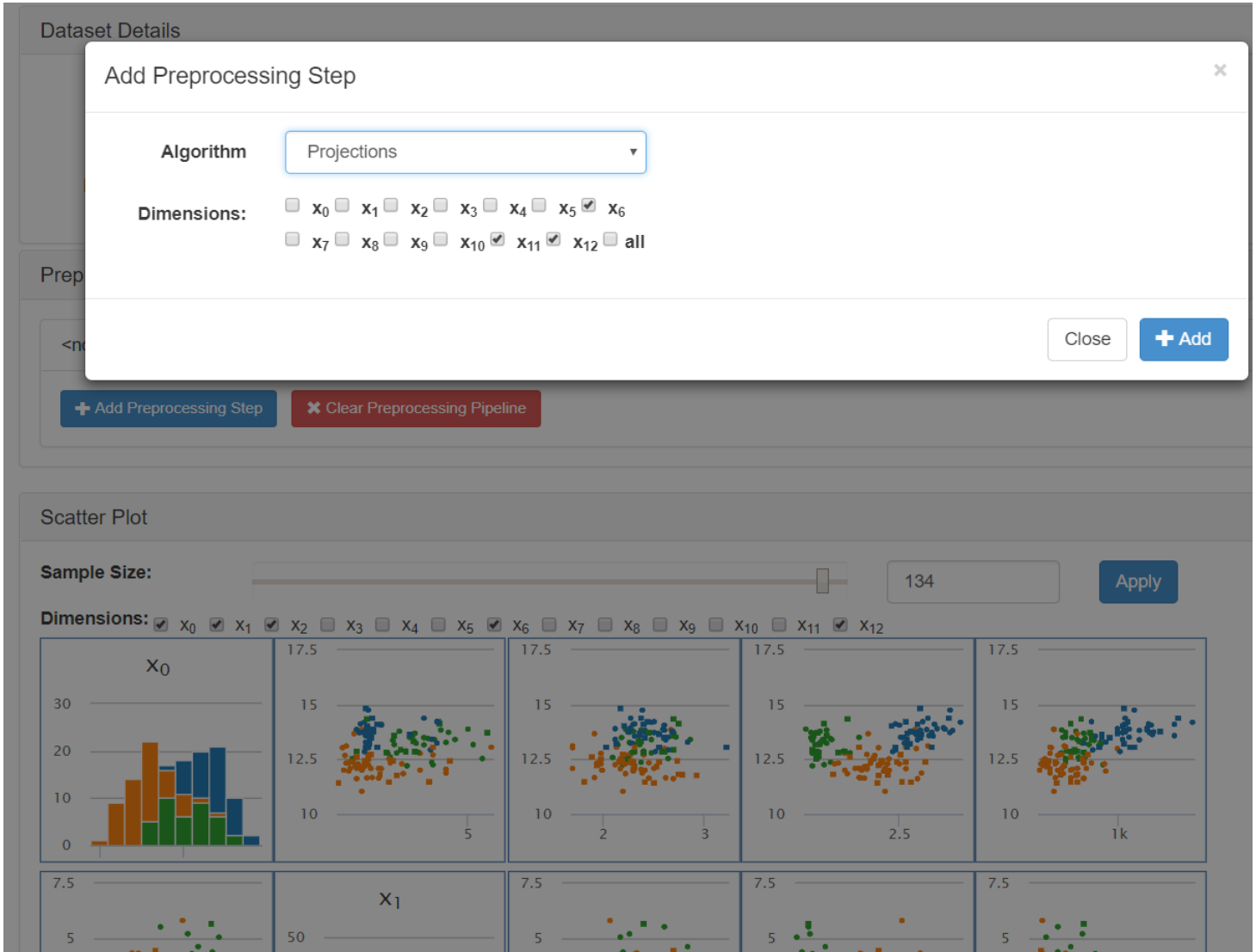
INF

FAKULTÄT FÜR
INFORMATIK

Feature Selection / Preprocessing

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

Was ist Feature Selection?



Warum Feature Selection?

- Mehr Variablen führen nicht automatisch zu besseren Ergebnissen.
  - Lernen von unwichtigen Daten
  - Mehr Daten notwendig
- Komplexität steigt.
- Datenmessung wird aufwändiger.
- Verständnispotenzial sinkt.

