Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

Übung 10 – Bias-Variance-Tradeoff und Unüberwachtes Lernen

Dr. Nikolai Stein
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

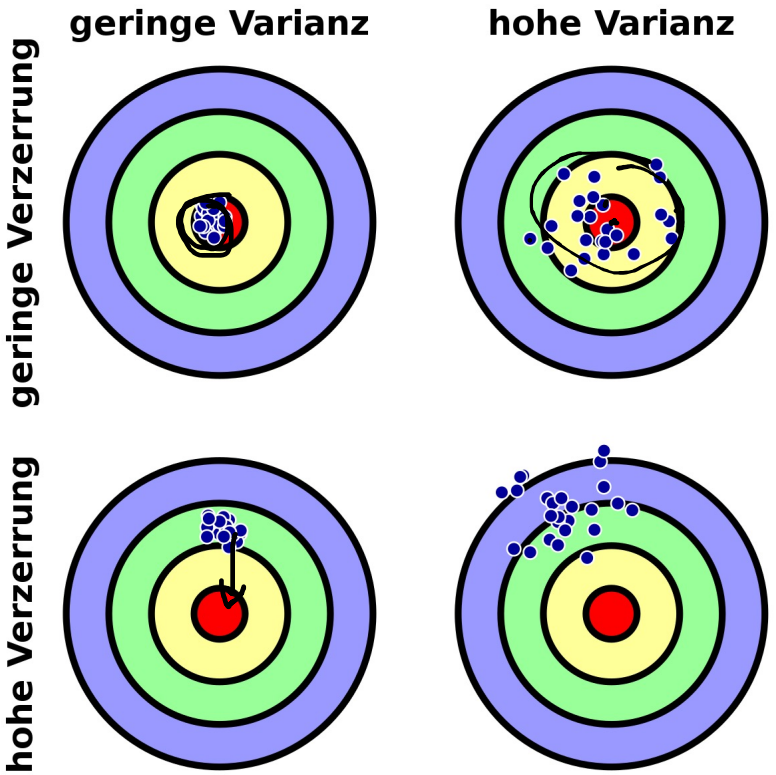
Sommersemester 2021

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Verzerrung-Varianz-Dilemma

- Es gibt allgemein zwei Fehlerquellen für Vorhersagen eines statistischen Modells:
  - Verzerrung (Bias): systematische Fehler in den Vorhersagen über alle Instanzen – das Modell liegt daneben
  - Varianz: Fehler durch Schwankung in der Vorhersagegüte über die verschiedenen Instanzen – das Modell ist manchmal besser, manchmal schlechter

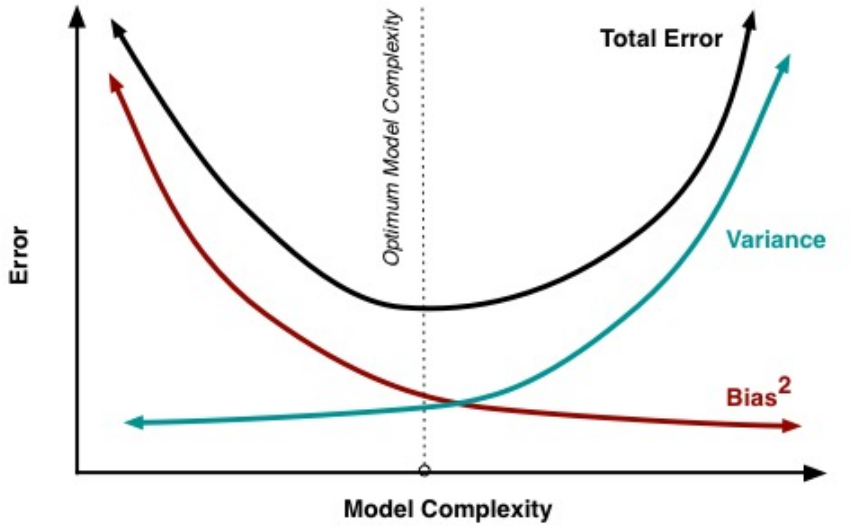
- Gemeinhin laufen die beiden Ziele entgegen:
  - Komplexere Modelle machen geringere systematische Fehler aber Fehler streuen stärker über die Instanzen
  - Einfache Modelle machen große systematische die homogen über die Instanzen sind



