Applied Analytical Data Science

Teil 2: KDD-Process, Data Understanding

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

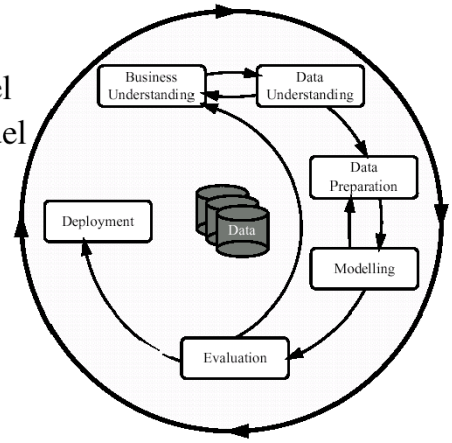
http://www.kietz.ch/AADS/

Heutiges Programm

KDD-Prozessmodel

- Was ist ein KDD-Prozessmodel
- Wozu dient ein KDD-Prozessmodel
- Das CRISP-DM KDD-Prozessmodel
  - Business understanding
  - Data understanding
  - Data preparation
  - Modelling
  - Evaluation
  - Deployment

Data Understanding



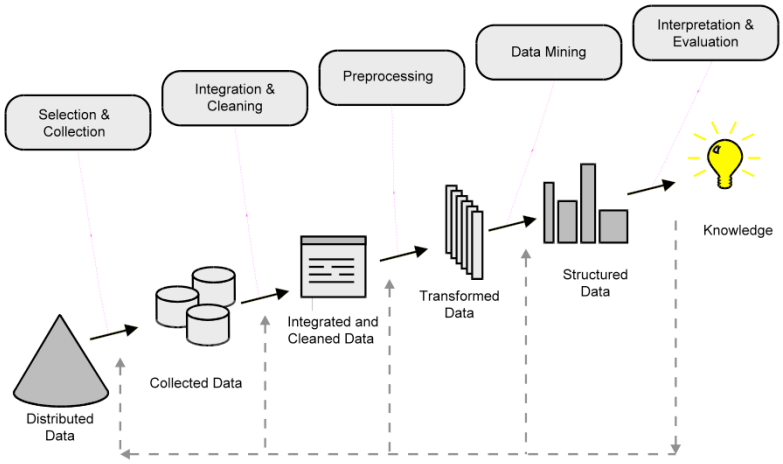
Was ist ein Prozessmodel

Prozessmodelle als formales Beschreibungsmittel für komplexe Abläufe:

- Zerlegung in überschaubarere, einfachere Teilschritte
- Abstraktion/Hierarchisierung und Parametrisierung erlauben Zusammenfassung und Wiederverwendung von Ablaufbeschreibungen
- Prozessmodelle stellen ein Framework dazu bereit und geben Strukturierung vor
- Verfeinerung/Konkretisierung des Prozessmodells (z.B. mit Operatoren) und Verbindung mit Modellelementen zur Beschreibung von Daten, Ergebnissen, Zielen erlaubt generische Implementierung von Prozessentwurfs- und Ausführungsumgebungen

⇒ Business Prozessmodellierung (engineering)

Was ist ein KDD-Prozessmodel



Wozu dient ein KDD-Prozessmodel

- Vereinfachung/verbesserung von
  - Planung
  - Durchführung, und
  - Controlling

  von KDD-Projekten
- Hinweis auf die üblichen Probleme von KDD-Projekten
- Definition des Formats von Ergebnissen von KDD-Projekten

⇒ Projekthandbuch für KDD-Projekte

⇒ Entwicklung von KDD Support Environments (KDDSE‘s)

