Maschinelles Lernen und Data Mining

Übungsblatt für den 19.1.2006

Aufgabe 1

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

Aufgabe 2

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?

Aufgabe 3

Ein Naive Bayes Klassifikator schätzt für eine Menge von 12 Test-Beispielen folgende Wahrscheinlichkeiten $P(+|x)$, daß das Beispiel der positiven Klasse angehört:

class	+	−	+	−	+	−	+	−	+	−	+	−
$P(+\|x)$	0.85	0.66	0.70	0.10	0.87	0.80	0.90	0.43	0.75	0.33	0.41	0.54

1. Stellen Sie die Konfusionsmatrix auf und bestimmen Sie Error, Accuracy, True Positive Rate und False Positive Rate für diese Vorhersagen.

2. Zeichnen Sie die ROC-Kurve, die diesen Vorhersagen entpricht. und berechnen Sie die Area under the ROC-Curve.

3. Welcher Punkt im ROC-Space entspricht dem Standard Naive Bayes Klassifikator? Unter welchem Kosten-Modell ist dieser Punkt optimal?

4. Bestimmen Sie anhand der Kurve, welcher Punkt die größte Genauigkeit (Accuracy) liefert. Geben Sie einen Klassifikations-Threshold an, für den dieses Optimum erreicht wird. Wie groß ist die erreichte Genauigkeit?

5. Ab welchem Kosten-Verhältnis von False Positives und False Negatives wird der Threshold 0.82 optimal? Ab welchem Verhältnis der Threshold 0.35?