Verzerrung-Varianz-Tradeoff

Unteranpassung
Underfitting

Überanpassung
Overfitting



Strategien zur Vermeidung von Überanpassung

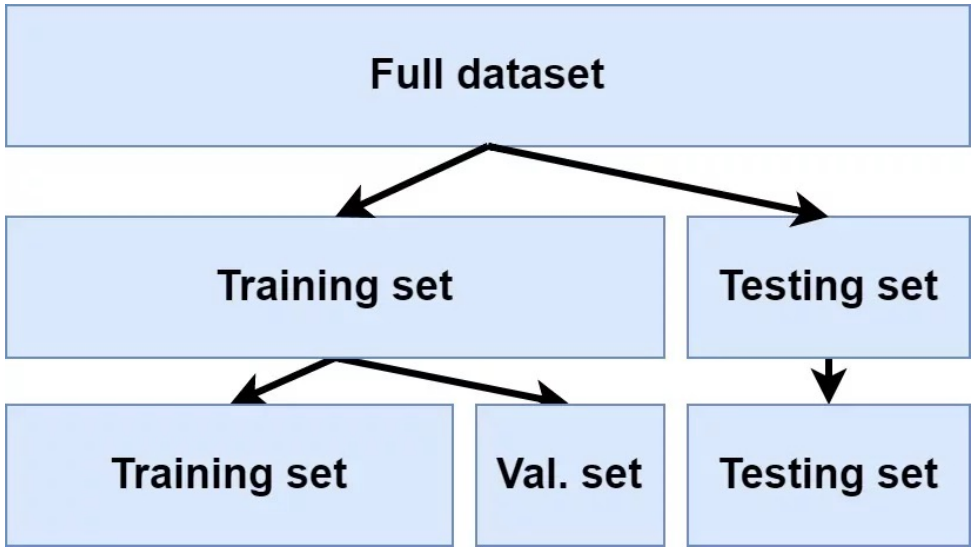
- Testdatenmanagement
- Regularisierung
- Ensemblemethoden

Nichtnutzung von Daten erlaubt eine neue Form der Evaluation

- Anstatt die kompletten Daten zu erklären zu wollen können wir auf Teildaten ein Modell lernen und dann auf den restlichen Daten objektiv dessen Güte bewerten: Train-Test Split

- Die Trainingsdaten werden wiederum in Training und Validation geteilt um eine Bewertung von Konfigurationsvarianten (andere Algorithmen, andere Parameter) zu ermöglichen

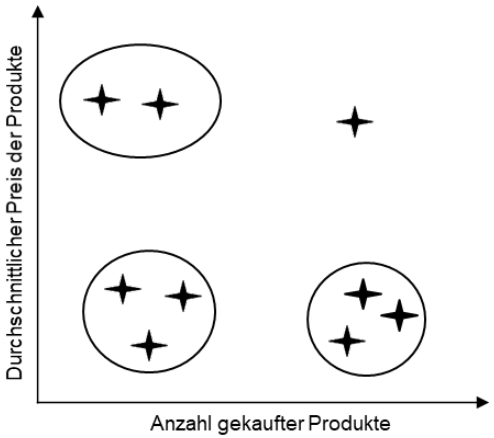
- Das Test Set dient nur der Bewertung der Generalisierungsfähigkeiten – es wird nicht für das Training benutzt
  - Insbesondere dürfen auf Basis der Testing Ergebnisse KEINE Anpassungen an den Algorihmen erfolgen



