Datamining - Versuch 4

Lukas Spieß, Mathis Hoffmann

25. November 2013

3 Fragen zum Versuch

1. *Was wird mit Evidenz bezeichnet und warum muss diese für die Klassikation nicht berücksichtigt werden?*

   Als *Evidenz* wird im Fall der Dokumentenklassifikation die Wahrscheinlichkeit für das Auftreten eines Wortes bezeichnet. Sie ist nicht relevant, weil sie für alle Klassen gleich groß ist.

2. *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 kongurieren)*

   Generell lässt sich der Einfluss des Wertes von *initprob* auf das Ergebnis wie folgt beschreiben: Ein großer Wert für *initprob* hat zur Folge, dass Dokumente mit vielen unbekannten Wörtern einen relativ hohen Wahrscheinlichkeitswert für eine Klasse bekommen. Ein niedriger Wert bewirkt genau das Gegenteil, also dass Dokumente mit vielen unbekannten Wörtern einen sehr niedrigen Wahrscheinlichkeitswert bekommen.

   Es lassen sich also durch gezielte Manipulation des Wertes bei der Wahrscheinlichkeitsberechnung für einzelne Kategorien, diese bei der Klassifikation von Dokumenten mit vielen unbekannten Wörtern gezielt unter- bzw. überbewerten. So lässt sich erreichen, dass Dokumente mit vielen unbekannten Wörtern vorzugsweise in eine definierte Klasse fallen (hoher Wert für *initprob* für diese Klasse) oder ihr herausgehalten werden (niedriger Wert).

3. *Was könnten Sie mit dem in dieser Übung implementierten Classifier noch klassizieren? Geben Sie eine für Sie interessante Anwendung an.* Eine offensichtliche weitere Anwendungsmöglichkeit für diesen Classifier wäre natürlich die Spamerkennung, wobei man hier praktisch gesehen nie an die Systeme großer Organisationen herankommen wird, da diese mit deutlich größeren Lerndatensätzen arbeiten können. Ein realistischerer, praktischer Anwendungsfall, der von uns auch in Ansätzen implementiert wurde, ist die tatsächliche Kategorisierung von RSS Feeds in mehrere Kategorien. Eine reine Unterteilung in *Tech* und *NonTech* ist auf Dauer wenig sinnvoll, eine Unterteilung in verschiedene Nachrichtenthemengebiete aber durchaus spannend und nützlich.