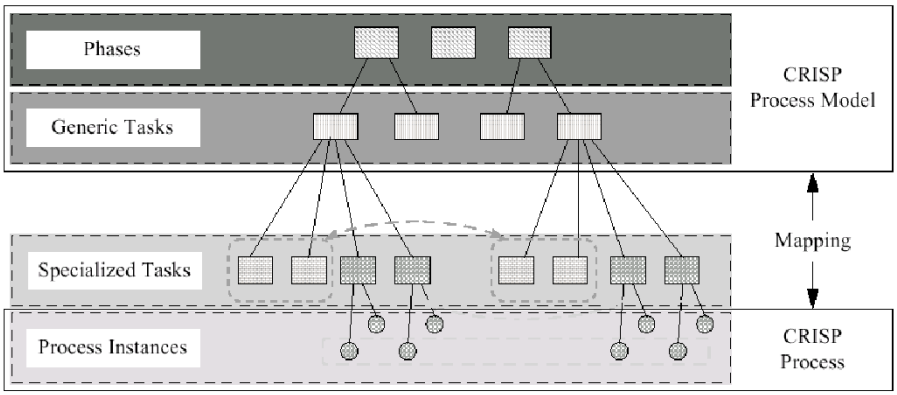
Das CRISP-DM KDD-Prozessmodel

- CRISP-DM ist ein hierarchisches Prozessmodell auf 4 Abstraktionsebenen:
  (1) Phasen oberste Ebene der Prozesszerlegung in 6 Phasen
  (2) Generische Aufgaben als Untergliederungen der Phasen
    - Umfassen den Gesamtprozess (complete)
    - Decken alle möglichen Anwendungen ab (stable)
    - Haben definierte Ergebnisse

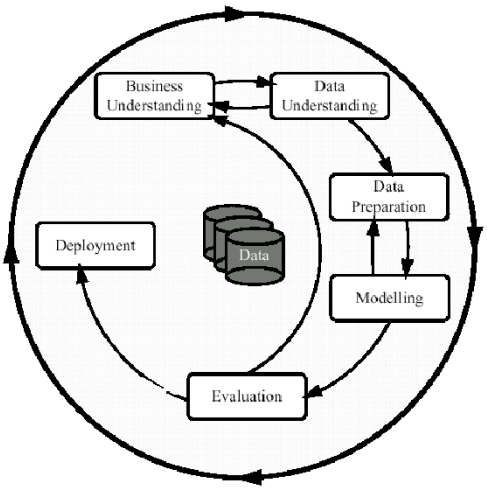
  (3) spezialisierte Aufgaben
    - Abbildung von generischen Aufgaben auf situationsbezogene, spezialisierte Aufgaben
    - bestimmt durch Data Mining Kontext

  (4) Prozessinstanzen:
    - Aufzeichnung von Aktionen, Entscheidungen und Ergebnissen eines tatsächlich ausgeführten KDD Prozesses

