Prof. Dr. Morik Dortmund, 4. Dezember 2013

Übung zur Vorlesung

Maschinelles Lernen

Wintersemester 2013/2014
Blatt 7

Das RapidMiner *Text-Plugin*[1] bietet eine Reihe von Operatoren, um unterschiedliche Text-Formate (ASCII, PDF, HTML, usw.) in Wort-Vektoren zu überführen.
Fügen Sie das Plugin zu Ihrer RapidMiner-Installation hinzu und erstellen Sie Experimente zu folgenden Aufgaben. Geben Sie als Ergebnis jeweils die Experimente sowie eine Tabelle mit den Performanz-Vektoren der Experimente an.

Aufgabe 1 — 5 Punkte

Auf der Übungsseite finden sie den Datensatz "Newsgroups", der News-Artikel aus unterschiedlichen Gruppen enthält. Der Datensatz enthält für jede Gruppe ein Verzeichnis, in dem die Artikel abgelegt sind.
Trainieren Sie einen Klassifizierer für die Zuordnung von Artikeln zu Newsgruppen und evaluieren Sie das verwendete Lernverfahren jeweils mit einer Kreuzvalidierung. Nutzen Sie als Klassifizierer je

1. Lineare Regression
2. Naive Bayes
3. SVM

Experimentieren Sie auch mit der Erzeugung der Wort-Vektoren (z.B. TF/IDF vs. Term-Occurrence) und nutzen Sie zum Vergleich zwei eigene Implementierungen von Lernverfahren aus den vorangegangenen Übungen.

Aufgabe 2 — 5 Punkte

Im Laufe Ihres Studiums haben sich sicherlich eine Menge von Artikeln, Wikipedia-Ausschnitten und anderen Materialien in digitaler Form angesammelt. (Falls nicht, können für die folgende Aufgabe auch das eigene EMail-Postfach, Vorlesungsskripte oder ähnliches herhalten.)
Erstellen Sie eine einfache (flache) Kategorisierung von gesammelten Materialien und legen sie eine Ordnerstruktur dazu an (für den wahrscheinlichen Fall, dass sie eine derartige Struktur bereits haben, können Sie diese natürlich auch verwenden).

1. Trainieren Sie auch auf dieser Struktur die SVM zur Klassifizierung von Dokumenten (vgl. Aufgabe 1).
2. Wenden Sie das Modell auf neue Dokumente an - passt die Klassifizierung zu Ihrer Ordnungsvorstellung?

[1] Das Text-Plugin ist verfügbar unter `http://rapid-i.com` → Downloads → RapidMiner Plugins oder über den Marketplace in RapidMiner.