Information Mining - Wintersemester 2014
Dipl.-Inform. Vu Tran, LF 139
vtran@is.inf.uni-due.de

Übungsblatt 7

Aufgabe 14: Evaluierungsmaße

(a) Bei einer Klassifikation können die folgenden vier Fälle auftreten. Erläutere kurz die Bedeutung der Begriffe und gib jeweils ein Beispiel an.

- Richtig positiv (true positive, TP)
- Richtig negativ (true negative, TN)
- Falsch negativ (false negative, FN)
- Falsch positiv (false positive, FP)

(b) Wie sind die nachfolgenden Maße definiert? Was bestimmen sie?

- Accuracy
- Kosten/Nutzen
- Mittlerer quadratischer Fehler
- Precision und Recall

(c) Welche der in (b) genannten Maße sind für die Evaluierung der Erkennung von Spam-E-Mails (vgl. Aufgabe 8) geeignet bzw. anwendbar? Gib jeweils eine kurze Begründung an.

Aufgabe 15: Evaluierung in *RapidMiner*

(a) Führe mit *RapidMiner* eine Evaluierung des Entscheidungsbaums aus Aufgabe 7 mit Hilfe einer 10-fachen Kreuzvalidierung durch[1]. Verwende als Evaluierungsmaße *Accuracy* und *Error*.

Welche Werte werden für die Evaluierungsmaße ermittelt?

Lösungshilfe: Es werden u. a. die folgenden Operatoren benötigt: `X-Validation`, `Apply Model`, `Performance (Classification)`.

(b) Welches der beiden Kriterien `information gain` und `gain ratio` zur Aufteilung von Instanzen beim Lernen des Entscheidungsbaums führt zu besseren Ergebnissen?

Vergleiche die Werte mithilfe des t-Tests (Operator `T-Test`), um zu überprüfen, ob die Unterschiede statistisch signifikant sind.

Lösungshilfe: Der folgende Operator ist hilfreich: `Multiply`.

[1] `http://www.is.inf.uni-due.de/courses/im_ws14/uebung/data_a15.csv`