25. November 2013

DATA MINING

- Versuch 4: Dokument Klassifikation / Spam Filter -

Teammitglieder

JanHorak

Ralf Palyov

Andreas Mayer

3 Fragen zum Versuch

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

Um zu berechnen mit welcher Wahrscheinlichkeit ein Dokument einer Klasse zugeordnet wird, wird die Bayes-Formel angewandt. Die Evidenz ist ein Teil dieser Formel und steht hier für die Wahrscheinlichkeit, mit welcher ein Dokument vorkommt. Genauer werden die Wahrscheinlichkeiten, dass jedes einzelne Wort in einer bestimmten Klasse vorkommt und, dass diese Klasse vorkommt (a-priori) multipliziert und das Ergebnis durch die Evidenz geteilt.

In unserem Anwendungsfall werden nur die Wahrscheinlichkeiten für ein und das selbe Dokument, aber unterschiedliche Klassen verglichen. Dadurch ist die Wahrscheinlichkeit für das Dokument (Evidenz) überall gleich und kann einfach ignoriert werden.

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 konfigurieren)

Dieser Wert wird bei Worten gewählt die im gesamten Trainingskorpus kein einziges mal vorkommen. Steht er auf 0.5, wird die Wahrscheinlichkeit, dass dieses Wort zu jeder beliebigen Klasse gehört auf 50% gesetzt. Wenn der Klassifizierer in zwei Klassen einteilt macht dieses Vorgehen auch Sinn. Sind allerdings mehrere Klassen zu unterscheiden, ist es besser, wenn die Summe aller Wahrscheinlichkeiten für dieses Wort und Klassen 100% ergibt. Initprob berechnet sich dann also:

$$initprob = 1 / Klassenanzahl$$

Ansonsten ist das Vorgehen der gewichteten Wahrscheinlichkeit aus meiner Sicht gut genug um alle Fälle abzudecken. Je öfter das Wort im Trainingskorpus vorhanden ist, desto geringer wird der Einfluss von initprob. Dies führt zu einer recht ausgeglichenen Verteilung bei relativ unbekannten Worten und einer Verteilung die fast nur noch auf den Trainingsdaten basiert bei Worten die häufig gelernt wurden.

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

Der implementierte Classifier könnte unverändert auch für andere Dokumente und Klassen verwendet werden, wenn entsprechende Trainingsdaten erstellt werden. Ein Beispiel wäre eine Spracherkennung die den Dokumenten jeweils eine Sprache zuordnet. Diese Funktion könnte beispielsweise für das Sortieren von Nachrichten, die auf einer Webseite angezeigt werden, nützlich sein. Ebenfalls Sinnvoll wäre, in einem globalen Unternehmen eine automatische Weiterleitung von Mails an den Supportmitarbeiter der die in der Mail verwendete Sprache versteht einzurichten. Dies würde unterschiedliche Mail Adressen für die einzelnen Sprachen überflüssig machen.

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

Die Trainingsdaten liegen in unserem Programm in einem Objekt und sind für die Methoden des Objektes formatiert. Diese Daten extern zu speichern würde einen Mechanismus nötig machen um sie beim importieren wieder interpretierbar zu machen. Um dies zu umgehen, wäre die einfachste Möglichkeit das gesamte Objekt der Klasse docclass zu serialisieren, also in einen String umzuwandeln. Das Ergebnis könnte dann einfach in einer Textdatei gespeichert werden. Für diesen Vorgang gibt es ein Python Package namens Python Pickle. Wieder eingelesene Objekte können dann ganz einfach wie vor der Serialisierung verwendet werden.