

Maschinelles Lernen: Symbolische Ansätze



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Wintersemester 2008/2009

4. Projektaufgabe für den 13.1.2009

Entscheidungsbäume

- Wählen Sie 2 der 5 Datensätze, die Sie hier finden, aus. Vergleichen Sie für diese Datensätze die ROC-Kurven bzw. die Fläche unter diesen Kurven für die Klassifizierer J48 einmal mit und einmal ohne Pruning (Option 'unpruned') und ID3. Bei J48 verwenden Sie für die anderen Optionen die Default-Werte.
- Vergleichen Sie die Klassifizierer ebenfalls mit den Accuracy Werten der Cross-Validation.
- Betrachten Sie auch die Größe der entstandenen Bäume (Anzahl Knoten und/oder Blätter im Baum) und setzen Sie diese in Zusammenhang mit der Güte der Klassifizierer.

Nearest Neighbour

- Verwenden Sie für diese Aufgabe die gleichen Datensets wie in der vorherigen Aufgabe. Finden Sie heraus für welches $k \in \{1, 3, 5, 7, 9, 11\}$ der Algorithmus k-NN (in weka heisst der Algorithmus IBk; verwenden Sie auch hier die Default Optionen) die höchste Cross Validation Accuracy bekommt. Ist der Algorithmus für diesen Wert von k besser als die Entscheidungsbäume der vorherigen Aufgabe?