Die 4 Ebenen von CRISP-DM



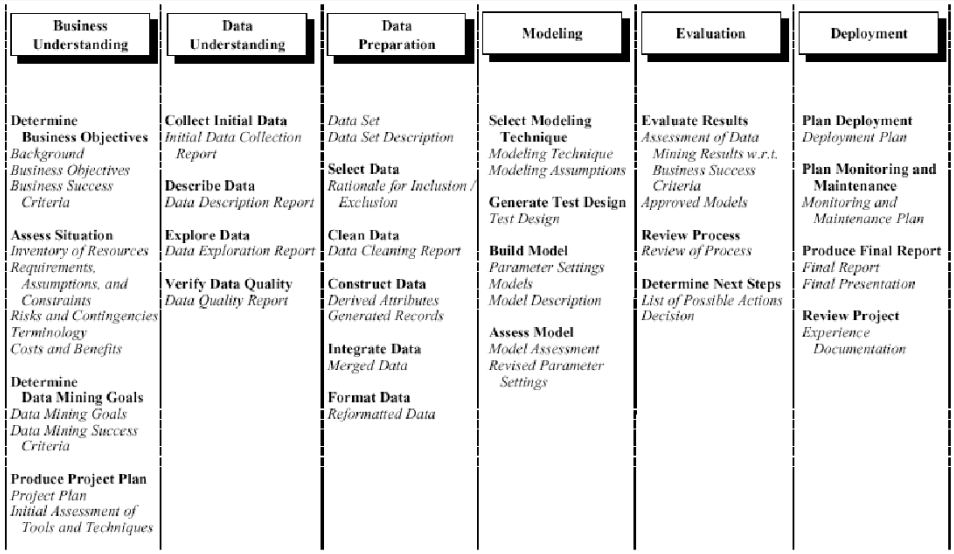
Die Phasen von CRISP-DM



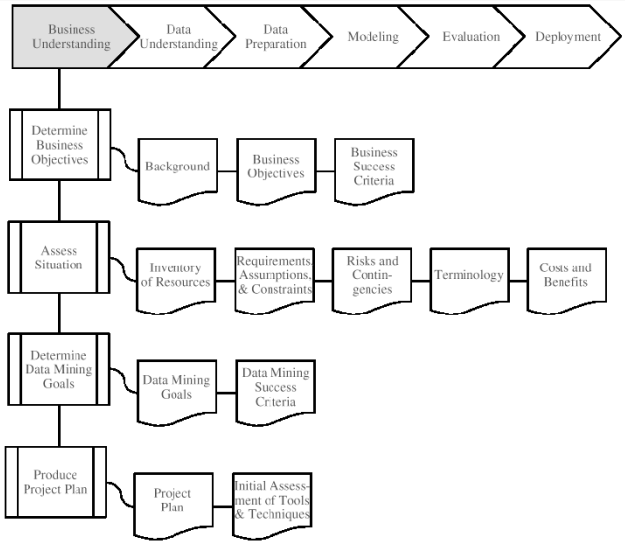
Typische Aufwandsverteilung:

5 - 20%	Business Understanding
15 - 30%	Data Understanding
40 - 70%	Data Preparation
5 - 10%	Modeling
5 - 10%	Evaluation
5 - 10%	Deployment

