- Was ist Overfitting?
- Was ist Change Detection? Wie unterschieden sich multivariate von univariaten Datenströmen?
- ‚Kategorisierung von Daten' (erste Folie dieses Kapitels) sprechend wiedergeben können.
- Gegeben Aggregationsfunktion X, ist sie distributiv/algebraisch/holistisch/self-maintainable?
- Sehen Sie einen allgemeinen Zusammenhang zwischen distributiv/algebraisch/holistisch auf der einen und self-maintainable auf der anderen Seite? (Es gibt einen.)
- Wie groß ist die Entropie, wenn alle Klassen gleich häufig sind? Schreiben Sie dafür eine allgemeine Formel hin.
- Welche Möglichkeiten kennen Sie, die Stärke des Zusammenhangs zwischen zwei Zufallsvariablen zu quantifizieren?
- Welche statistischen Tests kennen Sie?
- Wofür genau sind die einzelnen Tests gut?
- Was bedeutet Dimensionality Reduction?
- Welche Möglichkeiten der Dimensionality Reduction kennen Sie?
- Welche Diskretisierungsverfahren kennen Sie?
- Wie findet man jeweils den besten Merge bzw. Split?
- Warum kann man für räumliche Anfragen nicht ohne weiteres auswerten, wenn man für jede Dimension separat einen B-Baum angelegt hat?
- Wie ist der R-Baum aufgebaut?
- Wie funktioniert die Suche nach dem nächsten Nachbarn mit dem R-Baum? (Frage ich ganz gerne u. a. dann, wenn die Prüfung stockend verläuft...)
- Was ändert sich, wenn die Objekte eine räumliche Ausdehnung haben? Dto. Anfragen.
- Stören uns Überlappungen von Knoten des R-Baums? Wenn ja, warum?
- Wie unterscheiden sich R-Baum, kD-Baum und kDB-Baum? (Wie Balancierung für kDB-Baum funktioniert, muss man für Prüfung nicht wissen.)
- Wie funktioniert Einfügen in den R-Baum, inklusive Split?
- Wie baut man einen Entscheidungsbaum auf?
- Wie kann man Overfitting beim Aufbau eines Entscheidungsbaums berücksichtigen?

- Wie kann Aufbau des Entscheidungsbaums berücksichtigen, dass unterschiedliche Fehlerarten unterschiedlich schlimm sind?
- Was ist die „10-fold cross validation“?
- Wie haben wir die Erfolgsquote definiert?
- Was ist ein Lift Chart? Wie unterscheidet es sich von der ROC Kurve?
- Was für Fehlerarten gibt es bei Vorhersagen von Klassenzugehörigkeiten?
- Was für Kennzahlen kennen Sie, die diese Fehlerarten sämtlich berücksichtigen?
- Was ist Unterschied zwischen Kovarianz und Correlation Coefficient?
- Warum kommt bei der informational loss Funktion die Logarithmusfunktion zur Anwendung?
- Was sind Association Rules?
- Wie findet man sie?
- Wie überprüft man rasch für viele Transaktionen, welche Kandidaten sie enthalten?
- Geben Sie ein Beispiel für eine Association Rule mit hohem/niedrigem Support und hoher/niedriger Confidence.
- Was sind multidimensionale Association Rules?
- Was sind Multi-Level Association Rules, und wie findet man sie?
- In welchen Situationen ist Apriori teuer, und warum?
- Was kann man gegen diese Schwächen tun?
- Was sind FP-Trees, und wie lassen sie sich für die Suche nach Frequent Itemsets verwenden?
- Was kann man tun, wenn FP-Trees für den Hauptspeicher zu groß sind?
- Was ist Constraint-basiertes Mining? Was sind die Vorteile?
- Was für Arten von Constraints kennen sie? Beispiele hierfür.
- Was ist Anti-Monotonizität, Succinctness? <Für ein bestimmtes Constraint sagen/begründen, ob anti-monoton/succinct.>
- Wie lässt sich Apriori für das Mining von Teilfolgen verallgemeinern?
- Antagonismus von Support-basiertem und Constraint-basiertem Pruning erklären können.

- Alternativen für Constraint-basiertes Pruning (wenn Constraint nicht anti-monoton) erklären können.
- Welche Clustering-Verfahren kennen Sie?
- Gegeben Szenario X, welche Clustering-Verfahren sind sinnvoll, und warum?
- Erklären Sie Clustering-Verfahren XY.
- Warum funktionieren herkömmliche Clustering-Verfahren in hochdimensionalen Merkmalsräumen nicht?
- Skizzieren Sie eine mögliche Lösung.
- Erklären Sie, warum Clustering mit kategorischen Attributen besonders ist? Warum ist Link-basiertes Clustering hier hilfreich?
- Welche Definitionen für Outlier kennen Sie?
- Gegeben die abstandsbasierte Definition von Outlier, welche Techniken zur Ermittlung der Outlier kennen Sie?
- Warum kann man beim Dichte-basierten Clustering nicht einfach die Dichte um die Punkte herum vergleichen und die mit der geringsten Dichte zurückgeben?
- Sehen Sie einen Zusammenhang zwischen Clustering und Outlier Detection?
- Welche Anomalien hochdimensionaler Merkmalsräume kennen Sie?
- Wieso funktionieren hierarchische Indexstrukturen in hochdimensionalen Merkmalsräumen nicht?
- Was ist der Zusammenhang zwischen der Zelldichte und Outlier Detection in hochdimensionalen Merkmalsräumen?
- Wie groß sollten die Zellen sein?
- Geben Sie die Klassifizierung aus der LV in 'interessante' und 'weniger interessante' Outlier wieder.
- What do we mean with ‚physically consistent models'?
- What is matrix completion?
- What are residuals?
- Describe the objective and the main ingredients of GLUE.
- Please reproduce the classification of approaches for theory-guided data science from this chapter.