UNIVERSITÄT
LEIPZIG

Data Mining

INTERACTIVE VISUAL DATA MINING

Data Mining

- Warum?
  - Überbordernde Menge an Daten
  - Wachsende Lücke zwischen Produktion und Auswertung der Daten
- Ziel:
  - Extraktion von nützlichen und versteckten Informationen
- Methode:
  - Mustererkennung in den Daten
  - Nutzung von verschiedenen Data Mining Verfahren, um die Muster in den Daten (semi-)automatisch zu finden

Data Mining

- Data Mining ist definiert als Prozess zum Erkennen von Mustern in Daten
  - Muss automatisch oder zumindest semi-automatisch sein
  - Die gefundenen Muster müssen Informationen enthalten, die nützlich sind
  - Die gefundenen Muster ermöglichen nicht triviale Aussagen über neue Daten
  - Daten sind in großer Menge vorhanden

- Methodenarten:
  - Black Box: Innere Abläufe sind nicht einfach nachvollziehbar
  - Transparente Box: Zeigt die Struktur eines Musters offen und nachvollziehbar an

Data Mining

- Strukturierte Muster
  - Bilden die Struktur eindeutig ab
  - Sind untersuchbar
  - Sind begründbar
  - Sind nützlich für zukünftige Entscheidungen

Data Mining

Alter	Fehlsichtigkeit	Hornhaut-verkrümmung	Tränen-produktion	Empfohlene Linsen
Jung	Kurzsichtig	nein	reduziert	keine
Jung	Kurzsichtig	nein	normal	weich
Jung	Kurzsichtig	ja	reduziert	keine
Jung	Kurzsichtig	ja	normal	hart
Jung	weitsichtig	nein	reduziert	keine
Jung	weitsichtig	nein	normal	weiche
…	…	…	…	…

Data Mining

- Extrahierte Regeln:
  - Wenn Tränenproduktion = reduziert
    -> Empfehlung = keine
  - Ansonsten:
  - Wenn Alter = jung und
    Hornhautverkrümmung = nein
    -> weich
  - Darstellbar als Entscheidungsbaum

- Probleme:
  - Regeln bilden nur den Datensatz ab
  - Vollständige Datensätze sind aber nicht die Norm
  - Reale Datensätze weisen Lücken aufgrund von Messfehlern auf
  - Reale Daten beinhalten Fehler

Data Mining

- Weitere Probleme:
  - Die Anzahl der Regeln sollte nicht größer als die Anzahl der Datenpunkte sein
  - Die Aufzählung aller Regeln ist oft nicht erreichbar

  à Filtern einer Menge von Beschreibungen
    - Visualisierung?

- Bias
  - Wie beschreibt man die Daten?
  - In welcher Reihenfolge sucht man?
  - Overfitting?

Data Mining

- Beschreibungsprobleme
  - Weist die Sprache Einschränkungen auf, welche Muster erkannt werden können?
    - z.B. kein ODER
  - Nutzung von Vorwissen
    - Einteilung des Suchraums?

Data Mining

- Suchbias
  - Schwierigkeiten, den gesamten Suchraum abzuarbeiten
  - Nutzung von Heuristiken
    - Greedy Algorithmen
- Reihenfolge kann wichtig sein
  - Top down
  - Bottom up
- Instanz basierte Methode lernen einzelne Beispiele und generalisieren

Data Mining

- Overfitting
  - Zu genaue Training Daten führen zu überspezifischen Modellen!
  - Alle Datenpunkte als Regel zu lernen daher zu spezifisch

    à Kombination von Regeln

- Bottom Up Idee:
  - Starte mit einfachen Regeln
  - Erstelle darauf aufbauend komplexere Regeln
  - Stop, wenn die Regeln komplex genug sind
- Vermeidung von Overfitting
- Nennt man auch forward-pruning
  - Geht auch rückwärts
  - Backward-pruning
  - Start mit generellen Regeln zu simplen Regeln

Konzepte, Instanzen und Attribute

- Konzeptbeschreibungen: Das, was man lernen will
  - Verständlich
    - Nachvollziehbar
    - Diskutierbar
  - Anwendbar
    - Kann man auf Beispiele anwenden
- Instanzen
  - Einzelnes Beispiel für das zu lernende Konzept
  - Hintergrundwissen kann wichtig sein
- Attribute
  - Jede Instanz wird durch Attribute beschrieben
  - Data Mining bearbeitet zumeist numerische oder kategorische Daten

Konzepte

- Lernverfahren:
  - Klassifikation
  - Assoziation
  - Clustering
  - Numerische Vorhersage
- Output sind Konzeptbeschreibungen

Konzepte

- Klassifikation
  - Bsp. Kontaktlinsen:
    - Aggregiert die Daten
    - Gibt Empfehlungen, welche Linsen genutzt werden sollen

- Supervised Lernverfahren
  - Stellt iterativ neue Ergebnisse bereit für jeden neuen Datenpunkt
  - Ergebnis ist die „Klasse“ der Beispiele
  - Erfolg des Lernverfahrens kann beurteilt werden