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

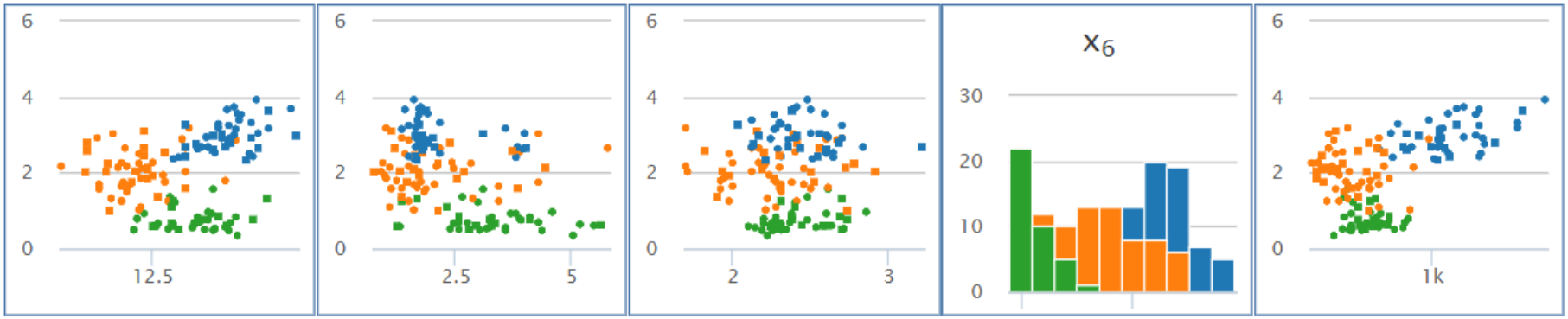
Wrapper-Methoden

- Testen Feature-Set anhand der Qualität einer Klassifikation.
- Klassifikator als Black Box
- Definiere:
  - Wie wird der Untermengen-Raum abgesucht?
  - Wie wird das Ergebnis einer Klassifikation evaluiert?
  - Welcher Klassifikator wird benutzt?
- Ergebnis an Klassifikator angepasst
- Nachteil: (Potenziell) sehr rechenaufwändig
- Beispiel: Beurteile Subset nach Test / Validierung

50530	Wine	kNN	distance: **"euclidean"** k: **5** weights: **"uniform"**	0.636	-	edit Clone Job Submit Job
50551	Wine	kNN	distance: **"euclidean"** k: **5** weights: **"distance"**	0.955	0.841	edit Clone Job

Filter-Methoden

- Testen Feature-Set anhand statistischer Eigenschaften in den Daten.
- Klassifikator nicht nötig.
- Definiere:
  - Wie wird der Untermengen-Raum abgesucht?
  - Wie wird die Nützlichkeit eines Subsets berechnet?
- Ergebnis für verschiedene Klassifikatoren nutzbar.
- Vorteil: (Potenziell) schnell
- Beispiel: Beurteile Subset nach visueller Begutachtung



