

Maschinelles Lernen: Symbolische Ansätze



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Wintersemester 2008/2009

7. Übungsblatt für den 16.12.2008

Aufgabe 1: Evaluierung und Kosten

Gegeben sei ein Datensatz mit 300 Beispielen, davon $\frac{2}{3}$ positiv und $\frac{1}{3}$ negativ.

- Ist die Steigung der Isometrien für Accuracy im Coverage Space für dieses Problem < 1 , $= 1$ und > 1 ?
- Ist die Steigung der Isometrien für Accuracy im ROC Space für dieses Problem < 1 , $= 1$ und > 1 ?
- Sie verwenden einen Entscheidungsbaum, um die Wahrscheinlichkeit für die positive Klasse zu schätzen. Sie evaluieren drei verschiedene Thresholds t (alle Beispiele mit einer geschätzten Wahrscheinlichkeit $> t$ werden als positiv, alle anderen als negativ klassifiziert) und messen folgende absolute Anzahlen von False Positives und False Negatives:

t	fn	fp
0.7	40	20
0.5	30	60
0.3	10	80

Geben Sie für jeden Threshold an, für welchen Bereich des Kostenverhältnisses $\frac{c(+|-)}{c(-|+)}$ der Threshold optimal ist.

- Wie hoch ist die maximale Genauigkeit (Accuracy), die Sie im Szenario von Punkt c bei einer False Positive Rate von maximal 30% erreichen können? Wie gehen Sie dabei vor?
- Sie erfahren, daß in Ihrer Anwendung ein False Positive 2 Cents kostet und ein False Negative 5 Cents kostet. Mit welchem Threshold können Sie die Kosten minimieren? Wie hoch sind die entstanden minimalen Kosten für diese 300 Beispiele?
- Sie bekommen die Möglichkeit, zusätzlich zu den vorhandenen 300 Beispielen noch 400 selbst auszuwählen. Wie würden Sie die Auswahl treffen, damit ein Lerner, der Kosten nicht berücksichtigen kann, unter den in Punkt e angegebenen Kosten möglichst effektiv wird?

Aufgabe 2: Nearest Neighbour

Hinweis: Diese Aufgabe bezieht sich auf die Vorlesung, die am Donnerstag gehalten wird.

Gegeben sei folgende Beispielmenge:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High		No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain		Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8		Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D11	Sunny	20	Normal	Strong	
D12	Overcast	21	High	Strong	Yes
D13		26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

- a) Überlegen Sie sich eine gute Abstandsfunktion für die einzelnen Attribute.
- b) Benutzen Sie 3-NN zum Ausfüllen der fehlenden Werte.
Beziehen Sie hier die Klassifikation mit ein oder nicht? Warum?
- c) Welchen Klassifikationswert gibt k -NN für die folgende Instanz aus?
1. Outlook=Sunny, Temperature=23, Humidity=High, Wind=Strong
- Testen Sie verschiedene k . Für welches k ändert sich die Klassifikation gegenüber $k = 1$?
- d) Berechnen Sie den Klassifikationswert obiger Instanz mittels abstandsgewichtetem NN (Inverse Distance Weighting).