Konzepte

- Asoziationen
  - Offenlegung von interessanten Strukturen in Daten
  - Kann jedes Attribut Vorhersagen, nicht nur die Klasse
  - Kann mehr als einen Attributwert vorhersagen
  - Es gibt wesentlich mehr Assoziationsregeln als Klassifikationsregeln
- Regeln basieren oft einem Minimum an notwendigen Daten
- Regeln brauchen oft eine minimales Level an Genauigkeit
- Notwendige Prüfung, ob die Regeln sinnvoll sind
- Typischerweise für kategorische Daten

Konzepte

- Clustering
  - Keine spezifischen Klassen
  - Gruppiert Daten in ähnliche Gruppen
  - Herausforderungen:
    - Finden der Cluster
    - Zuordnung von neuen Instanzen
- Wie viele Cluster braucht man?
- Subjektive Analyse, ob das Ergebnis sinnvoll ist
- Kann mit nachträglicher Klassifikation kombiniert werden, um neue Instanzen hinzuzufügen

Konzepte

- Numerische Vorhersage
  - Variante der Klassifikation
  - Ergebnis ist ein numerischer Wert statt eine Kategorie
  - Vorhersage von neuen Instanzen eher unwichtig
- Wichtiger ist meist die Struktur der Beschreibung
  - Wichtige Attribute
  - Relation der Attribute

Instanzen

- Input von ML Verfahren ist eine Menge von Instanzen
- Instanz
  - Einzelnes, unabhängiges Beispiel eines Konzeptes
  - Charakterisiert durch die Attribute
- Beispiel Datenbanken:
  - Viele Tabellen sind durch Relationen verbunden
  - Für ML werden diese in eine Tabelle gepresst, damit alle Attribute in einer Tabelle stehen

Instanzen

- Input von ML Verfahren ist eine Menge von Instanzen
- Instanz
  - Einzelnes, unabhängiges Beispiel eines Konzeptes
  - Charakterisiert durch die Attribute
- Beispiel Datenbanken:
  - Viele Tabellen sind durch Relationen verbunden
  - Für ML werden diese in eine Tabelle gepresst, damit alle Attribute in einer Tabelle stehen
    - Denormalisierung

Instanzen

- Beispiel Denormalisierung:
  - Tabelle 1: Name, Geschlecht, Eltern 1, Eltern 2
  - Tabelle 2: Erste Person, Zweite Person, ist Schwester?

- Denormalisiert:
- Erste Person: Name, Geschlecht, Eltern 1, Eltern 2
- Zweite Person: Name, Geschlecht, Eltern 1, Eltern 2
- Braucht man „ist Schwester?“ noch?

Instanzen

- Suche nach Relationen
  - Z.B. nach Vorfahren
  - Können beliebig lange Pfade werden
  - Rekursion oder als Teilfeld
    Induktive logische Programmierung
  - Schwierig bei fehlerhaften Daten