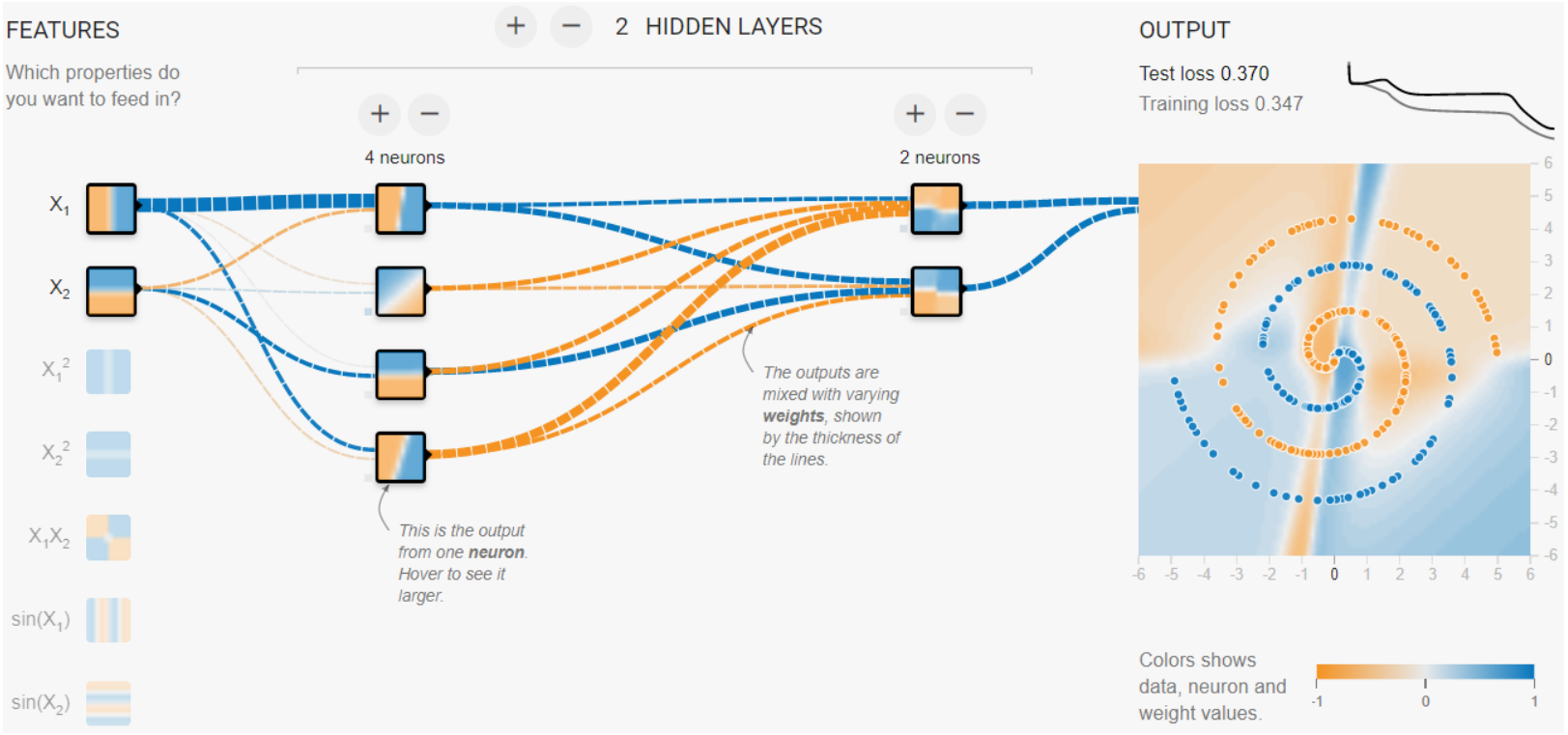
Filter-Methoden 2

- Transinformation (Mutual Information):
  - $$I(X,Y)=\sum_{y\in Y}\sum_{x\in X} p(x,y)\times\log\left(\frac{p(x,y)}{p(x)\times p(y)}\right)$$
  - Integral für kontinuierliche Werte
  - 0 für unabhängige Variablen
  - Steigt mit zunehmender Abhängigkeit
  - Vergleiche Transinformation zwischen jeder Variable und der Klassifikation.

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

Embedded-Methoden

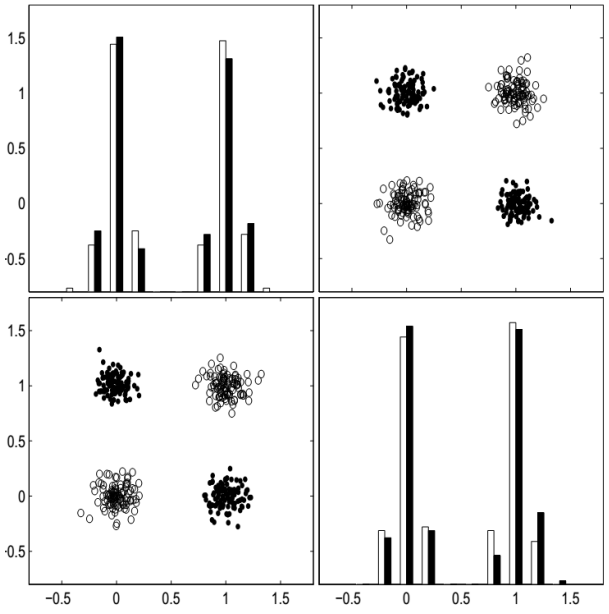
- Bauen Feature-Sets während des Lernens auf.
- In den Klassifikator integriert.
- Vorteil: (Potenziell) schnell
- Beispiel: Neuronale Netze



http://playground.tensorflow.org

Wie wählt man Feature Sets zum Testen aus?

