.beiske@udo.edu and tristan.gradetzke@udo.edu
Fr. 08:45–09:45	B	jonas.hackfeld@ruhr-uni-bochum.de and ludwig.neste@udo.edu
Fr. 10:00–11:00	C	stefan.froese@udo.edu and vincent.latko@udo.edu

Exercise 1 *Kurzfragen* **0 p.**

Dieses Blatt besteht auf Kontrollfragen, die bei der Klausurvorbereitung helfen sollen. Beantwortet werden sollen die Fragen mit dem Inhalt der Vorlesung und vorallem auch mit dem Inhalt der Übungsblätter. Die Aufgaben a) bis f) entsprechen den Kurzfragen aus SMD A und dienen der erneuten Wiederholung.

(a) Numerische Grundlagen:

- Wie werden die einzelnen Zahlen Typen im Computer gespeichert?
- Wie ist der Fehler von zweistelligen Operationen begrenzt?
- Wie können elementare Funktionen approximiert werden?
- Was ist der Unterschied zwischen Stabilität und Kondition?
- Auf was muss geachtet werden um eine Funktion stabil zu machen?

(b) Eindimensionale Verteilungen:

- Was ist der Unterschied zwischen bayesischer und frequentistischer Definition von Wahrscheinlichkeit?
- Was muss bei der Kombination von Wahrscheinlichkeiten beachtet werden?
- Was sind Verteilungsfunktionen und Wahrscheinlichkeitsdichte?
- Wie sind Erwartungswert und die Momente einer Verteilung definiert?
- Welche Wahrscheinlichkeitsverteilungen gibt es, wie sind diese miteinander verknüpft und was sind typische Anwendungen?

(c) Mehrdimensionale Verteilungen:

- Was ist die Kovarianz und wie ist sie definiert?
- Was ist der Korrelationskoeffizient und welche Werte kann er annehmen?
- Wie bestimmt man eine Randverteilung einer mehrdimensionalen Verteilung und was ist ihre Bedeutung?

(d) Generation von Zufallszahlen:

- Warum spricht man von Pseudo-Zufallszahlen-Generatoren?
- Welche Methoden gibt es um aus gleichverteilten Zufallszahlen, Zufallszahlen zu erzeugen, die einer bestimmten Verteilung folgen? Vergleichen sie diese.
- Welche Methoden gibt es um normalverteilte Zufallszahlen zu erzeugen?
- Welche Methoden gibt es um poisson-verteilte Zufallszahlen zu erzeugen?
- Welche Methoden gibt es um Chi-Quadrat-verteilte Zufallszahlen zu erzeugen?

(e) Fehlerfortpflanzung:

- Wie transformieren sich Varianzen in linearen Abbildungen?
- Wie geht man bei nicht-linearen Problemen vor?

(f) Data-Mining:

- Was sind die, in der Vorlesung vorgestellten, Schritte eines KDD-Prozesses?
- Wie fließt Hintergrundwissen in den KDD-Prozess mit ein?
- Was sind gängige Schritte der Datenbereinigung?
- Wann und wieso bedarf es einer Attributswahl?
- Welche Methoden zur Attributswahl kennen Sie, wie funktionieren Sie?
- Wie funktioniert die LDA und welche Bedingung gilt für die gesuchte Trennung?
- Was sind Qualitätsmaße für eine Separation? Was sagen Sie aus?
- Wie funktioniert eine Hauptkomponentenanalyse (PCA)?
- Worin liegt der Unterschied zwischen PCA und LDA?
- Welche beiden Arten des maschinellen Lernens wurden in der Vorlesung behandelt? Worin liegen die wesentlichen Unterschiede?
- Welche unterschiedlichen Ziele beim überwachten Lernen wurden vorgestellt?
- Was ist Clustering? Worin unterscheiden sich die verschiedenen Arten des Clusterings?
- Wieso ist es sinnvoll Lernverfahren zu validieren?
- Wie funktioniert die Naive-Bayes-Klassifikation und was ist die Laplace Korrektur?
- Wie funktioniert ein Entscheidungsbaum?
- Welche Qualitätskriterien zur Attributs-/Schnittwahl kennen Sie im Bezug auf Entscheidungsbäume?
- Wie kann Überanpassung bei Entscheidungsbäumen reduziert werden?
- Worin unterscheidet sich ein Random Forest von einem Entscheidungsbaum?
- Was ist Boosting?
- Machen Sie sich mit der groben Funktionsweise aller vorgestellten Algorithmen des überwachten Lernens vertraut.

(g) Schätzen:

- Was ist der Unterschied zwischen Punkt- und Intervallschätzung?
- Wie ist der n-te Momentenschätzer definiert?
- Welche Methoden zur Punktschätzung haben Sie in der Vorlesung kennengelernt?
- Was sind die vorgestellten Kriterien zur Beurteilung eines Schätzers? Wie sind sie definiert und was ist ihre Aussage?
- Was versteht man allgemein unter Fitten?
- Was ist die Methode der kleinsten Quadrate und wie sieht das Optimierungskriterium aus?
- Was verändert sich bei der Methode der kleinsten Quadrate, wenn Varianzen und Kovarianzen der Datenpunkte berücksichtig werden sollen?

