

Data Mining und maschinelles Lernen



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Wintersemester 2015/2016

2. Übungsblatt

Aufgabe 1 Konzepte, Regeln und Hypothesenräume

Gegeben sei ein Beispielraum, der durch n binäre Attribute aufgespannt wird. Hypothesenraum sind alle möglichen Regeln, die sich durch Konjunktionen von Tests der Form $Attribut = Wert$ ergeben.

- Wie viele mögliche Konzepte gibt es?
- Wie viele mögliche Regeln gibt es?
- Unter der Voraussetzung, daß das zu lernende Konzept im Hypothesenraum darstellbar ist, wie viele Fehler kann der Algorithmus Find-S beim Lernen dieses Konzepts maximal machen?

[Anm: Ein Fehler ist hier ein (Trainings-)Beispiel, das von der momentanen Theorie falsch klassifiziert wird.]

- Gegeben sei ein Hypothesenraum, der nur *Disjunktionen* von binären Attribut-Wert-Paaren erlaubt. Eine gültige Regel wäre also z.B.

$\text{if } (att_i = t) \text{ or } (att_j = f) \text{ or } (att_k = f) \text{ then } +$

Überlegen Sie sich Verallgemeinerungs- und Spezialisierungsvorschriften für diesen Hypothesenraum und geben Sie einen geeigneten Lernalgorithmus an.

Aufgabe 2 Find-S und Find-GSet

Sie wollen den Find-S Algorithmus auf numerische Daten anwenden, indem Sie Intervalle definieren.

- Wie sieht das spezifischste und generellste Element der Sprache aus, wenn Sie offene Intervalle verwenden?
- Wie sieht das spezifischste und generellste Element der Sprache aus, wenn Sie geschlossene Intervalle verwenden?
- Finden Sie eine passende Generalisierungsvorschrift und simulieren den Algorithmus Find-S auf folgenden Beispielen:

A1	Klasse
0.5	–
1.0	+
2.1	–
0.8	–
1.5	+
1.8	+

- Finden Sie eine passende Spezialisierungsvorschrift und simulieren Sie den Algorithmus Find-GSet auf denselben Beispielen.
- Skizzieren Sie beide Lösungen und vergleichen Sie die Allgemeinheit.