- Klassisch:
  - Forward Selection: Starte bei 0, füge das Element hinzu, das den größten Gewinn an Genauigkeit bringt
  - Backward Elimination: Starte bei allen, entferne das Element, bei dem die Genauigkeit am geringsten abnimmt.
  - Setze einen Grenzwert für Zu-/Abnahme der Genauigkeit, um das Verfahren zu beenden.
- Problem:
  - Variablen von geringem Einzelnutzen können gemeinsam nützlich sein.
  - Redundante Variablen erkennen ist aufwändig.
  - Lösung: Ensemble Methods



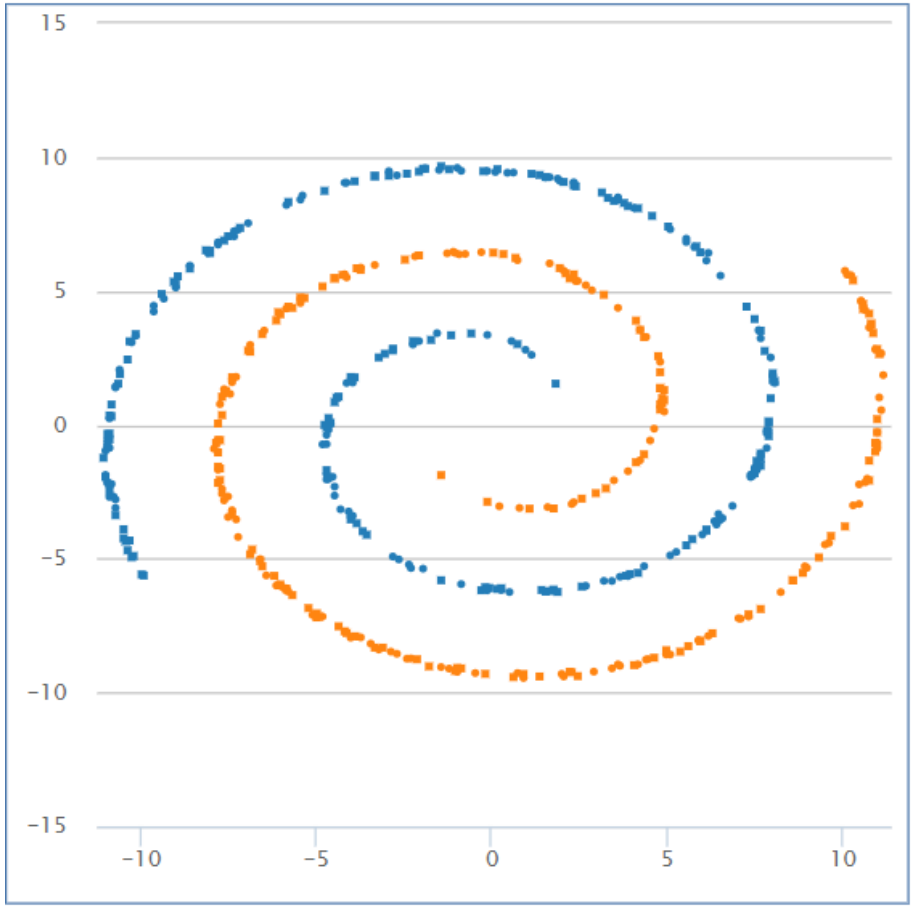
GUYON, Isabelle; ELISSEEFF, André. An introduction to variable and feature selection. Journal of machine learning research, 2003, 3. Jg., Nr. Mar, S. 1157-1182.

