Maschinelles Lernen: Symbolische Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2008/2009
5. Projektaufgabe für den 20.1.2009

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`).