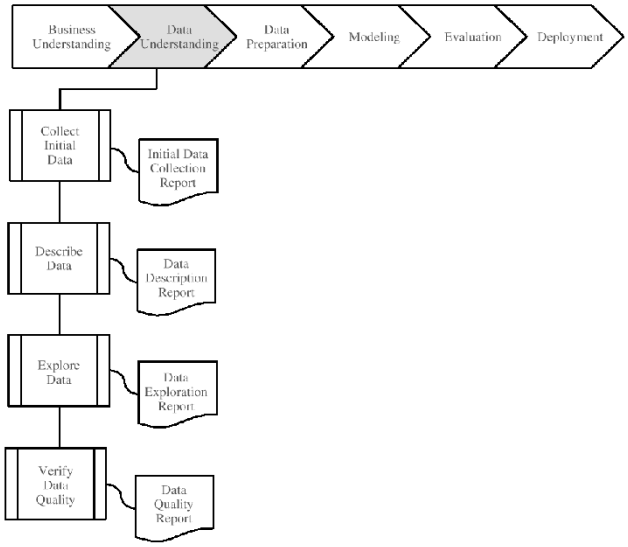
Generische Aufgaben und Ergebnisse



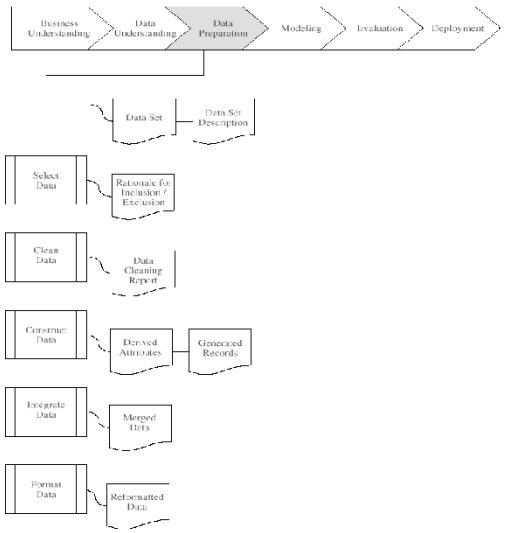
Business Understanding



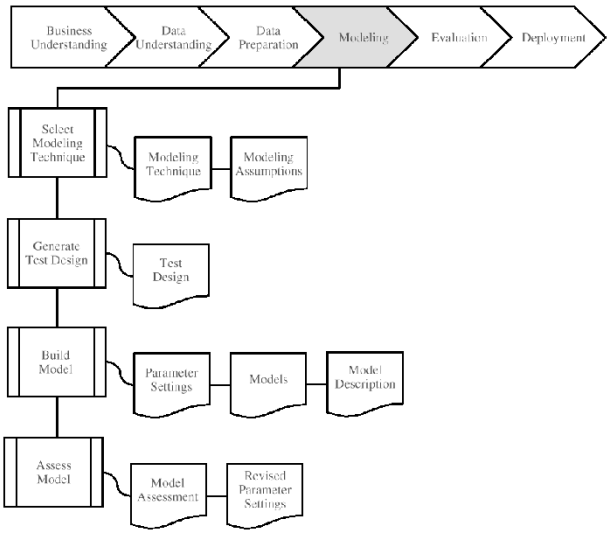
Data Understanding



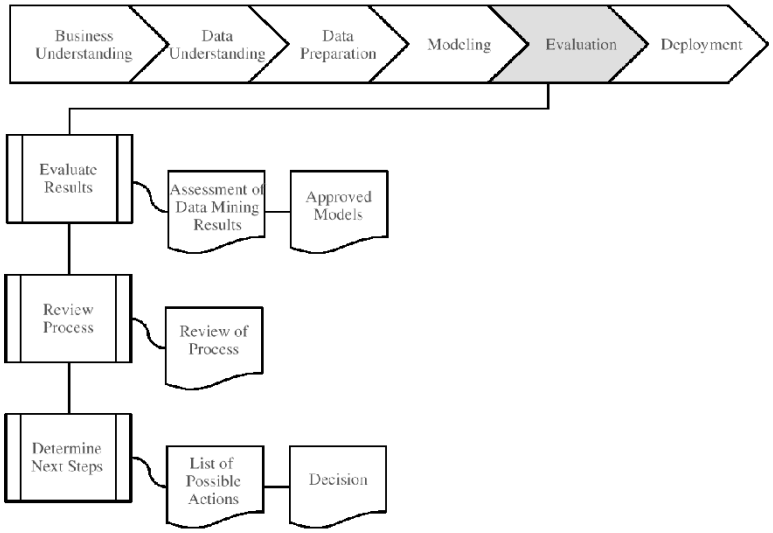
Data Preparation



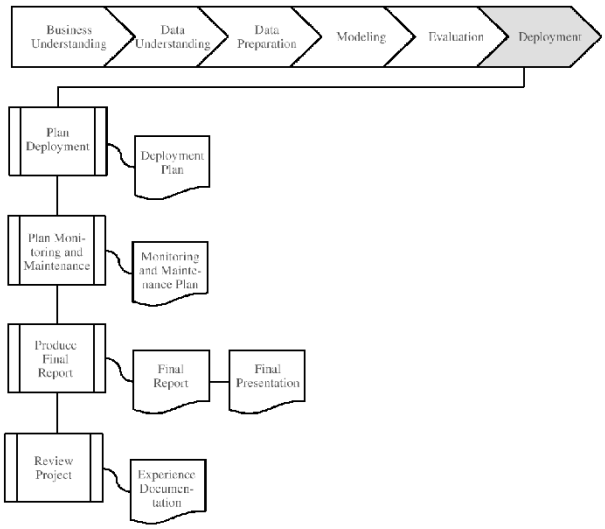
Modeling


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Select Modeling Technique
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Settings

Evaluation


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment of Data Mining Results
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision

Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

Fazit KDD-Prozess

- CRISP-DM unterteilt den KDD-Prozess in 6 Phasen:
  - Business Understanding (am wichtigsten)
  - Data Understanding
  - Data Preparation (am aufwendigsten)
  - Modeling
  - Evaluation
  - Deployment
- Vereinfachte(s) Planung, Durchführung und Controlling von KDD-Projekten
- Definition des Formats von Ergebnissen von KDD-Projekten

⇒ Projekthandbuch für KDD-Projekte

Literatur

- P. Chapman et al. CRISP-DM 1.0: Step-by-step data mining guide. ehemals http://www.crisp-dm.org, 2000. Zur Zeit z.B.: http://the-modeling-agency.com/crisp-dm.pdf

Heutiges Programm

KDD-Prozessmodel

Data Understanding

- Daten: Was ist das, Wo kommen sie her?
- Relevante Eigenschaften von Daten:
  - Eigenschaften von Attributen (DB: Spalten, Statistik: Variablen)
  - Eigenschaften von Beispielen (DB: Datensätze, Statistik: Objekte)
  - Eigenschaften von Beispielmengen (DB: Tabellen, Statistik: Populationen)
- Daten in Datenbanken und Daten fürs Data Mining
- Wie viele Daten braucht man zum Data Mining

Daten, was ist das

- Datensätze dienen zur Repräsentation einzelner Objekte.
- Spalten dienen zur Repräsentation von Eigenschaften, deren Werte für die Objekte gemessen und dann repräsentiert werden.
- Datensätze in einer Tabelle sollten Objekte gleichen Typs repräsentieren, d.h. alle den Spalten der Tabelle zugeordneten Eigenschaften sollten für alle in der Tabelle zu repräsentierenden Objekte prinzipiell messbar sein.

