Masterseminar Data Mining Themenliste

1. **Affinity Propagation: clustering objects via messages.**
   - Vorstellung des Clusterverfahrens *affinity propagation*.
   - Anwendung des Verfahrens auf verschieene Datensätze und
   - Vergleich mit verwandten Verfahren: k-means und k-medoids.
2. **Claims - Versicherungsdaten**
   - gewinnbringende und gewinnverringernde Kunden einer Versicherungsgesellschaft
   - Clusteringverfahren und deren Vergleich sind eine mögliche Anwendung
3. **Don't get kicked**
   - Daten aus einem DM-Wettbewerb
   - Ziel ist es, vorherzusagen, ob ein Auto ein "bad-buy" ist, d.h. ob es wiederverkauft werden kann oder nicht.
4. **Forest Covertype**
   - Anhand verschiedener Merkmale wie des Bodentyps soll die Art des Bewuchses von $30 \times 30\ cm^2$ Parzellen bestimmt werden.
   - Großer Datensatz zur Mehrgruppenklassifikation.
5. **Credit**
   - Datensatz zur Bestimmung der Kreditwürdigkeit verschiedener Personen
   - Aufgrund des kleineres Umfangs bietet sich der Vergleich verschiedener Verfahren an.
6. **Movies: IMDB vs. MovieLens**
   - Daten zu Spielfilmen auf den Webseiten `http://www.imdb.com` und `http://movielens.umn.edu`.
   - Schwerpunkt ist hier die explorative Datenanalyse und die Veränderungen im Zeitverlauf.
   - Die beiden Datensätze bieten auch Grundlage für Vergleiche.

7. **Plants**
    - Zu ca. 35.000 Pflanzen ist bekannt, in welchen Staaten (USA/Kanada) sie vorkommen.
    - Hier sind Matrixviuslisierung (heatmaps) sowie und Clustering und Sortierungsverfahren relevant.
    - Es können zusätzliche klimatische Daten und ggf. Kartenkoordinaten zu den Staaten herangezogen werden.
8. **Wine Quality**
    - Basierend auf verschiedenen Messgrößen wie zB dem Alkoholgehalt oder dem pH-Wert gilt es Rot- und Weiweine zu unterscheiden und insbesondere die Qualitätsstufe zu bestimmen.
9. **Zehnkampf**
    - Daten zu den Ergebnissen von Sportlern im Zehnkampf.
    - Interessant ist hier unter anderem die Umrechnung der Daten in ein Punktesystem und die Dimensionsreduktion: Gibt es gemeinsame Faktoren, die die Leistungen der Sportler erfassen können?
10. **Sonographie**
    - Ein Datensatz zu Geburten im Augsburger Zentralklinikum.
    - Von Interesse sind die Vorhersage des Kopfumfangs des Kindes mittels der Sonographie und der Einfluss auf die Umstände der Geburt.
11. **Typisierung von Fahrzeugen**
    - Hier geht es um die Typisierung von Fahrzeugen anhand von Umrissmerkmalen unterschiedlicher Perspektiven, wie sie von Kameras erfasstw erden können.
    - Thematisch ist dieser Datensatz also in der Mehrgruppenklassifikation anzusiedeln.
12. **Immobilien in Boston**
    - Ziel ist die Prognose des durchschnittlichen Immobilienpreises in Abhängigkeit verschiedener Einflussvariablen.

13. **Missing Values**

- Dieser Vortrag beschäftigt sich mit dem Thema “Fehlende Werte” (Missing Values, MV)
- MV kommen in sehr vielen Datensätzen vor und werden auch in Ihrer Art unterschieden.
- Neben speziellen Visualisierungsmethoden geht es auch um sog. Imputationsverfahren, bei denen fehlende Werte geschätzt werden, um so z.B. eine Modellberechnung zu ermöglichen.