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

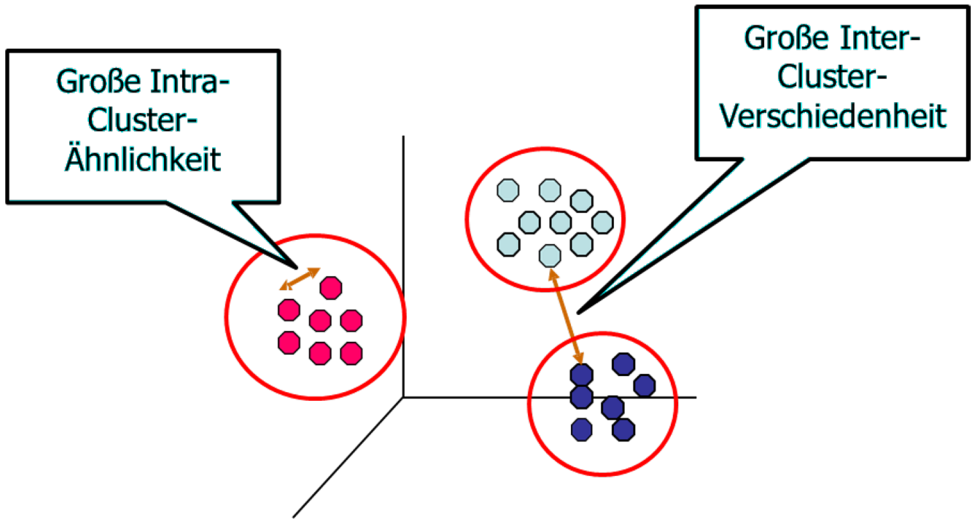
Clusteranalyse

- Idee
  - Bestimmung von Gruppen ähnlicher Tupel in multi-dimensionalen Datensätzen.
  - „Klassifizieren ohne die Klassen vorher zu kennen“ (nichtüberwachtes Lernen).
- Beispiel
  - Identifizieren von Kundengruppen zum Design passender Produkte.



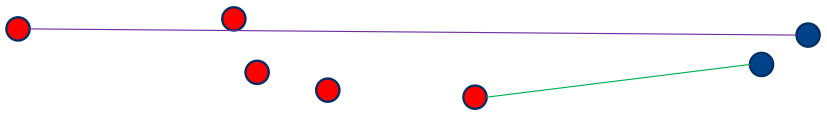
Eigenschaften „guter“ Cluster

- Als Ähnlichkeitsmaß wird oft Abstand verwendet
  - Geringer Abstand: Objekte sind ähnlich
  - Großer Abstand: Objekte sind unähnlich

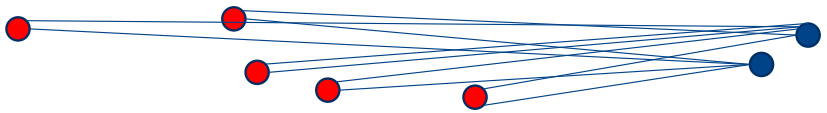


Berechnung des Abstands zwischen Clustern

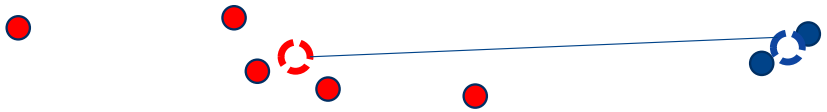
- **Single (complete) linkage**
  - Kleinster (größter) Abstand zweier Punkte von zwei Clustern. Sehr anfällig bei Ausreißern/Rauschen.



- **Average linkage**
  - Mittlerer Abstand zwischen allen Punkten von zwei Clustern. Sehr rechenaufwändig.



- **Centroid**
  - Abstand der Cluster-Zentren zueinander. Ähnlich wie Average, aber weniger aufwändig.



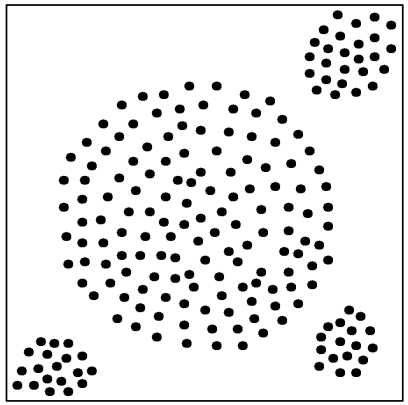
- **Medoid**
  - Abstand der zentralsten Repräsentanten der Cluster zueinander.
  - Ähnlich wie Centroid, aber weniger Anfällig für Ausreißer/Rauschen.