? Welche Objekte sollen repräsentiert werden?

? Welche Eigenschaften sollen für sie gemessen werden?

=> Festlegung durch den Anwendungskontext der Datenerhebung.

=> Bedarf im Kontext der Datenanalyse zu überprüfen.

Daten, wo kommen sie her

- In einem Unternehmen gibt es zahlreiche operative Datenbanken unterschiedlichster Art und unterschiedlichsten Inhalts, welche das tägliche Geschäft unterstützen.
- Die Datenbanken enthalten (meist implizit) wesentlich mehr Informationen, als durch die operative Arbeitsweise ausgenutzt wird.
- Diese Informationen können insbesondere der Entscheidungs-findung (Decision Support) im Unternehmen dienen.
- Aber, die Qualität der Daten hängt von ihrer operativen Nutzung ab, d.h. es wird nur das gepflegt, was auch operativ gebraucht wird.

Heutiges Programm

KDD-Prozessmodel

Data Understanding

- Daten: Was ist das, Wo kommen sie her?
- Relevante Eigenschaften von Daten:
  - Eigenschaften von Attributen (DB: Spalten, Statistik: Variablen)
  - Eigenschaften von Beispielen (DB: Datensätze, Statistik: Objekte)
  - Eigenschaften von Beispielmengen (DB: Tabellen, Statistik: Populationen)
- Daten in Datenbanken und Daten fürs Data Mining
- Wieviele Daten braucht man zum Data Mining

Eigenschaften von Attributen

- Typ des Attributes
- Verteilung der Attributwerte (univariate Statistiken)
- Anzahl fehlender Werte
- Qualität der Attribut-Wert Messungen
  - Toleranz der Messung
  - Wahrscheinlichkeit der Korrektheit der Messung

Typen von Attributen

- Nominale Attribute: Diese Attribute dienen zur Repräsentation von qualitativen Unterscheidungen / Klassifikationen, weder gibt es eine Anordnung der Werte, noch hat der Abstand zwischen den Werten eine Bedeutung.
- Skalare Attribute: Diese Attribute dienen zur Repräsentation von quantitativen Messungen, die Werte haben eine Bedeutungs-tragende Ordnung und auch der Abstand trägt eine Bedeutung.
- Ordinale Attribute: Diese Attribute dienen zur Repräsentation von qualitativen Ordnungen, d.h. die Anordnung hat eine Bedeutung, der Abstand nicht.

Nominale Attribute

Untertypen, gemäss der Anzahl der möglichen Werte und damit der **Selektivität** der Attribute:

- Schlüssel-Attribute: Diese Attribute dienen zur Benennung von repräsentierten Objekten, und damit zur (mehr oder weniger eindeutigen) Identifikation von Datensätzen.
- Kategoriale Attribute: Eine Zuordnung der Objekte in verschiedene Klassen wird repräsentiert.
- Binäre Attribute: Eine binäre Klassifikation/Entscheidung
- Konstante Attribute: Alle Datensätze haben den gleichen konstanten Wert.

Nominale Attribute und Data Mining

- Binäre und kategoriale Attribute sind gut für symbolische DM-Verfahren, d.h. sie können in Gleichheitstests benutzt werden:
  - Attribute = Wert
  - $\text{Attribute}_i = \text{Attribute}_j$

- binäre Attribute auch in numerischen DM-Verfahren benutzt werden, d.h. man kann für die binäre 0/1-Repräsentation auch Ordnung und Abstand halbwegs sinnvoll nutzen.
- Ein kategoriales Attribut mit *N* Werten kann in *N* binäre Attribute umgewandelt werden (Preprocessing).

Nominale Attribute und Data Mining

- Konstante und Schlüssel-Attribute sind **nicht brauchbar** für DM.
  - Die Grenze zwischen kategorialen und Schlüssel-Attributen ist fliessend.
    Z.B: Berufe ~ 9000 verschiedene Werte.
  - Fremdschlüssel können zur Denormalisierung benutzt werden (Preprocessing: Bereitstellung von Hintergrundwissen).
    Z.B. abstraktere Klassifikationen für Berufe.

  Aber: Das führt Abhängigkeiten (Korrelationen) ein und DM-Verfahren beruhen theoretisch auf unabhängigen Attributen.

