Maschinelles Lernen: Symbolische Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2009/2010
2. Projektaufgabe für den 12.1.2010

ROC-Kurven

- Wählen Sie einen der 5 Datensätze, die Sie hier finden, aus. Vergleichen Sie für einen ausgewählten Datensatz die ROC-Kurven bzw. die Fläche unter diesen Kurven für die Klassifizierer `J48` und `NaiveBayes`. Sie können die ROC-Kurven betrachten, indem Sie mit der rechten Maustaste im Fenster "Result List" den Menü-Punkt "Threshold List" auswählen.
- Interpretieren Sie die Resultate. Sie können die Werte, die zum Zeichnen der Kurve verwendet wurden, auch mit "Save" in ein ARFF-File exportieren, und dieses (nach Löschen des Headers) in Grafik-Programme importieren. So können Sie z.B. beide Kurven (für `J48` und `NaiveBayes`) übereinander legen.

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 (Datensets). 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?

Regressionsbäume

Benutzen Sie die Datensätze, die Sie hier finden, für diese Aufgabe (außer dem Datensatz `regression`). Für nominale Attribute beachten Sie bitte, dass der Lerner `M5P` eine Binarisierung der Daten vornimmt ($A = a <= 0.5$ bedeutet also: alle Instanzen wo A NICHT den Wert a hat). Die Gesamtanzahl der Instanzen ist n, der tatsächliche Wert einer Instanz j ist y_j und der vorhergesagte Wert einer Instanz j ist r_j (genau wie im Skript).

- Vergleichen Sie den *Mean Absolute Error* ($\frac{1}{n} \cdot \sum_j |y_j - r_j|$) und den *Root Mean Squared Error* ($\sqrt{Mean\ Squared\ Error}$) (10 CV oder Test Set wenn verfügbar), sowie die Modelle (Interpretierbarkeit/Größe) jeweils für den Regressionsbaumlerner `M5P`, einmal mit angeschaltetem Pruning und einmal ohne Pruning (Benutzen Sie Regressionsbäume, also setzen Sie die Option 'buildRegressionTree' auf 'True'). Bringt Pruning bei Regressionstasks eine Verbesserung?
- Verwenden Sie nun Model Trees (Option 'buildRegressionTree' auf 'False' setzen, ansonsten Default Optionen). Vergleichen Sie die Model Trees mit den Regressionsbäumen.
- Verwenden Sie nun den Datensatz `regression`. Dieser entspricht dem Datensatz aus der Übung. Vergleichen Sie den Baum aus der Übung mit einem Regressionsbaum, den Sie mit `M5P` gelernt haben. Verwenden Sie hier einen Regressionsbaum ohne Pruning, der min. 1 Instanz pro Blatt besitzen muss. Betrachten Sie wieder die Größe und z.B. den *Mean Absolute Error* jeweils auf dem Testset (`regression_test`).