Maschinelles Lernen Symbolische Ansätze: Projekt Aufgaben 4-6

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 4 - Entscheidungsbäume
- Benutzte Datensätze
- ROC Kurven
- Accuracy und Baumgröße

Aufgabe 5 - Nearest Neighbour
- Benutzte Datensätze
- Resultat

Aufgabe 6 - Regressionsbäume
- Benutzte Datensätze
- Pruning
- Model Trees
- Datensatz “regression”

Aufgabe 4 - Entscheidungsbäume
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- Breast Cancer Data
- 1984 United States Congressional Voting Records Database

$\rightarrow$ Auf beide Datensätze unsupervised Filter “ReplaceMissingValues” auf unvollständige Attribute anwenden
$\rightarrow$ Danach Lernen mit ID3 und J48 und Vergleich der Ergebnisse

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

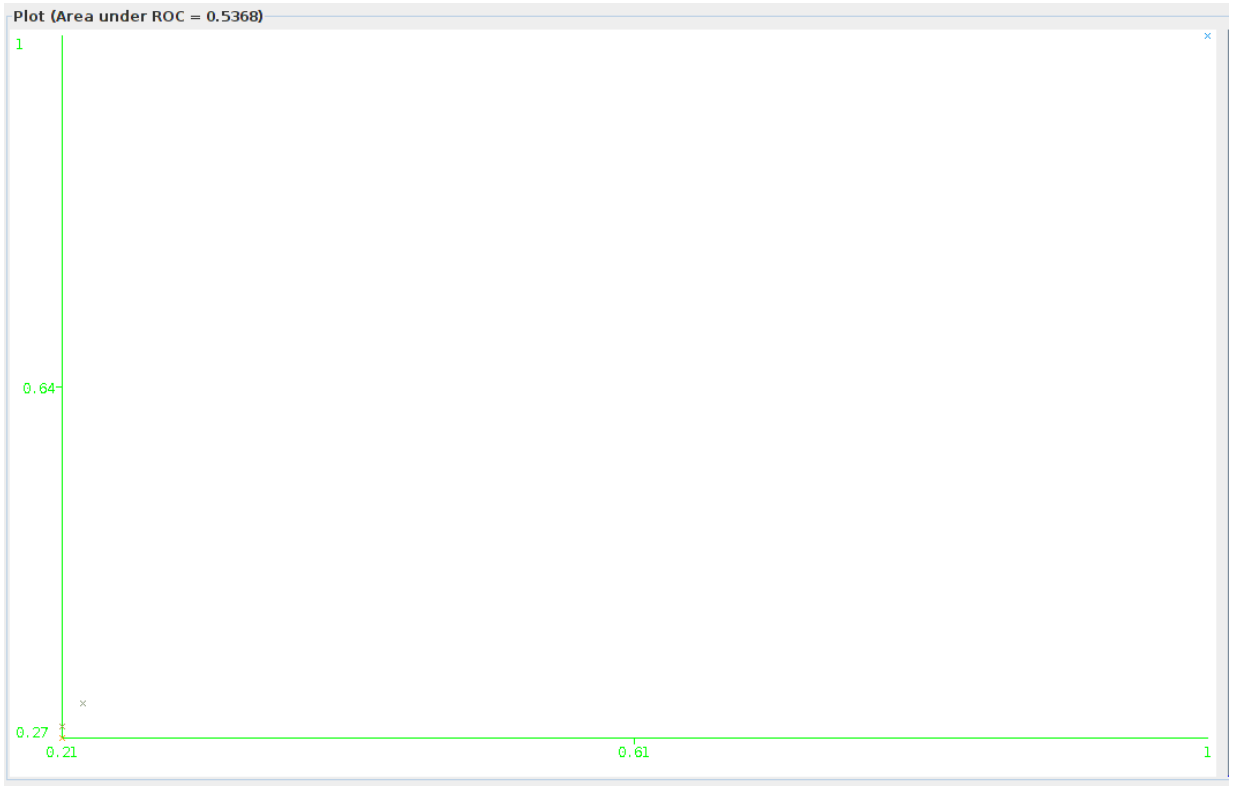

Abbildung : ID3 - ROC-Kurve für “breast-cancer”

