Praktikum Data Mining

Dokument Klassifikation / Spam Filter

Oliver Fesseler Maria Florusß Stefan Seibert

Daniel Grießhaber

17. Juni 2014

Durchführung

Implementierung: Dokument Klassifikation

Was wird mit Evidenz bezeichnet und warum muss diese für die Klassifikation nicht berücksichtigt werden?

Die Evidenz p(x) ist im Beispiel die Wahrscheinlichkeit dafür, dass das Wort x überhaupt vorkommt unabhängig davon, welcher Klasse die Dokumente, in denen es vorkommt, zugeordnet werden. Deshalb ist die Evidenz für alle Klassen gleich und muss für die Entscheidung nicht miteinbezogen werden.

Der eigentliche Wahrscheinlichkeitswert, mit der ein Dokument einer Klasse zugeordnet wird ist nicht relevant. Es ist nur relevant, für welche Klasse der Wert am größten ist. Deshalb ist es nicht wichtig, alle diese Werte noch einmal durch den selben Wert (die Evidenz) zu teilen.

Wann würden Sie in der Formel für die gewichtete Wahrscheinlichkeit den Wert von `initprob` kleiner, wann größer als 0.5 wählen? (Falls Sie die Möglichkeit haben diesen Wert für jedes Feature und jede Kategorie individuell zu konfigurieren)

- Wenn viele Dokumente eingelesen werden, die mit hoher Wahrscheinlichkeit viele unbekannte Worte enthalten sollte der Wert von `initprob` möglichst klein gewählt werden. Ansonsten würde ein Dokument eher einer Klasse zugeordnet, die viele seiner Worte nicht kennt, da diese einen relativ großen Wert zugeschrieben bekommen.
- Der Wert sollte auch dann niedirg gewählt werden, wenn die bekannten Worte relativ kleine Wahrscheinlichkeitswerte aufweisen. Wenn alle Worte im Durchschnitt nur mit einer Wahrscheinlichkeit von 0.1 vorkommen und für nicht bekannte Worte das `initprob` auf 0.5 gesetzt ist, kann dies das Ergebnis erheblich verfälschen.
- //TODO: wann initprob größer als 0.5?

Was könnten Sie mit dem in dieser Übung implementierten Classifier noch klassifizieren? Geben Sie eine für Sie interessante Anwendung an.

- Spamfilter o.ä.

Das einmal trainierte, sollte eigentlich persistent abgespeichert werden. Beschreiben Sie kurz wie Sie das für dieses Programm machen würden.

- Gelernte Daten serialisieren, in File speichern und bei Gebrauch laden.