**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
PD Dr. Matthias Schubert
Erich Schubert

# Knowledge Discovery in Databases II

WS 2013/2014

## Übungsblatt 11: Verteiltes und Paralleles Data Mining

### Aufgabe 11-1 Effiziente Cosinus-Ähnlichkeit für parallele Systeme

Die Cosinus-Ähnlichkeitsfunktion wird üblicherweise definiert als:

$$\cos(\varphi) := \frac{x \cdot y}{||x|| \cdot ||y||}$$

Der Winkel $\varphi$ kann auch als eine Pseudo-Distanz verwendet werden.

Besonders beliebt ist diese Distanzfunktionen auf Textdaten, die typischerweise hochdimensional sind, aber auch dünnbesetzt. Wird als Vorverarbeitung der Vektor normalisiert so dass $||v|| = 1$ gilt, so vereinfacht sich diese Formel weiter zu:

$$\cos_{\text{norm}}(\varphi) = x \cdot y = \sum_{i=0}^{n} x_i y_i$$

(a) Wie komplex ist die Berechnung dieser Distanzfunktion, wenn die Vektoren $x$ und $y$ beide dünnbesetzt sind, insbesondere im Vergleich zur Euklidischen Distanz, und wir sehr viele Dimensionen haben?

(b) Angenommen nur $x$ ist dünnbesetzt, $y$ aber (bspw. als Centroid) ist dichtbesetzt. Wie wirkt sich das auf den Berechnungsaufwand aus?

(c) Wir möchten die paarweisen Ähnlichkeiten einer großen Datenbank berechnen. Dazu transponieren die Daten und arbeiten komponentenweise (beispielsweise mit Hadoop). Welchen Vorteil bietet diese Art der Berechnung?

(d) Ein ähnlicher Trick ist auch für Euklidische Distanzen auf dünnbesetzten Vektoren möglich. Hierzu nutzt man die zweite binomische Gleichung aus: $(a - b)^2 = a^2 - 2 \cdot a \cdot b + b^2$. Überlegen Sie sich, wie man hier diese Formel anwendet.

### Aufgabe 11-2 Privacy Preservation in Standard-Klassifikatoren

Gegeben seien folgende vier Klassifikatoren: Entscheidungsbäume, Nächste-Nachbarn-Klassifikation, Support-Vector-Machines und Naive-Bayes.

- Untersuchen Sie die Frage, ob bereits trainierte Klassifikatoren an Dritte weitergegeben werden dürfen, ohne dass dabei Teile der Trainingsmenge offengelegt werden!
- Wie könnte man versuchen, etwaige Probleme bei den einzelnen Klassifikatoren zu lösen?

### Aufgabe 11-3 Parallele Assoziationsregeln

Überlegen Sie sich die Vorteile und Nachteile einer horizontalen und einer vertikalen Verteilung bei der nebenläufigen Berechnung von Assoziationsregeln!

### Aufgabe 11-4 Parallele Naive Bayes Klassifikation mit Map Reduce

Beschreiben Sie ein Programm, das mit Hilfe einer parallelen Verarbeitung in MapReduce alle notwendigen Wahrscheinlichkeiten für einen Naive Bayes Klassifikator berechnet.

Gehen Sie dabei davon aus, dass die einzelnen Klassen über multivariate achsenparallele Normalverteilungen modelliert werden und die Trainingsmenge $D$ durch Tupel der Form $< ID, object >$ gegeben sind, wobei object über die Felder $c$ und $v$ verfügt. Dabei ist $ID$ der Schlüssel des Objekts, $c \in C$ beschreibt die Klasse und $v \in \mathbb{R}^d$ einen Featurevektor.

Geben Sie hierzu eine Funktion für den Mapper und eine Funktion die den Reducer beschreibt in PseudoCode wieder.