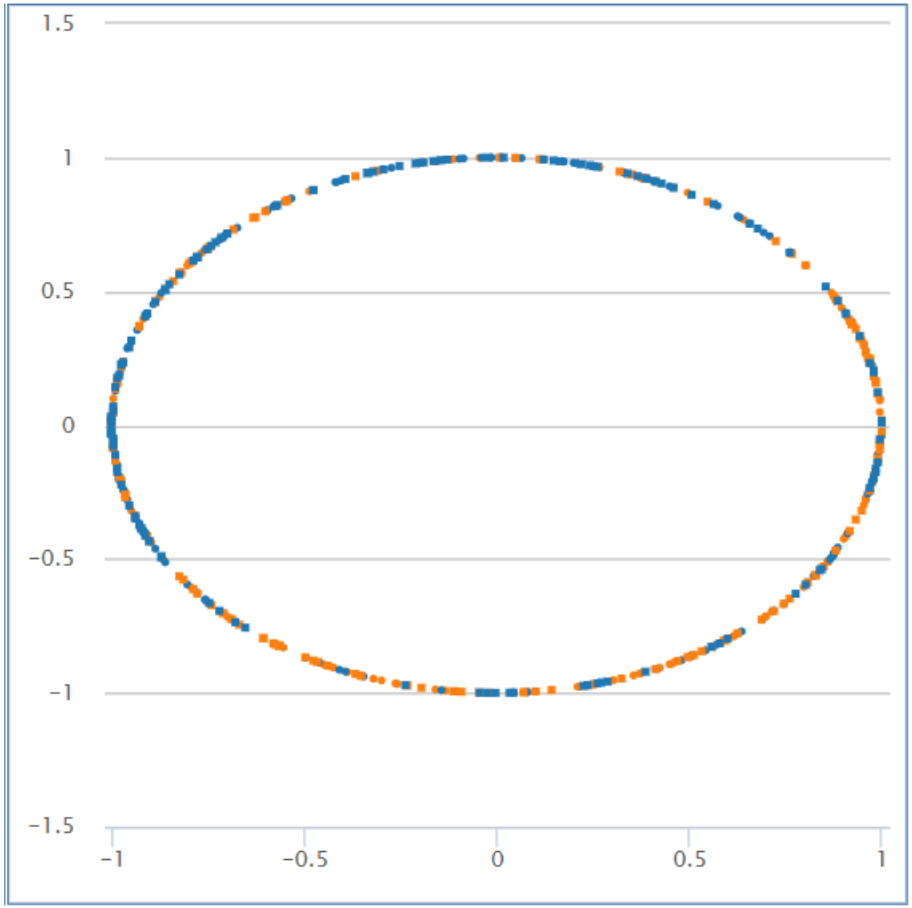
Weitere mögliche Preprocessing-Schritte

- Normalizer
- Standardizer
- Min-Max
- Principle Component Analysis (PCA)

Normalizer

- Bringt die Länge der Vektoren auf 1.
- Gut, wenn sie unwichtig ist, zum Beispiel bei Textklassifikation.

