Maschinelles Lernen: Symbolische Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2009/2010
3. Projektaufgabe für den 9.2.2010

Hinweise zu dieser Projektaufgabe:

- Hierbei handelt es sich um die letzte Projektaufgabe.
- Beachten Sie bitte, dass Sie, falls Sie die Aufgabe zu den Regressionsbäumen bereits für das vorherige Projekt bearbeitet haben, diese einfach in diese Abgabe kopieren sollen.
- Manche der Aufgaben kamen in der Vorlesung sowie in der Übung noch nicht dran. Selbstverständlich können Sie diese Aufgben auch bereits jetzt schon bearbeiten. Ansonsten können Sie auch mit der Bearbeitung so lange warten, bis die Aufgaben in der Vorlesung und in der Übung behandelt worden sind.
- Die Deadline für die Abgabe dieser Projektaufgabe ist Sonntag, der **7.2.2010, 23:59 Uhr**.

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 10. 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`).

Ensemble-Lernen

In dieser Aufgabe sollen unterschiedliche Ensemble-Methoden eingesetzt und deren Ergebnisse verglichen werden. Der Entscheidungsbaumlerner J48 soll als Basislerner verwendet werden. Benutzen Sie bitte die Datensätze, die Sie hier herunterladen können.

a) Bestimmen Sie die Genauigkeit des regulären J48.

b) Verwenden Sie nun Bagging mit J48 und AdaBoost mit J48. Benutzen Sie außerdem noch Random Forests. Bestimmen Sie für die so erhaltenen Klassifizierer die Genauigkeiten für eine stetig wachsende Anzahl von Iterationen (bei den Random Forest verändern Sie bitte die Anzahl der Bäume). Wie interpretieren Sie die Entwicklung der erzielten Genauigkeiten?

Entdecken von Assoziationsregeln

Das Datenset adult.arff enthält Daten von 48842 US Bürgern über Geschlecht, Ausbildung, Familienstand, Beruf, Einkommen (class Variable), etc. Versuchen Sie, mit dem Apriori-Algorithmus aus Weka in diesem Datenset *interessante* Regeln

zu finden. Sie können dabei sowohl die Optionen von Weka ausprobieren (z.B. `-T` das Maß, nach dem die Regeln sortiert werden) als auch das Datenset verändern (z.B. durch Entfernen einzelner Attribute). Beachten Sie, daß in der Version zum Download zwei numerische Attribute enthalten sind, die Sie diskretisieren oder einfach entfernen können. Falls die Laufzeiten zu lange werden (mehrere Minuten), können Sie auch auf einer Teilmenge der Daten arbeiten.

Pre-Processing

Wählen Sie drei Datenset aus den fünf hier verfügbaren Datensets aus. Erstellen Sie für jedes Datenset eine diskretisierte Version (`weka.filters.supervised.attribute.Discretize`).

a) Schätzen Sie die Genauigkeit von `J48` mittels Cross-validation auf den ursprünglichen Daten und auf den diskretisierten Daten ab.

b) Der Meta-Classifier `FilteredClassifier` erlaubt, eine Kombination einer Pre-processing Methode und eines Classifiers zu einem neuen Classifier zu machen. Erzeugen Sie die Kombination `Discretize` und `J48` und schätzen Sie deren Genauigkeit auf den ursprünglichen Daten ab.

Wie interpretieren Sie den Vergleich der Genauigkeiten und der Größe der gelernten Bäume dieser drei Experimente (die Ergebnisse können über die drei Datensets gemittelt werden)?