- Wie muss mit nicht linearen Modellen bei der Modellanpassung mittels der kleinste Quadrate Methode umgegangen werden?
- Was beschreibt eine Likelihood?
- Was versteht man unter dem Begriff Regularisierung? Wann wird diese erforderlich? Welche unterschiedlichen Methoden haben Sie in der Vorlesung kennengelernt?

(h) Testen:
Konfidenzintervalle:

- Ein Konfidenzintervall $[x_1, x_2]$ eines Parameters x zu einem Konfidenzniveau α sei gegeben. Was ist die frequentistische, was die bayesche Interpretation dieses Intervalls?
- Welche Bedeutung hat der Prior in der bayeschen Statistik?
- Welche Freiheiten gibt es bei der Wahl des Intervalls? Was passiert im Spezialfall symmetrischer pdf? Was ist der Unterschied von Intervallen und Upper/Lower Limits?
- Wie werden Konfidenzintervalle nach Neyman konstruiert? Was sind Probleme dieser Konstruktion?
- Wie werden die Probleme der Neyman-Konstruktion in der Feldman-Cousins-Konstruktion behandelt?

Hypothesentests:

- Was ist die Null- / Alternativhypothese?
- Was sind die vier möglichen Szenarien bei Hypothesentests?
- Was sind Typ I und Typ II Fehler? Wie hängen die Begriffe Signifikanz (*Significance*) und Trennkraft (*Power*) damit zusammen?
- Was ist ein p-Value?
- Wie funktioniert der Likelihood-Quotienten-Test? Was ist seine anschauliche Bedeutung?
- Was besagt das Wilk-Theorem?
- Was kann mit dem Kolmogorov-Smirnow-Test getestet werden? Wie wird die Teststatistik konstruiert?
- Wie ist der χ^2 Test konstruiert? Was sind typische Anwendungen?
- Wie ändert sich die Anzahl der Freiheitsgrade der χ^2 Teststatistik, wenn Parameter aus Daten bestimmt werden?
- Was erwarte ich für den χ^2/ndof Wert unter der Nullhypothese?
- Was muss beachtet werden, wenn ein χ^2 Test bei Histogrammen angewendet wird?

(i) Entfaltung:

- Was ist das Ziel der Entfaltung? Welche drei Eigenschaften charakterisieren das Enfaltungsproblem?
- Welche Prozesse erschweren die Messung der wahren Verteilung?
- Was ist der Vorteil der hier vorgestellten Entfaltungsmethode z.B. gegenüber dem *Forward Folding*?
- Wie wird das Entfaltungsproblem mathematisch beschrieben (Integralgleichung)?

- Was wird durch die Responsefunktion (bzw. Responsematrix) beschrieben? Wie wird die Matrix in der Praxis bestimmt?
- Was ist das Problem einer Entfaltung durch einfache Invertierung der Responsematrix? Welche Ursache haben diese Probleme? Wie verhalten sich die Eigenwerte der Responsematrix?
- Mit welcher Technik werden diese Probleme behandelt? Wie funktioniert sie?
- Im Fall einer Regularisierung, wie verhält sich die Korrelationsmatrix der entfalteten Verteilung im Vergleich zum unregularisierten Fall?
- Wie werden kleine Statistiken in den Bins bei der Entfaltung korrekt berücksichtigt?

(j) Analyse: Sie haben die Aufgabe bekommen, ein Energiespektrum zu messen. Anschließend sollen verschiedene Theoriemodelle, welche jeweils leicht unterschiedliche Energiespektren vorhersagen, getestet werden. Im Folgenden sollen Sie die Analyseschritte kurz erläutern, die für diese Analyse notwendig sind. Manche Teilaufgaben haben mehrere mögliche Lösungen.
Tipp: Erinnern Sie sich auch an die Themen des vorherigen Semesters (SMD A).

a) Sie stellen fest, dass Ihr Detektor hauptsächlich Untergrundereignisse und nur sehr wenige Signalereignisse misst, was eine direkte Analyse unmöglich macht. Was können Sie tun, um dieses Problem zu lösen?

b) Sie haben nun ein Datenset von Signalkandidaten. Für jedes Event misst der Detektor viele verschiedene Variablen. Leider ist keine dieser Variablen die Energie des gemessenen Teilchens. Wie können Sie die Energie eines gemessenen Teilchens schätzen?

c) Was gibt es bei den beiden vorherigen Schritten zu beachten und wie können Sie überprüfen, dass die gewählten Methoden ordnungsgemäß funktionieren?

d) Nun haben Sie zu jedem gemessenem Ereignis auch eine Energieschätzung. Wie können Sie das Energiespektrum erstellen?

e) Was für eine Art von Problem wird im vorherigen Schritt gelöst und wie lautet die zugrunde liegende Gleichung?

f) Sie haben nun die Datenpunkte Ihres gemessenen Energiespektrums. Wie können Sie testen, welches Theoriemodel am besten passt?