

Maschinelles Lernen: Symbolische Ansätze



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Wintersemester 2008/2009

Musterlösung für das 6. Übungsblatt

Aufgabe 1: Evaluierungsmethoden

Ein Datenset enthält $2 \times n$ Beispiele, wobei genau n Beispiele positiv sind und n Beispiele negativ sind. Der einfache Algorithmus ZeroRule betrachtet nur die Klassenverteilung der Trainings-Daten und sagt für alle Beispiele die Klasse + voraus, wenn mehr positive als negative Beispiele in den Trainings-Daten enthalten sind, und die Klasse – falls es umgekehrt ist. Bei Gleichverteilung entscheidet er sich zufällig für eine der beiden Klassen, die er dann immer vorhersagt.

- Wie groß ist die Genauigkeit dieses Klassifizierers, wenn die Verteilung der Trainings-Daten der Gesamt-Verteilung entspricht (d.h., wenn die Trainings-Daten repräsentativ sind)?

Lösung: Geht man davon aus, dass die Verteilung der Trainings-Daten repräsentativ ist, so erreicht der Klassifizierer eine Genauigkeit von 50%, da er zufällig klassifiziert.

- Schätzen Sie die Genauigkeit von ZeroRule mittels Leave-One-Out Cross-Validation ab.

Lösung: Bei Leave-One-Out CV wird ein Beispiel aus der Trainingsmenge entfernt, auf dem Rest gelernt und dann das eine Beispiel klassifiziert. Als Genauigkeit nimmt man den Mittelwert aller Beispiele. Nimmt man zB ein negatives Beispiel aus der Trainingsmenge heraus und lernt auf dem Rest, dann sagt der Klassifizierer die Klasse positiv vorher, klassifiziert also falsch (analog bei der Entnahme eines positives Beispiels). Der Klassifizierer ZeroRule wird also 0% Genauigkeit erreichen.

Aufgabe 2: Vorzeichentest

Sie vergleichen zwei Algorithmen A und B auf 20 Datensets und beobachten folgende Genauigkeitswerte:

Datenset	1	2	3	4	5	6	7	8	9	10
Algorithm A	0,91	0,86	0,93	0,74	0,65	0,91	0,87	0,95	0,78	0,86
Algorithm B	0,94	0,80	0,96	0,88	0,84	0,94	0,97	0,67	0,86	0,89
Datenset	11	12	13	14	15	16	17	18	19	20
Algorithm A	0,98	0,96	0,74	0,53	0,95	0,67	0,98	0,96	0,97	0,91
Algorithm B	0,87	0,90	0,79	0,51	0,96	0,69	0,79	0,98	0,98	0,76

Läßt sich mit Hilfe des Vorzeichentests nachweisen, ob einer der beiden Algorithmen A oder B signifikant besser ist als der andere? Folgt daraus, daß er nicht besser ist?

Lösung: Als erstes zählt man die Siege und Niederlagen eines Algorithmus, z.B. von A. A gewinnt 7 mal und verliert 13 mal. Wie in der Vorlesung beschrieben (Evaluierung und kosten-sensitives Lernen, Folie 15) geht man von einer Binomialverteilung aus. Man kann nun in der Tabelle der Kritischen Häufigkeiten (Evaluierung und kosten-sensitives Lernen, Folie 16) nachschauen, ab welchem Wert man den Bereich unter der Kurve, der als kritisch angesehen wird, verlässt. Da die Nullhypothese von einer Gleichheit der beiden Algorithmen ausgeht, kann man sicherer sein, eine korrekte Aussage getroffen zu haben, je kleiner die Fläche unter der Kurve ist. So muss ein Algorithmus z.B. auf 30 Datenmengen mindestens auf 21 Mengen besser sein, um bei einer Irrtumswahrscheinlichkeit von 5% signifikant besser als der andere zu sein und bei einer Wahrscheinlichkeit von 1% auf 23 Mengen.

Schaut man in der Tabelle nach, so sieht man, dass B nicht signifikant besser als A ist, da er bei einer Irrtumswahrscheinlichkeit von 5% mindestens auf 15 Mengen gewinnen hätte müssen. Da man keine Aussage über die Güte des Algorithmus treffen kann, folgt daraus auch nicht, dass er nicht besser ist.