Skalare Attribute

- Skalare Attribute repräsentieren mehr oder weniger genaue quantitative Messungen.
  Z.B. Geburtsdatum und Uhrzeit, Jahrgang, Jahrhundert der Geburt.
- Sowohl die Ordnung (Jemand ist Älter) als auch der Abstand (Jemand ist doppelt so alt) ist von Bedeutung.
- Skalare Attribute haben eine feste Skalierung.
  - Mit Masseinheit, z.B.: Grösse in cm, Einkommen in CHF
  - Ohne Masseinheit, z.B.: Verhältnis von Einkommen und Verschuldung, Verhältnis von Höhe zu Breite, ...

Skalare Attribute und Data Mining

- Die Benutzung von skalaren Attributen ist vielfältig:
  - Schwellwerte: Attribute ≤ T
  - Intervalle: $T_{min} \leq$ Attribute $\leq T_{max}$
  - Lineare Modelle: $w_1A_1 + ... + w_nA_n \leq T$
  - Nicht-lineare Modelle, z.B.: Neuronale Netze
  - Distanz- und Dichtebasierte Verfahren (Clustering und Nearest Neighbours), z.B.: geometrischer Abstand zwischen Objekten

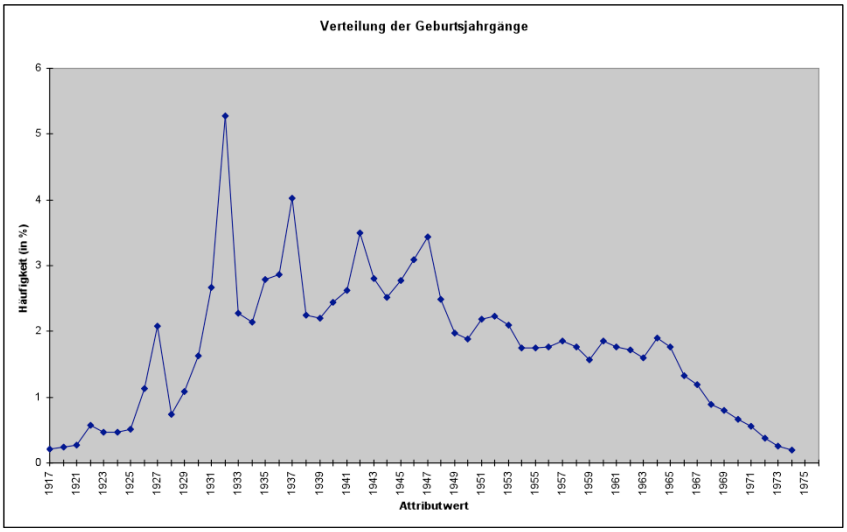
Aber:

- Gleichheit (und zu kleine Intervalle) kann zu selektiv sein.
- Die Distanz von Objekten hängt von der Skalierung ab

