Intelligente Datenanalyse in Matlab

Prüfung: Projektaufgabe 11

Michael Brückner
Thomas Vanck
Prof. Tobias Scheffer

Datum: 07.12.2010

Die Projektaufgabe ist Teil der Prüfung *Intelligente Datenanalyse in Matlab*. Jede Aufgabe soll durch einen Studenten selbstständig bearbeitet und die Lösung innerhalb der mündlichen Prüfung vorgestellt werden. Ein Ausdruck des MATLAB-Programmcodes und der Ergebnisse in Form eines Diagramms, Tabelle o.ä. werden vorausgesetzt; die Art der Präsentation der Ergebnisse ist dem Studenten freigestellt.

Problemstellung

Sie wurden von der IT-Abteilung eines mittelgroßen Unternehmens damit beauftragt einen Email-Spamfilter zu trainieren welcher die eingehenden Emails aller Mitarbeiter als Spam oder Nicht-Spam markieren soll. Die Emails werden dabei von einem Modul geparst und in die Bag-of-Words-Darstellung konvertiert. Insgesamt werden dabei 57.173 verschiedene Wörter (Attribute) unterschieden. Ziel des Filters ist es maximal viele Spam-Emails zu identifizieren wobei höchstens 0,2% aller legitimen Emails falsch klassifiziert werden dürfen. Außerdem möchte das Unternehmen eine Aussage über die Effektivität des Filters auf zukünftigen Emails, d.h. wie viel Prozent eingehender Spam-Emails zukünftig erkannt werden.

Aufgabe

Aus den Inboxen der Mitarbeiter wurden zum Trainieren 10.000 markierte Emails extrahiert (siehe emails.mat). Dabei seien X die Lerndaten mit den zugehörigen Klassenlabeln Y ($+1$ steht für Spam, -1 bedeutet Nicht-Spam). Identifizieren sie ein geeignetes Lernverfahren zur Konstruktion eines Spam-Filters. Implementieren sie dieses Verfahren in MATLAB. Bestimmen sie die Modellparameter; trainieren und evaluieren sie das Modell. Treffen sie eine Aussage über die zu erwartende Qualität des Filters und stellen sie sicher, dass nicht mehr als 0,2% aller legitimen Emails gefiltert werden. Plotten sie zu diesem Zweck eine Precision/Recall-Kurve und tragen sie das Modell (für den gewählten Schwellwert) ein. Begründen und dokumentieren sie kurz alle durchgeführten Schritte.