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

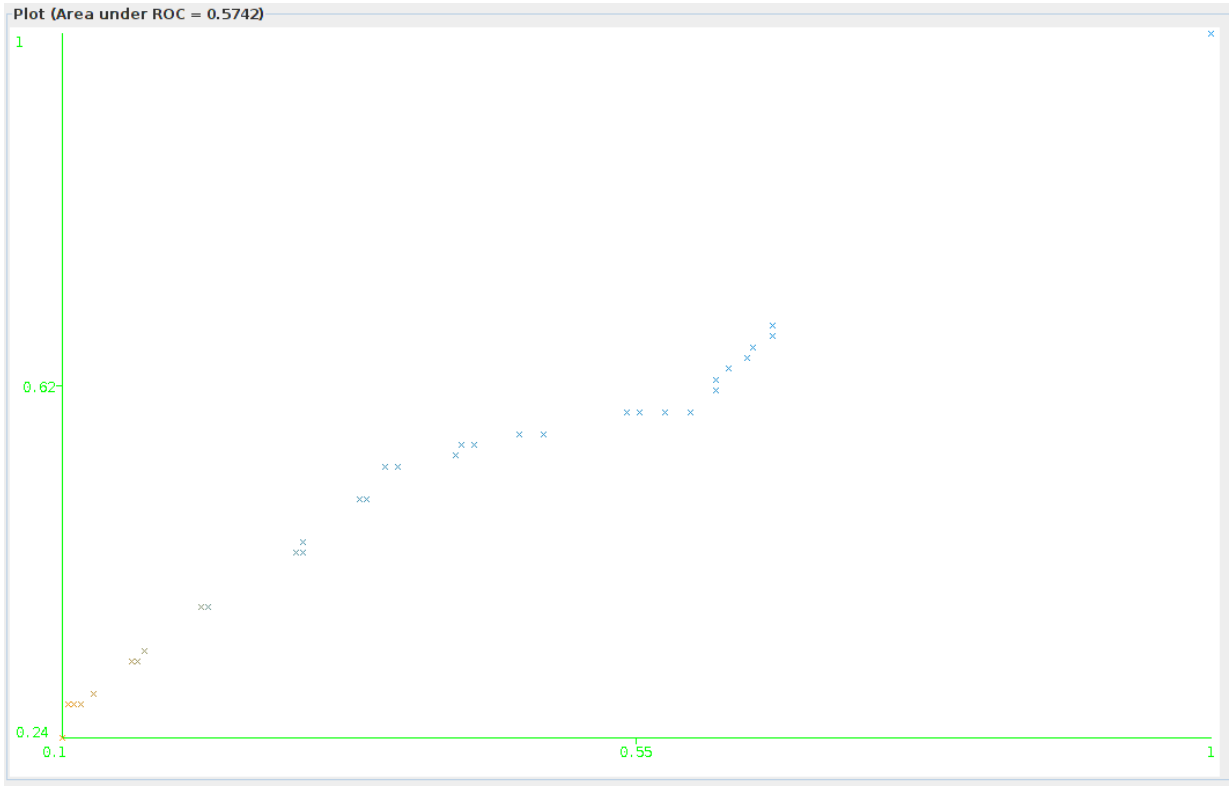

Abbildung : J48 (unpruned) - ROC-Kurve für “breast-cancer”