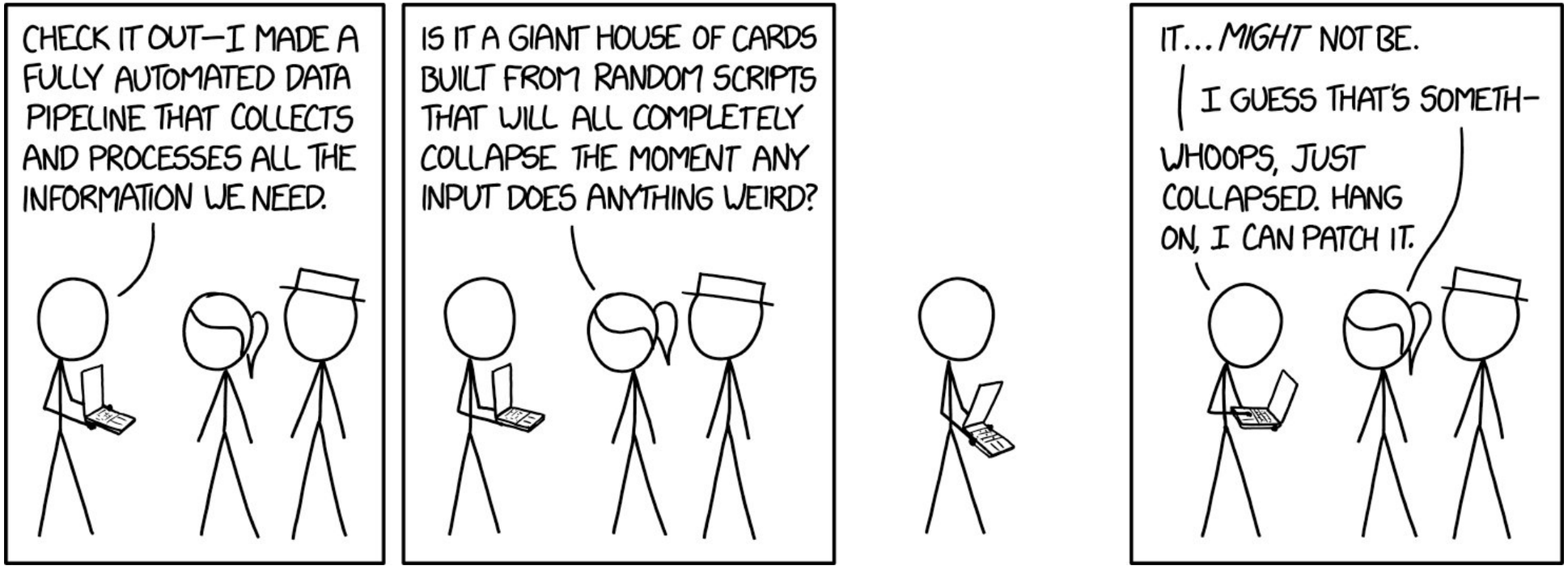
Attribute

- Feste, vordefininierte Menge von Merkmalen
- Verschiedene Attribute können in verschiedenen Instanzen gültig sein
- Beispiel Fahrzeuge
  - Landfahrzeuge: Anzahl der Räder
  - Schiffe: Anzahl der Masten
- Standardlösung: Markiere Attribute als irrelevant, wenn sie nicht anwendbar sind
- Es kann Abhängigkeiten zwischen Attributen geben
  - „Geburtsname“ und „Familienstand“

Attribute

- Attributarten:
  - Kategorial
    - Ungeordnet
    - Z.B. Apfel, Birne
  - Ordinal
    - Geordnete Kategorien
    - T-Shirt in S, M, L, XL, XXL
  - Quantativ
    - Unterstützen arithmetische Operationen
    - 1, 2, 3, 4, 5
- Ordnungen:
  - Sequentiell
  - Divergierend
  - Zyklisch

Data Preparation



Data Preparation

- Benötigt häufig am meisten Zeit
- Daten sind meistens von enttäuschend schlechter Qualität
- Datenaufbereitung ist zwingend notwendig!
- Schlechter Input resultiert in schlechten Output

- Sind die Daten dokumentiert?
- Gibt es fehlende Werte?
- Gibt es fehlerhafte Werte?
- Wie sind Strings codiert?
- Ist das Format konsistent?

Data Preparation

Verknüpfung von Daten:

- Aus verschiedenen Quellen
  - Syntax der Quellen
  - Zeitbereiche
  - Primärschlüssel
  - Fehlerarten
  - Aggregationsstufen

- Data Warehouse
  - Datenbankintegration
  - Nützlicher Schritt vor der Analyse
- Nutzung von Schematemplates für Daten
  - Daten haben definierte Syntax

Data Preparation

Sparse Data

- Tabellen beschreiben vollständige Daten gut
- Z.B. Speicherung von Graphen kann in einer sparse besetzten Adjazenzmatrix resultieren
  - Speicherplatzprobleme

- Speicherung in Key Value Listen
  - Graphen: Edgelist mit Knoten, Knoten
  - Wörter in Dokument: Position, Wort
  - Aber missing values müssen explizit markiert werden

Data Preparation

Verarbeitung von Attributen

- Numerische Werte können als natürlicher oder reele Zahlen interpretiert werden
- Normalisierung kann das Intervall eingrenzen
  - Nützlich für reele Zahlen
- Normalisierung zwischen [0 ,1]

$$v' = \frac{v - min}{\max - min}$$

- Standardisierung
  - Berechne Mittelwert *m* und Standardabweichung *d* der Daten

$$v' = \frac{v - m}{d}$$

- Kategorische Daten können als numerische Werte interpretiert werden
  - Markiere nominale Daten!

Data Preparation

Missing values

- Werden oft mit out of range Einträgen codiert
  - Numerisch: -1, 999, etc
  - Kategorisch: “ “, “-”
- Unterschiedliche Arten von missing values:
  - Unbekannt
  - Nicht aufgezeichnet
  - Irrelevant

- Fehlende Werte müssen bewertet werden
  - Was ist der Grund?
  - Ist der fehlende Wert selber eine wichtige Information?
  - ML Verfahren behandeln fehlende Werte meistens ohne Bedeutung

Data Preparation

Fehlerhafte Werte

- Falsch geschriebene Werte
- Abkürzungen vs. vollständige Namen
  - Personennetzwerke basieren auf Namen
  - D. Zeckzer <-> Dirk Zeckzer
- Numerische Werte können Messfehler haben
- Duplikate
- Veraltete Daten
  - Daten können sich verändern

Daniel Gerighausen <-> Daniel Wiegreffe

Data Preparation

Zusammenfassung

- Datenaufbereitung ist extrem wichtig
- Muss vor der eigentliche Analyse gemacht werden
- Analyse der Verteilungen, z.B. mit Histogrammen
- Visualisierung der Daten, um Outlier zu erkennen
- Sehr aufwendig und zeitintensiv
- Muss sauber dokumentiert werden
- Domänenwissen ist notwendig
  - Fehlerhafte Werte erkennen
  - Verteilungen sinnvoll?