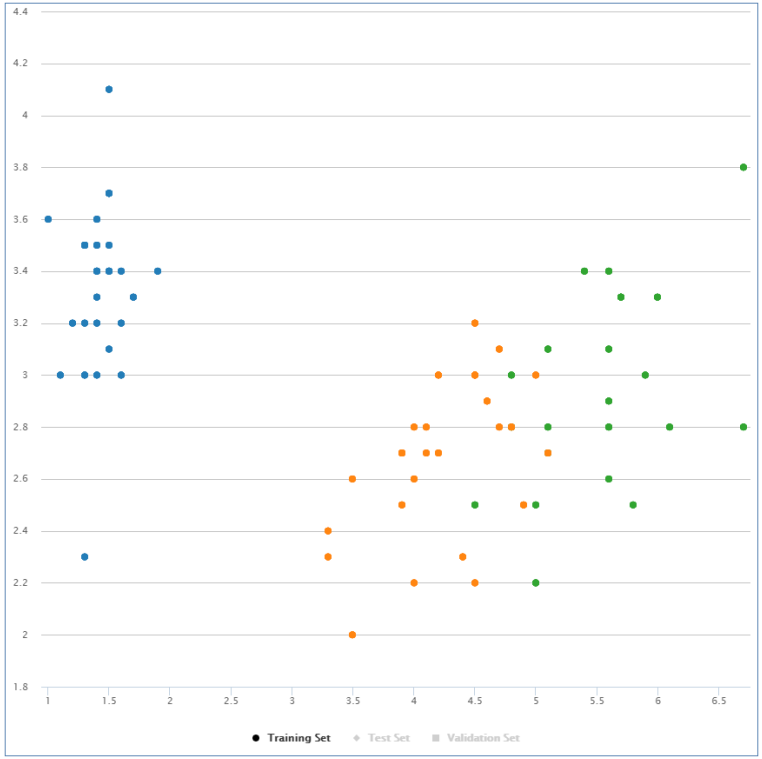


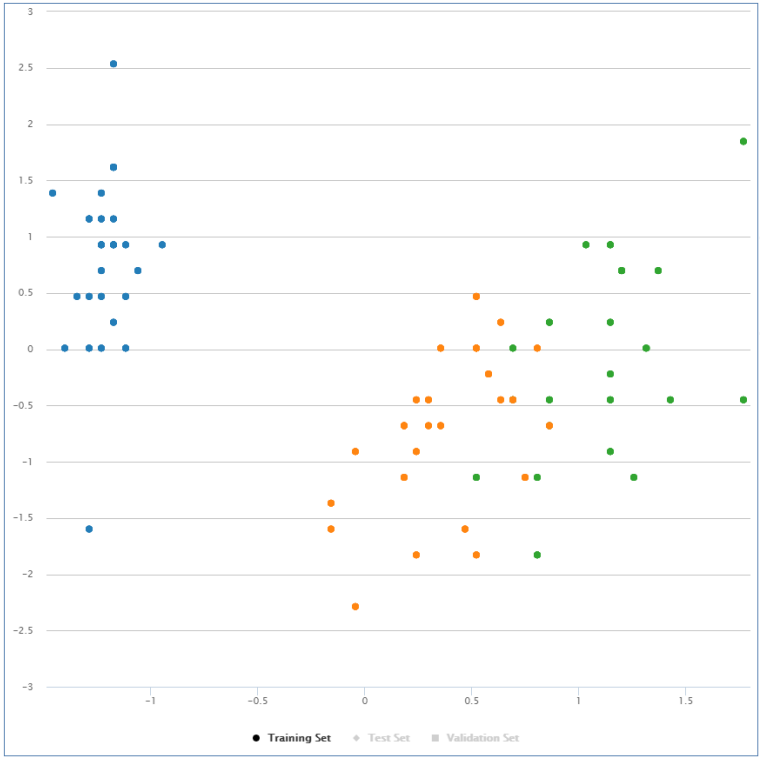


OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

Standardizer

- Bringt den Durchschnitt der Trainingsdaten auf 0, die Varianz auf 1.
- Macht die Dimensionen vergleichbar.
- Vorteil: Ausreißer beeinflussen die Skalierung kaum.
- Nachteil: Die Intervalle müssen nicht gleich sein.



