# Praktische Informatik

Gedächtnisprotokoll vom 02.03.2006

Prüfer: Prof. Seidl
Prüfling: Dennis Meichsner
Note: 1,3
Dauer: ca. 40 Minuten

Fächer: Data Mining Algorithms, Datenexploration, Logikprogrammierung

**S:** Guten Tag... Womit wollen Sie anfangen?

**D**: Data Mining oder Datenexploration

**Data Mining** (etwa 15 Minuten)

**S**: Was sind denn so die Aufgaben des Data Mining, bzw. was versteht man unter dem KDD-Prozess?

**D**: Daten sammeln, reinigen, fehlende Werte bearbeiten und danach Clustering, Association Rules, Classification...

**S**: Da gibt es doch Verfahren mit festem k. Und wofür steht das k?

**D**: k-Means und k-Medoid erklärt

**S**: Kann man immer beide anwenden?

**D**: Nein, für k-Means muss Mittelwert der Objekte definiert sein.

**S**: Wo ist das denn nicht der Fall?

**D**: Beispiele aufgezählt – keine kontinuierlichen Werte

**S**: Was würde denn k=1 oder k=n bedeuten? Mit welchem k fängt man denn so an?

**D:** Alle Daten in einem Cluster, jeder Datenpunkt ein Cluster, muss man "austesten"

**S**: Wie denn austesten? Wie kann man die Qualität eines k Messen?

**D**: Hier bin ich leider nicht direkt auf den Silhouetten Koeffizienten gekommen, kleinen Hinweis vom Herrn Seidl erhalten und den erklärt.

**S**: Ok, was unterscheidet denn so das Clustering von der Classification?

**D**: Beim Clustering will man Gruppen finden, bei der Classification Objekte einer Klasse zuordnen.

**S**: Wie kann man denn so eine Classifikation machen?

**D**: Den Aufbau eines Decision-Tree erklärt. Dann wollte Herr Seidl noch etwas über Pruning dieses

Baumes hören.

**S**: Was ist denn die NearestNeighbour-Classification?

**D**: Erklärt und ein Beispiel aufgemalt mit k=1 wird a gewählt, k=3 wird b gewählt, mit k=3 und Gewichtung wird wieder a gewählt... siehe Skript

**Datenexploration** (etwa 15 Min.)

**S**: Ok, dann kommen wir mal zur Datenexploration. Was haben wir denn da so für Ähnlichkeitsmodelle gehabt?

**D**: Bilder, Formen, Sequenzen,...

**S**: Wie kann man denn Bilder durch Farben vergleichen?

**D**: Pixelweise, aber nicht gut. Besser: Histogramme. Diese kurz erläutert. Die LP-Distanzen erklärt. Dann wollte er noch auf die Ähnlichkeiten zwischen den einzelnen Histogrammen hinaus. Distanzmatrix, da ansonsten kleine Farbverschiebung genauso schlecht wie ganz andere Farben.

**S**: Schreiben sie doch mal die Summenformel auf die man bei der Berechnung braucht wenn man die Ähnlichkeiten der Farben hat.

**D**: Tja, da habe ich ein wenig Hilfe gebraucht. Eigentlich war sie ja klar, aber ich habe einen blöden Fehler eingebaut den ich mit Herrn Seidls Hilfe aber behoben habe.

**S**: Wir hatten da so eine Earth Mover Distance. Was ist das und wo liegt der Unterschied bei den Gewichten?

**D**: EMD erklärt. Bei Histogrammdistanzen: Distanzmatrix hat einsen auf der Diagonalen, bei EMD sind es Nullen, da es sich hier nicht um Ähnlichkeit handelt sondern um Kosten von i nach j.

**S**: Wir haben es oft mit riesen Datenmengen zu tun. Was kann man da machen um es effizient zu halten?

**D**: Er wollte auf die PCA und Dimensionsreduktion hinaus, die ich dann erklärt habe.

**S**: Was sind eigentlich Momente?

**D**: Statistische Werte über die Daten...

**Logikprogrammierung** (etwa 10 Minuten)

**S**: Gut, dann wollen wir uns nun zuletzt noch der Logikprogrammierung widmen. Die ist ja deklarativ, was bedeutet das?

**D**: Nur die Fakten und Relationen angeben, Programm findet selber Lösung... Auch direkt Erklärt warum das ganze in Prolog nicht ganz erfüllt ist (Reihenfolge der Regeln).

**S**: Wie sieht das denn in Prolog aus. Wird also immer ein Ergebnis geliefert oder gezeigt dass es keines gibt?

**D**: Nein, unendliche Zweige im Baum → unendliche Berechnung

**S**: Kann man die denn nicht erkennen?

**D**: Nein geht nicht, ginge nur unter gewissen Einschränkungen – siehe Skript. Habe direkt drauf hingewiesen, dass in DataLog diese Eigenschaften erfüllt sind. (Skript)

**S**: Wie kann ich denn ausdrücken, dass etwas nicht gilt?

**D**: Simulation des **not** mit Hilfe des Cut erklärt und entsprechende Regel+Fakt aufgeschrieben. Dann die Arbeitsweise des Cut erläutert.
Als nächstes wollte Herr Seidel auf die ClosedWorldAssumption hinaus, die im Skript aber so gar nicht vor kam. Sie ist vielleicht besser aus der KI bekannt.

**S**: Was können Sie mir denn zu Differenzlisten sagen?

**D**: Erklärt wie sie aussehen, wie man mit ihnen Konkatenation zweier Listen durchführt. Darauf eingegangen, wie diese aussehen muss damit man auch nachher noch eine Differenzliste hat.

**S**: Gut, dann warten Sie jetzt bitte kurz draußen.

**Anmerkung**: Dieses Protokoll ist etwa eine Woche nach der Prüfung entstanden. Ich denke, dass ich alle wichtigen Fragen sinngemäß – nicht wörtlich – wiedergegeben habe. Aber es kann durchaus sein, dass kleine Nebenfragen zu den Gebieten nicht einzeln aufgeführt sind. Wer den Rest kann, kann diese aber auch leicht beantworten. Zu der Atmosphäre kann ich nur sagen, dass Herr Seidl ein sehr angenehmer Prüfer ist und einem bei kleinen Schwierigkeiten auch mit kurzen Hinweisen weiterhilft.

Bei Fehlern, Anmerkungen oder Fragen: mailto:corran@d-meichsner.de