Prof. Dr. Morik
Christian Bockermann

Dortmund, 21. Oktober 2008

Übung zur Vorlesung

Maschinelles Lernen

Wintersemester 2008/2009
Blatt 2

Wiederholung In diesem Teil der Aufgabeblätter sollen Sie sich kurz an die bisherigen Dinge erinnern und ggf. einige vertiefende Überlegungen anstellen:

1. Welche Gründe sprechen für die quadratische Fehlerfunktion RSS? Warum verwendet man nicht die einfache Summe der Fehler?

2. Welcher Code ist häufig effizienter und warum?

```
for( Example ex : examples ){
  for( Attribute a : ex.getAttributes() ){
     if( ! a.isNominal() )
       sum += ex.getValue( a );
  }
}
```

```
for( Attribute a : examples.getAttibutes ){
  if( ! a.isNominal() ){
    for( Example ex : examples )
       sum += ex.getValue( a );
  }
}
```

Aufgabe 1 6 Punkte

Sie finden im Verzeichnis `samples` des *subversion*-Repositorys zu den Übungen ein Experiment `LinReg.xml`. Dieses Experiment generiert Daten, welche zum Lernen einer lineare Regression genutzt werden sollen.

1. Probieren Sie verschiedene Visualisierungen der Daten mit Hilfe der Plotter-Ansicht aus.

2. Erweitern Sie das Experiment um einen Lerner *Lineare Regression* und lernen Sie damit ein lineares Modell. Welches konkrete Modell ergibt sich?

3. Verändern Sie in der Kette *Datengenerierung* den Parameter `local_random_seed` des Operators *ExampleSetGenerator* und wiederholen Sie das Experiment mit den Werten 19 und 23142.

4. Erweitern Sie das Experiment um eine Kreuzvalidierung und geben Sie die Evaluierungsergebnisse der gelernten Modelle für die unterschiedlichen generierten Daten (für die `local_random_seed`-Werte 1976, 19 und 23142) an.

Geben Sie als Lösung zu den Experimenten lediglich das jeweils gelernte, lineare Modell an. Als Lösung der Kreuzvalidierungsaufgabe geben Sie bitte zusätzlich die Experiment-Datei (xml) ab.

Hinweis: Der Operator *XValidation* verlangt als Eingabe ein ExampleSet und benötigt zwei innere Operatoren: einen Lerner, der ein Modell zurückliefert, und eine Operator-Kette, die die Evaluierung durchführt.
Der *XValidation*-Operator teilt seine Eingabe in Teilmengen auf, verwendet eine als Eingabe des ersten inneren Operators und die Ausgabe dieses ersten Operators (im Falle eines Lerner ist das das Modell) wird zusammen mit der übriggebliebenen Menge an den zweiten Operator (bzw. der Operator-Kette) als Eingabe weitergereicht.

Aufgabe 2

4 Punkte

In dieser Aufgabe sollen Sie sich nochmal etwas näher mit der linearen Regression befassen. Gegeben sind Datenpunkte im $\mathbb{R}^2$ (siehe Tabelle 1) und das allgemeine lineare Modell:

$$y = m \cdot x + b$$

1. Stellen Sie die zu minimierende Funktion (bzgl. des RSS-Kriteriums) auf, um die Regressionsgerade auf den unten angegeben Punkten zu finden.

2. Lösen Sie das Minimierungsproblem und geben Sie als Lösung eine allgemeine Gleichung zur Bestimmung der Regressionsgeraden im $\mathbb{R}^2$ an. Geben Sie dazu auch Ihren Rechenweg in der Abgabe an!

3. Wie groß ist der Trainingsfehler, den Ihr lineares Modell macht?

4. Geben Sie eine Laufzeit-Abschätzung für die Berechnung des linearen Modells an!

x	y
1	2
2	4
3	2
4	3
5	5
6	7
7	4
8	9
9	7

Tabelle 1