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

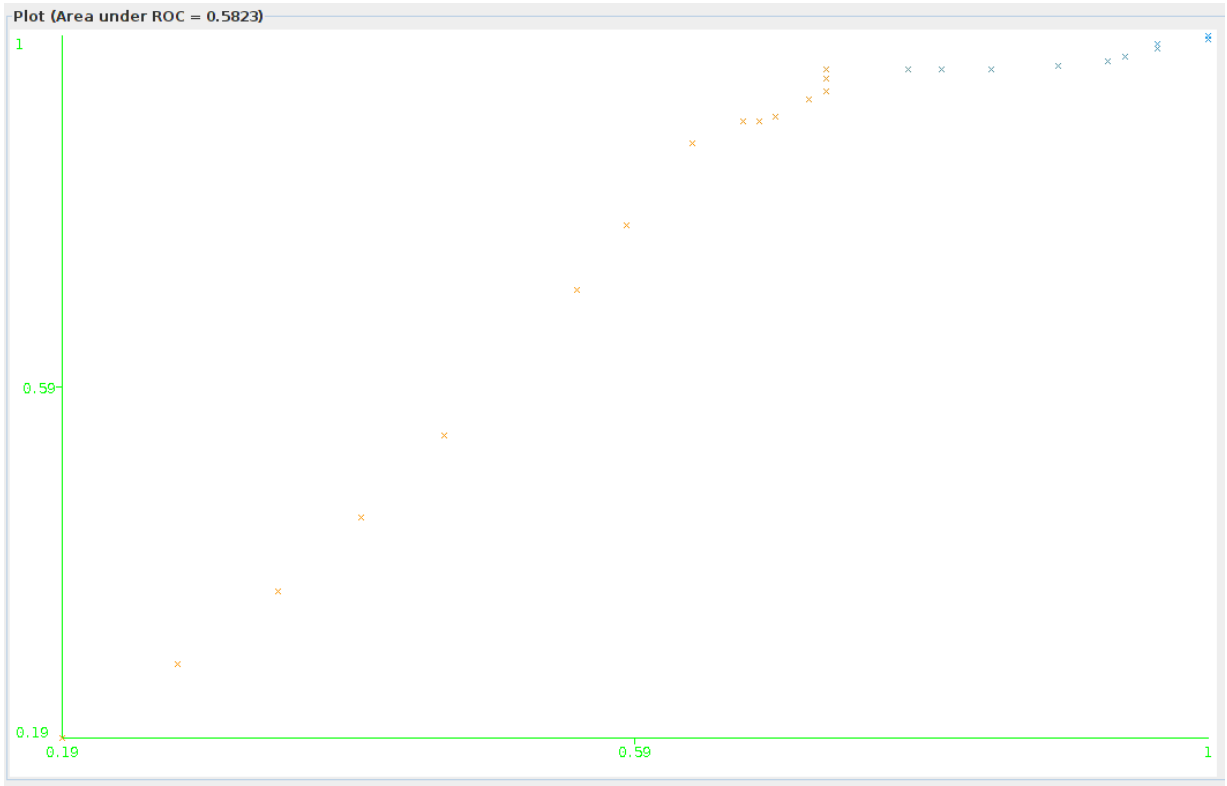

Abbildung : J48 (pruned) - ROC-Kurve für “breast-cancer”

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

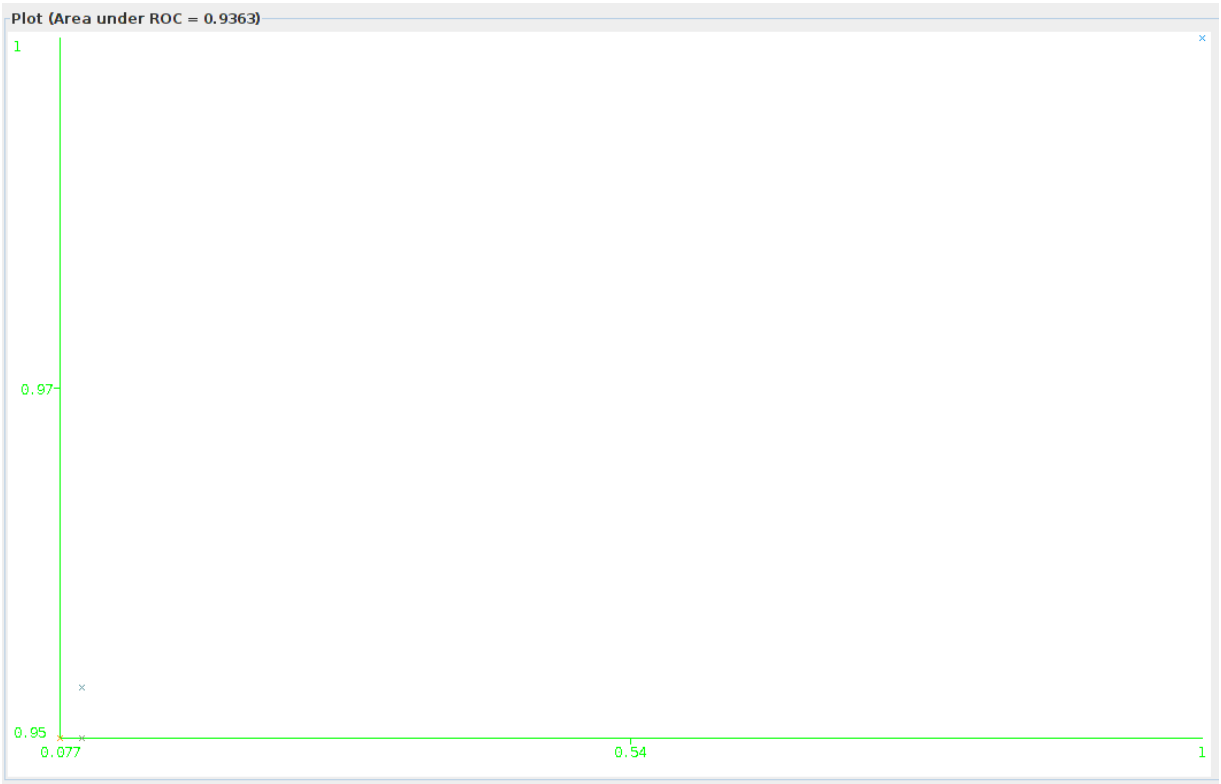

Abbildung : ID3 - ROC-Kurve für “vote”

