**Data Mining Cup 2005**
**Ilja Bezrukov**

# 1 Vorgehensweise

Um Overfitting zu vermeiden, haben wir als erstes die Trainingsdaten zufällig in zwei Teile getrennt: ein Trainingset mit 23910 Datensätzen und ein Validierungsset mit 6090 Datensätzen.
Wir testeten verschiedene Möglichkeiten der Featureerzegung und -vorverarbeitung. Als Klassifikator wurde aufgrund der Geschwindigkeit hauptsächlich logistische Regression verwendet. Wenn sich eine Featureauswahl als vielversprechend herausstellte, wurden weitere Experimente mit anderen Klassifikatoren durchgeführt.
Alle Experimente wurden in 5-fach-Kreuzvalidierung auf dem Trainingsset durchgeführt. Kurz vor dem Abgabetermin wählten wir die besten 10 Setups aus und testeten deren Performance auf dem Validierungsset. Wenn ein Setup auf dem Validierungset einen deutlich schlechteren Score erbrachte, als auf dem Trainingsset, so deutet das auf Overfitting hin.
Ein Setup umfasste die verwendeten Features, deren Vorverarbeitung und den Klassifikator. Das Setup mit dem besten Score auf dem Validierungs- und Trainingsset wurde für die Klassifikation der Testdaten benutzt.

# 2 Datenvorverarbeitung

Unsere bisherige Erfahrung hat gezeigt, dass die Vorverarbeitung der Daten sehr wichtig ist, sogar wichtiger als die Auswahl des Klassifikators. Wir haben deshalb intensiv verschiedene Möglichkeiten der Vorverarbeitung getestet. Wir haben neue Features erzeugt und bereits vorhandene mit Hilfe von Histogrammen und binären Features transformiert.
Die in den Daten enthaltenen binären Features haben wir ohne weitere Transformation übernommen.
Die folgenden Features verdienen besondere Beachtung, weil sie sich nicht ohne weiteres mit den Standardverfahren verarbeiten liessen:
Das Feature `TIME_BEST`, das die Uhrzeit der Bestellung enthielt, haben wir in die Anzahl der Minuten seit Mitternacht umgewandelt und mit Histogrammen weitertransformiert.
Aus den `ANUMMER`-Features, welche die Artikelnummern von bis zu 10 gekauften Artikeln enthielten, haben wir folgende neue Features erzeugt:

- 6 Features für die Anzahl der gekauften Artikel der Gruppen 1-6. Wir sind davon ausgegangen, dass die ersten beiden Ziffern der Artikelnnummer die Gruppe räpresentieren.
- 9 Features für die Gruppe des Artikels im Warenkorb. In den Trainingsdaten hat kein Kunde mehr als 9 Artikel gekauft.
- Für jede Artikelgruppe 1 Feature, welches die Nummer des Artikels innerhalb dieser Gruppe enthielt. Die Nummern wurden für jede Gruppe neuberechnet.

- 6 Features für die Anzahl der gekauften Artikel in selbsterzeugten Gruppen. Die Gruppen wurden so gewählt, dass auf den gesamten Daten sich in jeder Gruppe in etwa gleichviele Artikel befanden.

Die weitere Verarbeitung erfolgte, wie in den folgenden Abschnitten beschrieben, mit Hilfe von binären Features und Histogrammen.

## 2.1 Binäre Features

Durch Erzeugung von binären Features können bestimmte Merkmalausprägungen hervorgehoben oder verallgemeinert werden. Kategorische Features können durch Dimensionsersetzung in binäre Features überführt werden.
Wenn Features fehlende Werte enthielten, haben wir ein enstsprechendes binäres Feature erstellt. Bei kategorischen Features wie der Zahlungsmethode erzeugten wir für wichtige Ausprägungen binäre Features.
Bei numerischen Features wie dem Geburtsjahr und dem Bestellwert haben wir den Wertebereich in Intervalle eingeteilt und für die Intervalle binäre Features erzeugt.

## 2.2 Histogramme

Für die Transformation der ANUMMER-Features haben wir Equi-Depth Histogramme mit 10 und 20 Bins verwendet [1]. Dabei wurde ein Equi-Depth-Histogramm des Merkmals erstellt. Jedes Datum eines Merkmals wurde einem Bin zugewiesen, und die Bin-Abstände wurden so angepasst, dass jeder möglichst gleichviele Einträge aufweist.
Jedem Merkmalswert wurde anschliessend die [0,1]-normierte Mitte des zugehörigen Bins zugewiesen.
Neben Equi-Depth-Histogrammen haben wir die Verwendung von Quantilhistogrammen getestet. Dabei wurde jeder Merkmalswert ersetzt durch den Anteil der Werte, die kleiner als dieser sind.
Die Equi-Depth-Histogramme brachten im Allgemeinen bessere Ergebnisse, so dass wir in den für die Abgabe erzeugten Modellen die Quantilhistogramme nicht verwendet haben.

# 3 Verwendete Klassifikatoren

Wir haben verschiedene Klassifikationsverfahren getestet, die alle eine ähnliche Performance aufwiesen: Logistische Regression, neuronale Netze, Naive Bayes+Maximum Entropy.
Für logistische Regression und neuronale Netze verwendeten wir die Implementierungen von Netlab [4], Naive Bayes und Maximum Entropy waren Eigenimplementierungen [3] [2].
Zusätzlich kombinierten wir die Ergebnisse der Klassifikatoren, um die Gesamtperformance zu verbessern.
Zum einen kombinierten wir die Ergebnisse der Klassifikatoren mit Hilfe der Summenregel. Dabei wurden die Posteriori-Wahrscheinlichkeiten der Klassifikatoren aufsummiert und die Klasse mit der maximalen Summe als Klassifikationsergebnis gewählt.
Zum anderen verwendeten wir eine Kombination aus Naive Bayes und Maximum Entropy. Hierbei wurden keine Annahmen über die Verteilung der Features gemacht, und die

Wahrscheinlichkeiten der Feature-Ausprägungen aus den relativen Frequenzen geschätzt. Ungesehene Werte im Test wurden durch ihren nächsten Nachbar ersetzt.
Die Verteilungen der einzelnen Features wurden anschliessend mit Hilfe des Maximum-Entropy Klassifikators gewichtet.
Die Kombination Logistische Regression, Neuronale Netze, Naive Bayes+Maximum Entropy erreichte im Wettbewerb mit 12297 Punkten den ersten Platz.

# Literatur


[1] I. Bezrukov, T. Deselaers, A. Hegerath, D. Keysers, and A. Mauser. Gfkl data mining competition 2005: Predicting liquidity crises of companies part i: Data preprocessing. In *Proceedings GfKl 2005*, page 152, Magdeburg, Germany, 2005.

[2] A. Mauser, I. Bezrukov, T. Deselaers, and D. Keysers. Predicting customer behavior using naive bayes and maximum entropy – winning the data-mining-cup 2004. In *Proceedings Informatiktage 2005.*, St. Augustin, Germany, 2005.

[3] A. Mauser, D. Keysers, A. Hegerath, T. Deselaers, and I. Bezrukov. Gfkl data mining competition 2005: Predicting liquidity crises of companies part ii: Training and classification. In *Proceedings GfKl 2005*, page 153, Magdeburg, Germany, 2005.

[4] I. T. Nabney. *Netlab. Algorithms for Pattern Recognition*. Advances in Pattern Recognition. Springer-Verlag Telos, 1st edition, 2001.