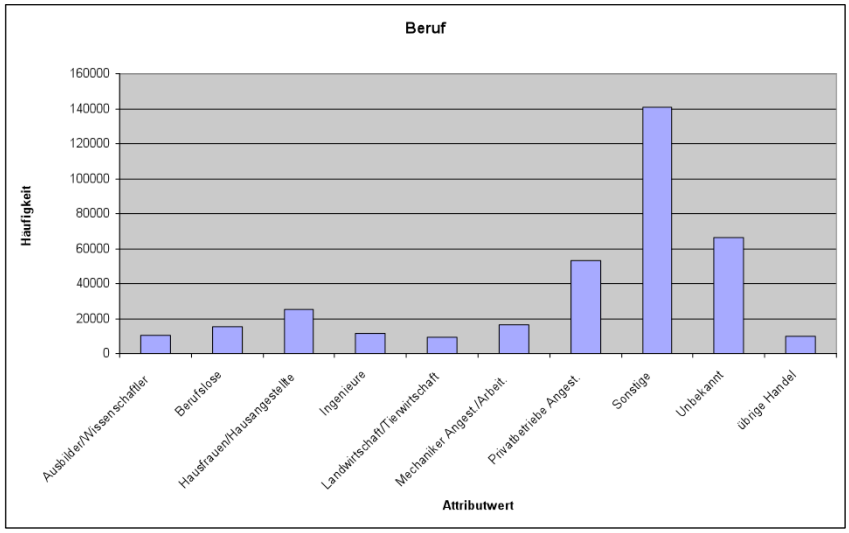
Ordinale Attribute

- Ordinale Attribute sind eine Abstraktion von skalaren Attributen. Z.B. Einkommensklasse: „gering“, „durchschnitt“, „hoch“ (oder auch „1“, „2“, „3“) als Abstraktion des realen Einkommens.
- Sie können in Gleichheits- und Ordnungstests benutzt werden:
  - Attribute = Value Attribute ≤ Value
  - $\text{Attribute}_i = \text{Attribute}_j$ $\text{Attribute}_i \leq \text{Attribute}_j$
- **Aber**: Der Abstand oder die Verrechnung in Funktionen ist **nicht sinnvoll** (auch wenn sie als „1“, „2“, „3“ codiert sind).
- Bei Ordnungstests ist die Selektivität nicht so wichtig für die Nützlichkeit fürs Data Mining, wie bei nominalen Attributen.

Verteilung der Attributwerte (skalar)



Verteilung der Attributwerte (nominal)



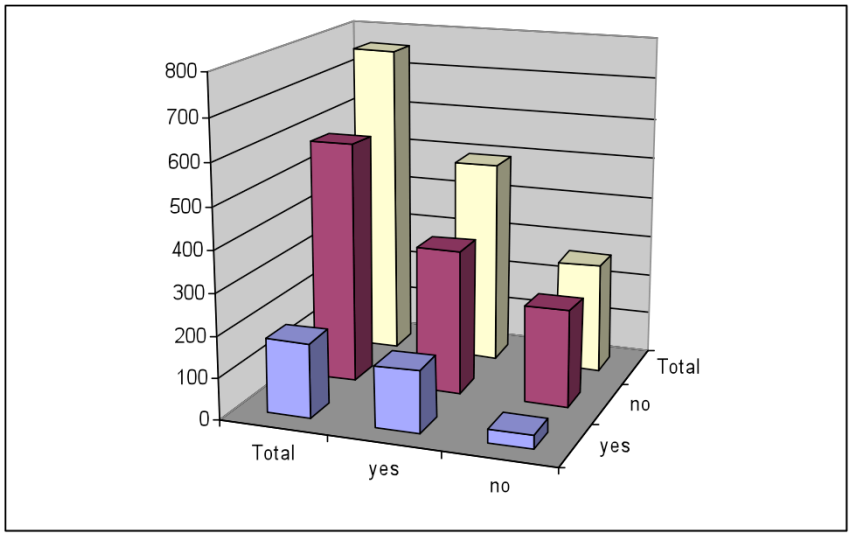
Korrelation/Unabhängigkeit von Attributen

- Data Mining „funktioniert am Besten“, wenn
  - die Eingabevariablen mit der Zielvariablen korreliert sind.
  - die Eingabevariablen unabhängig, d.h. paarweise unkorreliert sind.

Bivariate Statistik:

- nominale Attribute: χ^2-Test zum Test des Grads der Unabhängigkeit über Kontigenztabellen.
- skalare Attribute: Diskretisierung und χ^2-Test, oder Berechnung des Korrelationskoeffizienten.
- für gemischte Attribute, z.B. Diskretisierung und χ^2-Test.

Kontigenztabelle zwei binärer Attribute



Heutiges Programm

KDD-Prozessmodel

Data Understanding

- Daten: Was ist das, Wo kommen sie her?
- Relevante Eigenschaften von Daten:
  - Eigenschaften von Attributen (DB: Spalten, Statistik: Variablen)
  - Eigenschaften von Beispielen (DB: Datensätze, Statistik: Objekte)
  - Eigenschaften von Beispielmengen (DB: Tabellen, Statistik: Populationen)
- Daten in Datenbanken und Daten fürs Data Mining
- Wieviele Daten braucht man zum Data Mining

Eigenschaften von Beispielen

- Ein Beispiel (Datensatz) repräsentiert das Objekt der Analyse.

=> Die Granularität der Daten muss zur Analyseaufgabe passen.

- Es gibt zwei Gründe für fehlende Werte bei Datensätzen:
  - Unbekannte Werte, der Wert existiert, er ist nur nicht bekannt
  - Das Attribut ist auf dieses Objekt nicht anwendbar
- Qualität eines Datensatzes
  - Messfehler/Ausreisser
  - Fehlerhafte Werte

=> Datensätze mit vielen fehlenden Werten oder zweifelhafter Qualität sollten weggelassen werden.