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

Abbildung : J48 (unpruned) - ROC-Kurve für “vote”

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

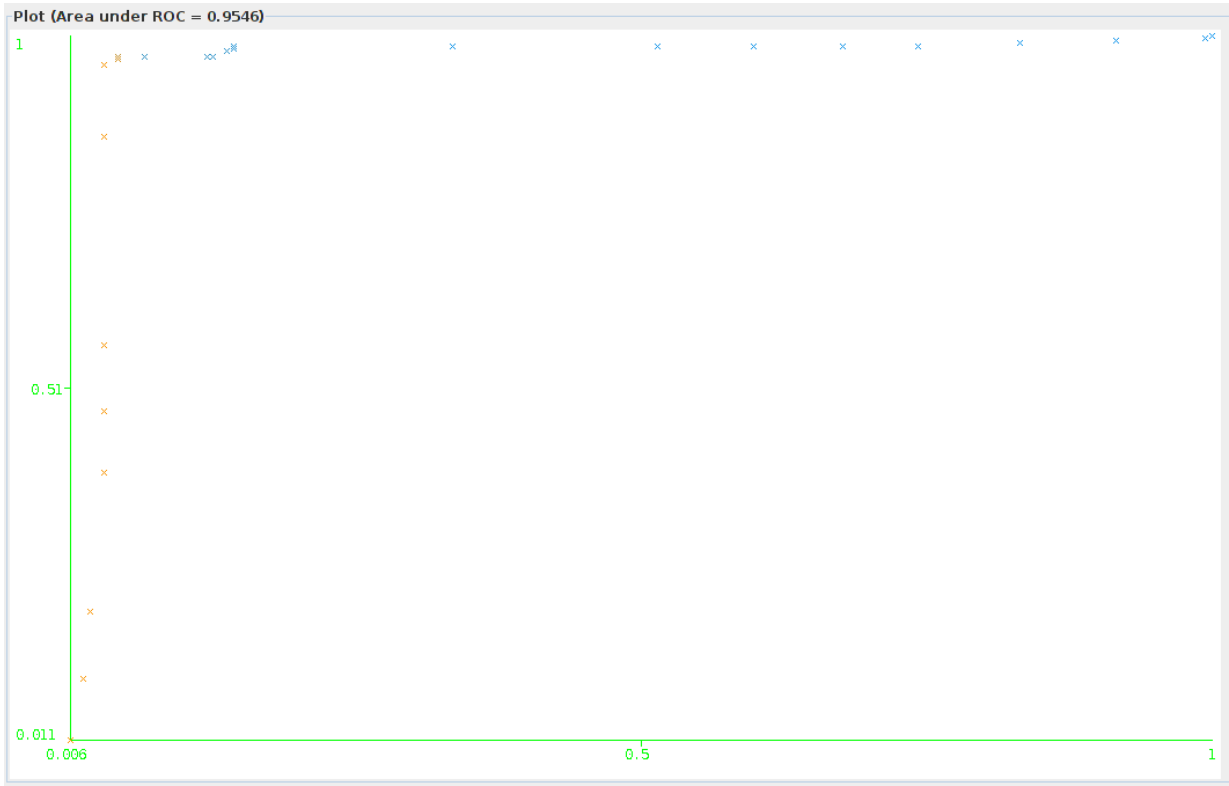

Abbildung : J48 (pruned) - ROC-Kurve für “vote”

Aufgabe 4 - Entscheidungsbäume
ROC Kurven

TECHNISCHE UNIVERSITÄT DARMSTADT

Area Under ROC:

Datensatz	ID3	J48 (unpruned)	J48 (pruned)
breast-cancer	0.5368	0.5742	0.5823
vote	0.9363	0.9647	0.9546

- Der Datensatz *breast-cancer* scheint mit den verwendeten Klassifizieralgorithmen nicht gut lernbar zu sein
- J48 hat generell eine deutlich bessere Performance als ID3
  - Aber hier ist zumindest im Bezug auf die Area under ROC kein eindeutiger Unterschied bei Pruning erkennbar

Aufgabe 4 - Entscheidungsbäume
Accuracy und Baumgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensatz	ID3	J48 (unpruned)	J48 (pruned)
breast-cancer	57% Accuracy ~465 Knoten	68.5% Accuracy ~180 Knoten	75.5% Accuracy 6 Knoten
vote	93% Accuracy ~60 Knoten	95% Accuracy 25 Knoten	96% Accuracy 6 Knoten