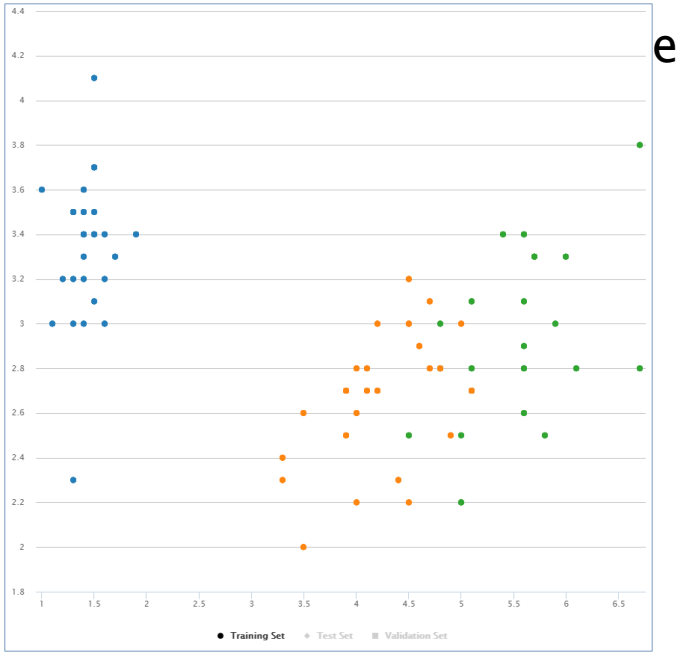


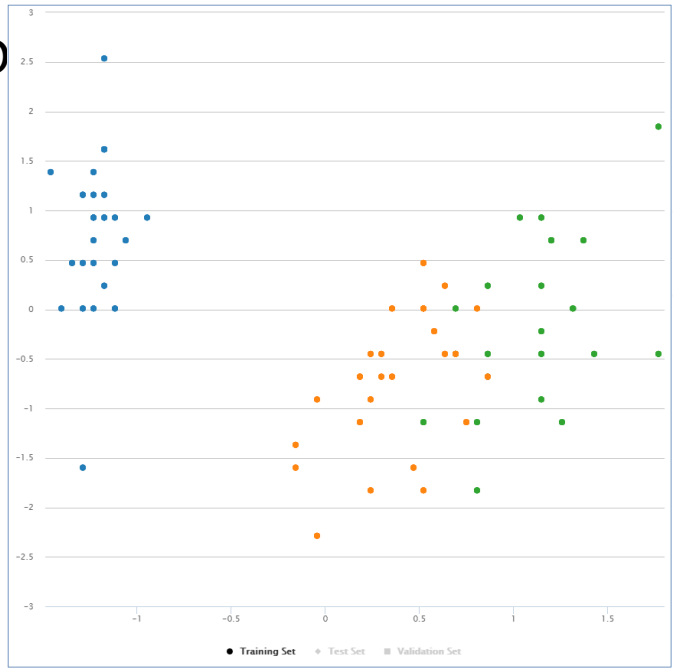
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

Min-Max Scaling

- Bringt die Trainingsdaten auf ein vorgegebenes Intervall.
- Macht die Dimensionen vergleichbar.
- Nachteil: Ausreißer beeinflussen die Skalierung stark.
- Nachteil: Validierungsdaten können außerhalb des Intervalls liegen.
- Nachteil: Für viele Algorithmen unbrauchbar, z.B. SVM, kNN

