

# Maschinelles Lernen: Symbolische Ansätze



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Wintersemester 2009/2010

5. Übungsblatt für den 26.11.2009

## Aufgabe 1: Covering-Algorithmus und Coverage-Space

Visualisieren Sie den Ablauf des Covering-Algorithmus mit den Daten des letzten Übungsblatts Aufgabe 1b). Veranschaulichen Sie das Lernen jeder einzelnen Regel im Coverage-Space. Zeichnen Sie auch alle untersuchten Kandidaten-Regeln ein und skizzieren Sie zusätzlich die Linien, die dem jeweiligen Bewertungsmaß entsprechen. Sie sollten sowohl einen Graphen für jede Regel als auch für das Lernen der gesamten Theorie anfertigen

- für das Bewertungsmaß Accuracy, wobei die Regel mit der höchsten Bewertung ausgewählt wird und
- für das Bewertungsmaß Precision (zumindest für die zweite gelernte Regel, da die erste Regel nur einmal verfeinert wird).

## Aufgabe 2: Heuristiken und Äquivalenzen

In der Vorlesung haben Sie die Heuristiken Precision, Accuracy, Weighted Relative Accuracy, Gini-Index und ihre äquivalenten Berechnungen kennengelernt.

a) Zeigen Sie die Äquivalenz von

- Accuracy  $h_{Acc} = \frac{p+(N-n)}{p+N}$  und  $p - n$
- WRA  $h_{WRA} = \frac{p+n}{p+N} \left( \frac{p}{p+n} - \frac{p}{p+N} \right)$  und  $\frac{p}{p} - \frac{n}{N}$
- Gini-Index  $h_{Gini} = 1 - \left( \frac{p}{p+n} \right)^2 - \left( \frac{n}{p+n} \right)^2$  und  $\frac{pn}{(p+n)^2}$

b) In der vorangegangenen Teilaufgabe haben Sie die folgende Äquivalenz

$$h_{Gini} = 1 - \left( \frac{p}{p+n} \right)^2 - \left( \frac{n}{p+n} \right)^2 \equiv \frac{pn}{(p+n)^2}$$

bewiesen. Zeigen Sie anhand dieses Ergebnisses, daß der Gini-Index äquivalent zu Precision ist, falls  $p < n$  gilt, bzw. äquivalent zur negierten Precision, falls  $p \geq n$  gilt.

## Aufgabe 3: CN2's likelihood ratio statistics

Signifikanz-Niveau	0,9	0,95	0,975	0,99	0,995
Schwellen-Wert	2,71	3,84	5,02	6,64	7,88

Gegeben sei ein Datensatz, der aus 60 positiven und 40 negativen Beispielen besteht.

- a) Berechnen Sie für die folgenden Regeln, von denen Ihnen nur die Abdeckung bekannt ist, die "CN2's likelihood ratio statistics" und bestimmen Sie für die oben gegebenen Signifikanz-Niveaus, ob die Regeln gepruned werden würde.
- R1:  $p=11$  und  $n=3$
  - R2:  $p=15$  und  $n=2$
  - R3:  $p=22$  und  $n=6$
- b) Überlegen Sie sich ohne Berechnung der "CN2's likelihood ratio statistics", warum eine Regel, die 9 positive und 6 negative Beispiele abdeckt, für alle Signifikanz-Niveaus gepruned wird.