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

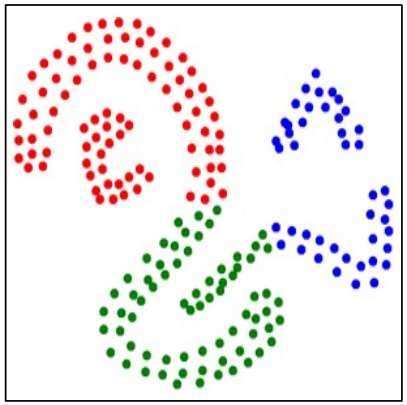
Probleme mit abstandsbasiertem Clustering

- Oft werden nur konvexe Cluster gefunden
  - Abhängig von Berechnungsmethode zum Abstand zweier Cluster (letzte Folie).
  - Single link findet am ehesten konkave Cluster.
  - Mögliche Lösung: dichtebasierte Maße

- Hohe Dimensionen / dünn besetzte Attribute verringern die Bedeutung des Abstands.
  - Ein Ansatz: Subspace-Clustering
  - Suche Cluster unter Verwendung eines Subsets von Attributen.
  - Subsets schwer zu bestimmen, wenn Bedeutung der Attribute unbekannt.
  - Achtung: Nicht unbedingt globale Lösung!

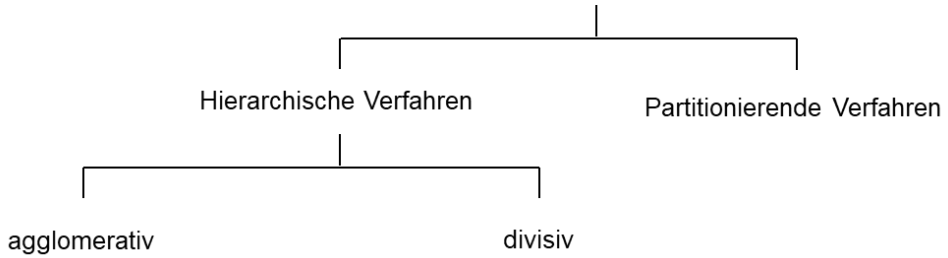


Konvexe Punktwolken



Konkave Punktwolken

Kategorisierung von Clustering-Methoden



- Partitionierende Verfahren (# Cluster fest)
  - Zunächst: Zufällige Clustereinteilung
  - Dann: Sukzessiver Wechsel der Clusterzugehörigkeit der einzelnen Tupel
- Hierarchische Verfahren (# Cluster variabel)
  - Agglomerative Verfahren
    - Zunächst: Jedes Tupel ein Cluster
    - Dann: Schrittweise Zusammenfassung ähnlicher Cluster
  - Divisive Verfahren
    - Zunächst: Alle Tupel in selbem Cluster
    - Dann: Schrittweise Aufteilung in möglichst unähnliche Teilcluster

Partitionierendes Clustering: k-means

- Gegeben:
  - Lerndatensatz
  - Anzahl der zu findenden Cluster k
  - Abstandsmaß
- Gesucht:
  - Partitionierung des Datensatz in k Cluster

Algorithmus

1. Wähle zufällig *k* Cluster-Zentren
2. Weise allen Objekten Cluster zu, basierend auf den Abständen zu den Cluster-Zentren
3. Berechne die neuen Zentren der Cluster
4. Wiederhole Schritt 2 und 3 bis Zentren stabil

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Bemerkungen zu k-means

- Sehr unterschiedliche Ergebnisse möglich.
  - Ggf. lokales Minimum bei “schlechtem” Seed.
  - Abhilfe: Mehrfache Ausführung, unterschiedliche Seeds.

- k wird nicht automatisch ermittelt
  - 1.) Mehrfache Ausführung mit verschiedenen k
  - 2.) Bestimmung der Ähnlichkeit von jedem Cluster
  - 3.) Entscheidung für k mit homogensten Clustern

