# Maschinelles Lernen und Data Mining

Übungsblatt für den 30.01.2007

## Aufgabe 1

Gegeben sei ein Datenset mit 2 numerischen Attributen $x$, $y$ mit jeweils Wertebereich [0,1]. Das gesuchte Konzept sei $y > 1 - x$ (wie auf Folie 16 dargestellt). Nehmen Sie weiters an, die Trainingsbeispiele sind gleichmäßig über den Bereich verteilt (d.h. alle $x$ und $y$-Koordinaten sind gleich wahrscheinlich). Ihr Lernalgorithmus findet Decision Stumps mit nur einem binären Knoten, wobei er versucht, die Anzahl der falsch klassifizierten Beispiele zu minimieren.

- Bestimmen Sie den Fehler eines Splits, abhängig von seiner Position auf der $x$-Achse (bzw. $y$-Achse). Den Fehler können Sie durch die Größe der Fläche der Region von falsch klassifizierten Beispielen bestimmen. Wo werden die Splits in etwa gesetzt werden? Was ist der erwartete Fehler dieses Algorithmus?
- Nehmen Sie an, Sie verwenden Bagging, um die Performanz zu verbessern. Überlegen Sie sich, wie die Klassifizierer der einzelnen Iterationen aussehen werden.
- Überlegen Sie sich grafisch, wie der Gesamt-Klassifizierer nach einigen wenigen Iterationen aussehen muß.
- Wenn Sie statt der Decision Stumps einen Lernalgorithmus haben, der völlig zufällig einen Split auswählt, und dann die beiden Blätter so markiert, daß der Fehler minimiert wird. Was ist der erwartete Fehler dieses Algorithmus? Wenn Sie mit diesem Algorithmus mehrere Theorien lernen und diese kombinieren, erwarten Sie bessere oder schlechtere Ergebnisse als mit Bagging?

# Aufgabe 2

Rechnen Sie das AdaBoost-Beispiel aus der Vorlesung nach. Verwenden Sie für die einzelnen Datenpunkte die folgenden Koordinaten (x, y, Klasse):

```
1,  5,  +          3,  1,  -
2,  2,  +          4,  6,  -
5,  8,  +          7,  4,  -
6, 10,  +          9,  3,  -
8,  7,  +         10,  9,  -
```

Der Basis-Lerner wählt unter allen möglichen Splits jenen aus, bei dem die Gesamtsumme der Gewichte der falsch klassifizierten Beispiele minimiert wird.