Übung 6

Prof. Dr. Alexandros Nanopoulos, Christoph Freudenthaler
Wirtschaftsinformatik und Maschinelles Lernen (ISMLL)
Universität Hildesheim

Abgabe: Juli 2009

Aufgabe 1: Klassifikation (20 Punkte) - Welche Kreditkarte darf es denn sein?

Aufgrund Deiner tollen Arbeit hast Du Dein Ziel erreicht und bist nun optimal sozialversichterter Angestellter bei der Firma *FoodMart Inc.*. Nach der OLAP-Analyse sollst Du nun anhand der zur Verfügung stehenden Daten und Sql Server 2008 herausfinden, welchen Kunden welchen Kundenkartenstatus besitzen. Analysiere die Kundendaten dahingehend, welche kundenindividuellen Parameter wie z.B. Bildung, Geschlecht,... den Status ihrer *Member Card (Golden, Silver,...)* bestimmen. Danach erstelle mit den drei Modellklassen *Entscheidungsbäume*, *Neuronale Netze* und *Naive Bayes* Vorhersagemodelle, die diese Membercard-Stufe möglichst präzise vorhersagen. Für die Evaluation der Modelle verwende eine Aufteilung in Test- und Trainingsdaten im Verhältnis 3:7 und Cross-Validation. Beantworte dann folgende Fragen:

- Welche Parameter scheinen Einfluss zu haben und vor allem wie beeinflussen sie den Membercard-Status?
- Warum verwendet man Cross-Validation?
- Wie unterscheiden sich die drei verwendeten Modellklassen hinsichtlich
  1. Vorhersagegenauigkeit?
  2. Interpretierbarkeit der Resultate?
  3. Schwierigkeitsgrad der Modellparametersuche?
  4. Zugänglichkeit für nominale, ordinale und stetige Variablen (als Ziel- sowie als Inputvariable)?
  5. Geschwindigkeit der Berechnung?

Aufgabe 2: Clustering (20 Punkte) - Kosten senken

Als nächstes gibt Dir Dein neuer Chef die Aufgabe die Produktpalette dahingehend zu untersuchen, ob es in der Gesamtmenge der angebotenen Produkte größenmäßige Übereinstimmungen gibt, sodass in Zukunft ähnlich große Produkte in dann standardisierten Regalen dem Kunden angeboten werden könnten. Seine famose Idee ist nämlich ein neues Store-Konzept, wonach die Produkte nicht mehr nach Kategorie angeordnet werden, sondern nach ihrer Größe. Das hätte neben einem Kostenvorteil durch standardisierte Regale den Vorteil, dass die Marke *FoodMart Inc.* unverwechselbarer werden würde.
Untersuche daher anhand der Attribute *shelf-width*, *shelf-height* und *shelf-depth* folgende Aufgabenstellung:

- Welche Produkte sind größenmäßig ähnlich und in Clustern zusammenfassbar?
- Wie viele Cluster hast Du gewählt und warum?
- Versuche die Einteilung der Produkte in die von Dir gefundenen Cluster zu verstehen. Wie kommt es zu diesem Ergebnis? Unterscheide zwischen inhaltlichen und Modellierungsgründen.

Aufgabe 3: Assoziationsanalyse (20 Punkte) - Umsätze erhöhen

Als letzte Aufgabe vor Deinem großen Sommerurlaub hat Dein neuer Chef noch eine letzte Bitte. Aufgrund der Wirtschaftskrise und der dadurch gesunkenen Erträge soll er nächste Woche dem Vorstand ein Konzept vorlegen, wie neben den Kostensenkungen auch die Umsätze in Zukunft wieder erhöht werden könnten. Er dachte neben der unkonventionellen Art die Produkte nach Größe zu sortieren und so Kosten zu sparen auch an eine Ausdehnung der Umsätze durch ein innovativeres Produktangebot. Produkte, die häufig zusammen gekauft werden, sollen möglichst weit voneinander entfernt im Store zu finden sein, sodass der Kunde so oft wie möglich den Laden durchqueren muss und so die Chance steigt, dass er das ein oder andere zusätzlichen Produkt erwirbt. Deshalb sollst Du noch rasch eine Assoziationsanalyse durchführen und jene Produkttupel finden, die häufig gemeinsam gekauft werden. Beantworte danach folgende Fragen:

- Welche Zusammenhänge hast Du gefunden?
- Welche Support- und Konfidenzgrenzen hast Du gewählt, um die Produkttupel zu finden?
- Wie beeinflusst die Wähl der beiden Grenzen, die Geschwindigkeit mit der die Berechnungen in SQL Server terminieren?
- Warum ist es wichtig beide Grenzen zu betrachten und sich stattdessen nicht nur auf eine Grenze zu konzentrieren?