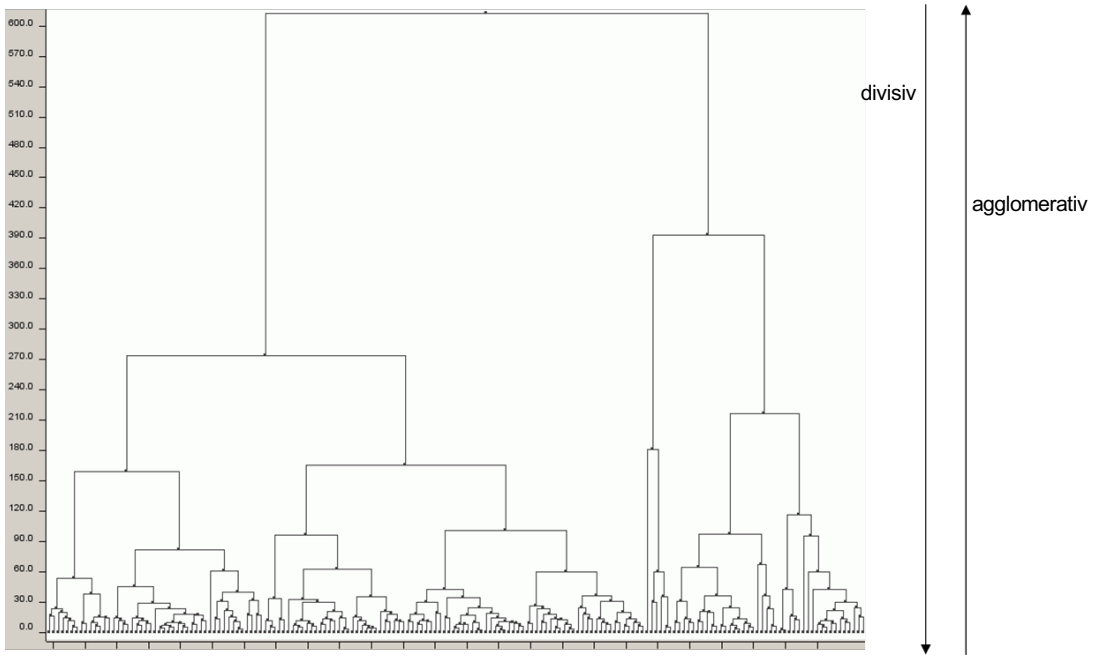
- Anfällig für Outlier und Rauschen, da Verwendung von Mittelwerten.
  - *k*-medoids anstatt von *k*-means (reale Repräsentanten für die Cluster anstatt rechnerischer Mittelpunkte).

Hierarchisches Clustering: Dendrogramme

- Ein Dendrogramm ist ein Binärbaum.
  - Wurzel repräsentiert alle Tupel.
  - Blätter repräsentieren die einzelnen Tupel.

- Nachfolgerknoten repräsentieren die zwei Cluster, aus denen der Vorgängerknoten entstanden ist bzw. in die er aufgespalten wurde.
  - Jede horizontale Ebene ist eine Clustereinteilung.

- Die Kantenlänge repräsentiert die Unähnlichkeit der Cluster.
  - Vorteil: Anzahl der Cluster kann mit Hilfe eines Dendrogramms bestimmt werden.
  - Die Ebene mit der längsten Kante wird als Clustereinteilung genommen (Maximierung der Unähnlichkeit).



Hierarchisches Clustering

- Es wird ein geeignetes Abbruchkriterium benötigt.
  - k ist nur eine, einfache Möglichkeit.
  - Besser: Dann abbrechen, wenn Unterschied zwischen Supercluster und Kindclustern besonders groß.
  - Es kann auch die gesamte Hierarchie berechnet werden um anschließend die Anzahl der Cluster zu bestimmen.

- Agglomeratives Clustering
  - Gegeben: Lerndatensatz, Abstandsmaß,
  - Strategie zur Cluster-Abstand-Berechnung
  - Wichtige Datenstruktur: Matrix mit Distanzen zwischen den Clustern
  - Gesucht: Gesamte Cluster-Hierarchie (Dendrogramm)

Algorithmus Agglomeratives Clustering

1. Berechne die Distanz-Matrix
2. Betrachte jeden Datenpunkt als Cluster
3. Vereinige die zwei Cluster mit minimalem Abstand
4. Aktualisiere die Distanz-Matrix
5. Wiederhole Schritt 3 und 4 solange mehr als ein Cluster

Anschließend: Bestimmung der Anzahl der Cluster mittels Dendrogramm.