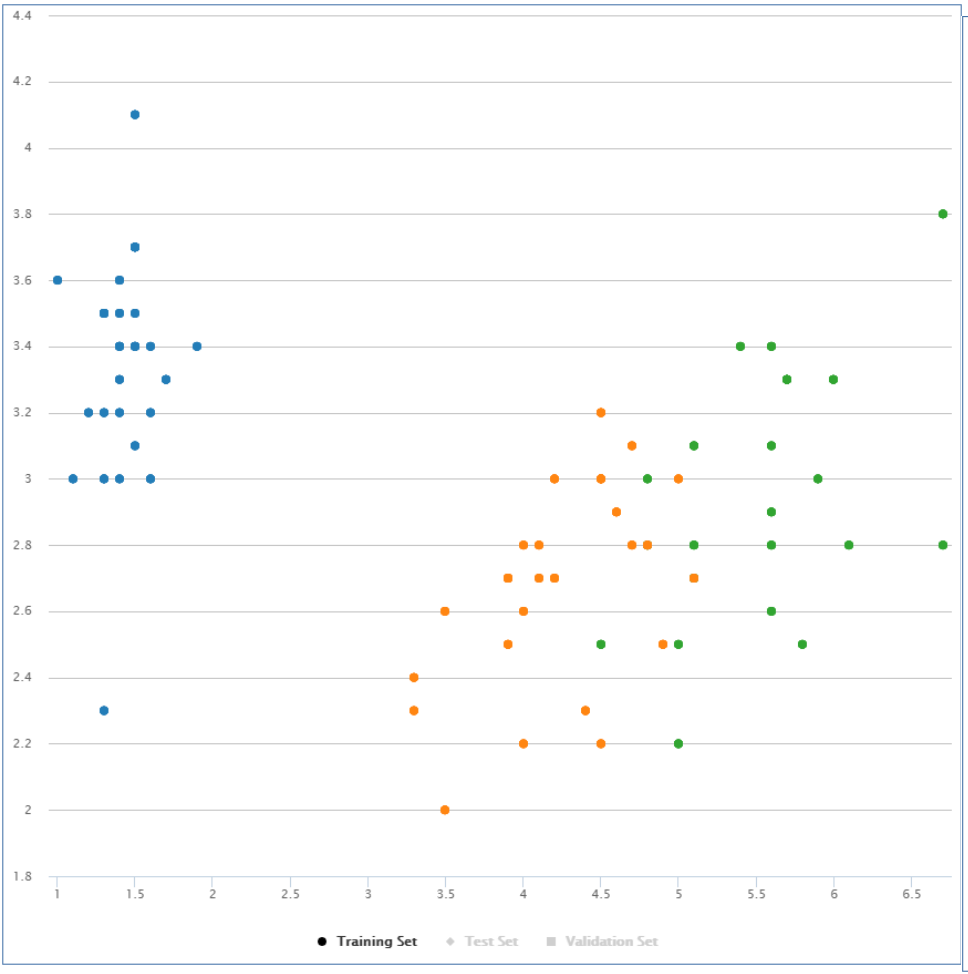
en Netzen: Brauchen D

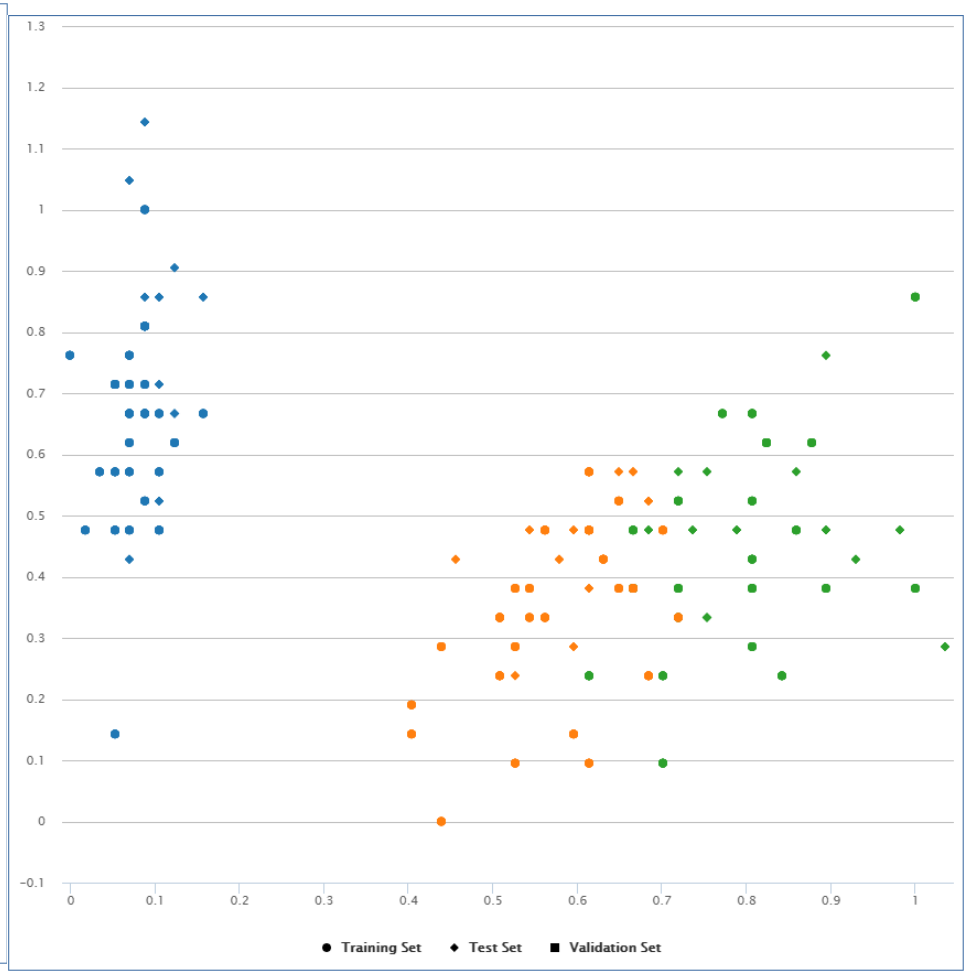




OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

Min-Max Scaling





Vielen Dank für Ihre Aufmerksamkeit!

www.ovgu.de