- J48 erreicht eine höhere Genauigkeit als ID3
- Pruning erhöht die erzielte Genauigkeit weiter
  - Resultierender Baum generalisiert besser und vermeidet Overfitting
- Die Bäume von ID3 sind viel größer als die von J48
  - Ziel von ID3: In jedem Blatt nur Beispiele einer einzigen Klasse
- (Post-)Pruning bei J48 reduziert die Baumgröße deutlich
  - Knoten entfernen, wenn dadurch der erwartete Fehler geringer wird
  - Verhindern von Fragmentierung (Minimalanzahl an Instanzen in Knoten)

Aufgabe 5 - Nearest Neighbour
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- Breast Cancer Data
- 1984 United States Congressional Voting Records Database

$\rightarrow$ Auf beide Datensätzen den unsupervised Filter “ReplaceMissingValues” anwenden $\rightarrow$ Danach anwenden und evaluieren von IBk

Aufgabe 5 - Nearest Neighbour
Resultat

TECHNISCHE UNIVERSITÄT DARMSTADT

Cross Validation Accuracy:

Datensatz	k=1	k=3	k=5	k=7	k=9	k=11	Aufg. 4
breast-cancer	71.7%	74.1%	73.8%	73.8%	73.8%	72.7%	75.5%
vote	93.6%	93.1%	94.0%	93.3%	92.9%	92.6%	96%

- Es gibt keinen allgemeinen besten Wert für k
  - Dieser muss experimentell festgestellt werden
- Bei beiden Datensätzen führen zu kleine und zu große Werte für k zu einer schlechteren Genauigkeit (Noise bzw. zu große Neighborhood)

Aufgabe 6 - Regressionsbäume
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- Auto Price Dataset
- Concrete Compressive Strength
- Boston Housing Data
- Stock Prices Dataset
- Wine Quality

$\rightarrow$ Benutzen von M5P mit verschiedenen Optionen

Aufgabe 6 - Regressionsbäume
Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensatz	Unpruned	Pruned
autoprice	MAE 2075, RMSE 3287 Number of Rules 62	MAE 2096, RMSE 3336 Number of Rules 8
concrete	MAE 6.5, RMSE 8.3 Number of Rules 405	MAE 6.9, RMSE 8.7 Number of Rules 60
housing	MAE 3.2, RMSE 4.7 Number of Rules 193	MAE 3.3, RMSE 4.8 Number of Rules 26
stock	MAE 1.2, RMSE 1.6 Number of Rules 253	MAE 1.2, RMSE 1.6 Number of Rules 88
winequality	MAE 0.5, RMSE 0.7 Number of Rules 1562	MAE 0.6, RMSE 0.7 Number of Rules 73

- Pruning verringert die Größe des Baumes deutlich, während der Fehler nur geringfügig größer wird
- Bei Regression Trees ist Pruning sinnvoll, um fast ohne Performanceverlust die Interpretierbarkeit zu erhöhen

Aufgabe 6 - Regressionsbäume
Model Trees

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensatz	Pruned	Model Tree
autoprice	MAE 2096, RMSE 3336 Number of Rules 8	MAE 1467, RMSE 2171 Number of Rules 10
concrete	MAE 6.9, RMSE 8.7 Number of Rules 60	MAE 4.7, RMSE 6.4 Number of Rules 10
housing	MAE 3.3, RMSE 4.8 Number of Rules 26	MAE 2.5, RMSE 3.8 Number of Rules 19
stock	MAE 1.2, RMSE 1.6 Number of Rules 88	MAE 0.7, RMSE 0.9 Number of Rules 47
winequality	MAE 0.6, RMSE 0.7 Number of Rules 73	MAE 0.5, RMSE 0.7 Number of Rules 24

- Model Trees scheinen noch besser zu sein
  - fast immer kleiner und zusätzlich weisen sie kleineren Fehler auf
- Ursache könnte umfassendere Betrachtung der einzelnen Attribute im linearen Modell sein, anstatt nur den Mittelwert der Instanzen zu verwenden

Abschlussüberblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aufgabe 4 - Entscheidungsbäume
  - Benutzte Datensätze
  - ROC Kurven
  - Accuracy und Baumgröße
- Aufgabe 5 - Nearest Neighbour
  - Benutzte Datensätze
  - Resultat
- Aufgabe 6 - Regressionsbäume
  - Benutzte Datensätze
  - Pruning
  - Model Trees
  - Datensatz “regression”

Gruppenmitglieder

TECHNISCHE UNIVERSITÄT DARMSTADT

Joachim Brehmer, 1766932

Jeannine Endreß, 1669152

Uli Fahrer, 1664571