Zwischenpräsentation – Data Mining Cup 2010
Gruppe 1

David, Philippe, Michael,
Alexander, Thomas M, Elvi

INSTITUTE FOR PROGRAM STRUCTURES AND DATA ORGANIZATION, FACULTY OF INFORMATICS

Gliederung

- Vorverarbeitung
  - Typänderungen
  - Bereinigungen
  - Ableitungen
- Getestete Klassifikatoren
  - Zahlen
- Probleme
- Ausblick

Vorverarbeitung - Typen

- Flag:
  - title
  - newsletter
  - delivertype
  - voucher
  - gift
  - entry
  - points
  - shippingcosts
  - target90

Vorverarbeitung - Typen

- Set:
  - salutation
  - domain
  - model
  - paymenttype
  - invoice- & deliv- postcode
  - advertisingdatacode

Vorverarbeitung - Typen

- Sorted Set:
  - case
  - verspätungsklasse

Vorverarbeitung - Attributfilter

- customernumber
- Deliverytype (bei michael weg)
- Invoicepostcode (bei michael weg)
- delivpostcode
- points

Vorverarbeitung – Bereinigung

- 9 Datensätze mit deliverydatepromised

  Jahr 4746, dann auf deliverydatereal
- deliverydatereal == null:

  “download oder alle Waren storniert”, dann auf deliverydatepromised
- Kleine Bereinigungen von Inkonsistenzen wie
  - numberItems<cancel
  - wi = 0, aber numberItems nicht 0

Vorverarbeitung – Ableitungen I

- Datumsfelder zu Integer in Tagen nach 01-01-2008 (teilweise)
- Verspätung in Tagen als neues Attribut
- Anzahl Tage zu Feiertagen (Ostern/Weihnachten)
- Flag, ob deliverpostcode == invoicepostcode
- Flag, ob alle Items einer Bestellung storniert

Vorverarbeitung – Ableitungen II

- Wenn deliverydatereal zu früh (z.B. 250 Tage), dann reduziere deliverydatepromised um 365d
  - Annahme Tippfehler deliverydatepromised um ein Jahr zu groß
- Verspätungsklassen:
  - Kategorisieren der Verspätungen/Verfrühungen
  - 5 ordinale Klassen
- Kategorisieren der Mail-Domains in 3 Klassen
  - Willkürlich!
- advertisingdatacode in Flag konvertiert

Klassifikatoren

- Support Vector Maschine
  - Mit Standardeinstellungen wird alles als 0 klassifiziert
- Neuronales Netz
  - Standardwerte -> Score: 9300
  - Alles als 0 klassifiziert => alle Kunden bekommen einen Gutschein
- Entscheidungsbaum mit C 5.0 Algorithmus
  - Partitionierung der Trainingsdaten 90/10 in Trainings-/Testdaten
  - Maximal erreichte Score: ca. 12160 (kreuzvalidiert)
- Automatischer Klassifizierer (C5+Netz, gewichtetes Voting)
  - Ca. 10.000

Probleme

- Kaum Domänenwissen
- Lange Ausführungszeiten beim klassifizieren
- Verbesserung bei Hinzu-/Wegnahme intuitiv (un)wichtiger Attribute beim Klassifizieren
- Score der Kreuzvalidierung schwankt um ca. **600** Punkte bei verschiedenen Ausführungen
- Behalten der Übersicht über abgeleitete Daten von abgeleiteten Daten

Ausblick

- Weitere Klassifikatoren testen
- Automatische Klassifikatoren testen
- Knime einbeziehen
- Mehrere Klassifikatoren mit Hilfe der *confidence* kombinieren (Mehrheitsentscheider)
  - Test der einzelnen Klassifikatoren im Voraus
  - …rechnet gerade…

Fragen?