Prof. Dr. Morik
Christian Bockermann

Dortmund, 20. Januar 2009

Übung zur Vorlesung

Maschinelles Lernen

Wintersemester 2008/2009
Blatt 12

Naive Bayes Klassifikation In den vorangegangenen Übungen beschäftigten sich einige Aufgaben mit der Text-Klassifikation. Ein Standard-Verfahren, welches direkt in der Vorlesung nicht behandelt wurde, ist die Klassifikation nach dem *Satz von Bayes*, die insbesondere bei SPAM-Filtern sehr verbreitet ist.
Dabei werden Attribute $A_1, \ldots, A_k$ von E-Mails erfaßt (z.B. A_j als Auftreten eines Wortes w_j) und die Mails mit einem (binären) Label C (z.B. *SPAM* oder *HAM*) versehen. Die Wahrscheinlichkeiten

$$P(C), P(A_i|C), P(A_i|\overline{C}) \ , i = 1, \ldots, k$$

werden über die Häufigkeiten in der Trainingsmenge "gelernt". Dabei geht man von der bedingten Unabhängigkeit der Wahrscheinlichkeitsmaße aus (daher *naive* Bayes).

Sei eine E-Mail M mit den Attribut-Werten $A_{m_1}, \ldots, A_{m_l}$ mit $\{m_1, \ldots, m_l\} \subseteq \{1, \ldots, k\}$ gegeben (z.B. durch auf $[0; 1]$ normierte Häufigkeit vorher gelernter Worte w_j). Mit Hilfe des Satzes von Bayes und der angenommenen bedingten Unabhängigkeit erhält man eine Wahrscheinlichkeit für die Klasse C für M als

$$P(C|A_{m_1} \cap \ldots \cap A_{m_k}) = \frac{P(C) \prod_{j=1}^{l} P(A_{m_j}|C)}{P(C) \prod_{j=1}^{l} P(A_{m_j}|C) + (1 - P(C)) \prod_{j=1}^{l} P(A_{m_j}|\overline{C})}$$

Dabei entspricht $\overline{C}$ dem Komplementär-Ereignis von C, also $\neg C$.
Der Bayesche Klassifikator liefert zu einem gegebenen Beispiel M dann die wahrscheinlichste Klasse

$$prediction(M) = \arg\max_{C} P(C|A_{m_1} \cap \ldots \cap A_{m_l}).$$

Aufgabe

10 Punkte

1. Implementieren Sie einen Operator `NaiveBayes`, der nach dem Satz von Bayes eine Menge von Beispielen mit numerischen Attributen klassifiziert.

2. Erstellen Sie ein Sammlung von Texten für die Klassen SPAM und HAM, und lesen diese mit dem `TextInput`-Operator in RapidMiner ein. Evaluieren Sie ihren Operator mit Hilfe einer Kreuzvalidierung über diesen Texten.

3. Vergleichen Sie ihren Operator mit dem bereits vorhandenen `NaiveBayes`-Operator.