Heutiges Programm

KDD-Prozessmodel

Data Understanding

- Daten: Was ist das, Wo kommen sie her?
- Relevante Eigenschaften von Daten:
  - Eigenschaften von Attributen (DB: Spalten, Statistik: Variablen)
  - Eigenschaften von Beispielen (DB: Datensätze, Statistik: Objekte)
  - Eigenschaften von Beispielmengen (DB: Tabellen, Statistik: Populationen)
- Daten in Datenbanken und Daten fürs Data Mining
- Wieviele Daten braucht man zum Data Mining

Eigenschaften von Beispielmengen

- Eine Beispielmenge (Population, Datenbank) wird auf eine bestimmte Weise erzeugt/gesammelt (Sampling Bias).
- Eine Beispielmenge kann repräsentative für eine Fragestellung sein oder auch nicht.
- Die Grösse einer Beispielmenge ist weniger entscheidend für ihre Repräsentativität als ihr Sampling Bias.

=> Ohne Wissen über den Sampling Bias sind Analysen von Daten nicht interpretierbar.

=> Random-Sampling ausreichender Grösse ändert die Repräsentativität einer Beispielmenge nicht.

Heutiges Programm

KDD-Prozessmodel

Data Understanding

- Daten: Was ist das, Wo kommen sie her?
- Relevante Eigenschaften von Daten:
  - Eigenschaften von Attributen (DB: Spalten, Statistik: Variablen)
  - Eigenschaften von Beispielen (DB: Datensätze, Statistik: Objekte)
  - Eigenschaften von Beispielmengen (DB: Tabellen, Statistik: Populationen)

Daten in Datenbanken und Daten fürs Data Mining

- Wieviele Daten braucht man zum Data Mining

Daten in Datenbanken und Data Mining

- Datenbanken sind für eine bestimmte Anwendung gedacht, diese bestimmt die Objekte, ihre Attribute, den Sampling Bias und die Pflegequalität der Daten, nicht jedoch die Nützlichkeit für spätere Analysen.
- Die Codierung von Daten ist nicht immer geeignet fürs Data Mining:
  - Data Mining Tools erwarten, dass skalare Attribute numerisch und nominale Attribute alphanumerisch codiert sind.
  - Dynamische Werte werden nicht gespeichert, sondern durch Anwendungsprogramme berechnet.
  - Data Mining braucht denormalisierte Daten.

Heutiges Programm

KDD-Prozessmodel

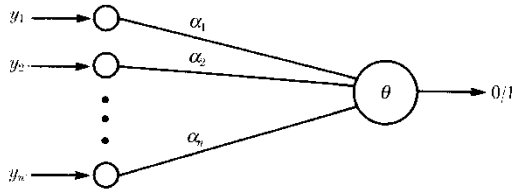
Data Understanding

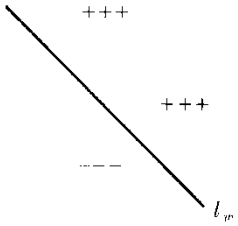
- Daten: Was ist das, Wo kommen sie her?
- Relevante Eigenschaften von Daten:
  - Eigenschaften von Attributen (DB: Spalten, Statistik: Variablen)
  - Eigenschaften von Beispielen (DB: Datensätze, Statistik: Objekte)
  - Eigenschaften von Beispielmengen (DB: Tabellen, Statistik: Populationen)
- Daten in Datenbanken und Daten fürs Data Mining
- Wieviele Daten braucht man zum Data Mining

Benötigte Menge von Daten fürs Data Mining

- Die Benötigte Menge von Daten hängt von
  - der Data Mining Aufgabe
  - der Komplexität des Ergebnisses, und
  - der benötigten Wahrscheinlichkeit der Korrektheit ab.
- Theorie des Maschinellen Lernens: Für welche **Klassifikations-systeme** sind **wahrscheinlich annähernd korrekte Ergebnisse** mit **polynomialer Anzahl von Beispielen** und **polynomialem Aufwand** zu erreichen. Sie ermöglicht unter anderem
  - Die Kapazität von Klassifikationen zu untersuchen (VC-Dimension).
  - Die Anzahl der Beispiele die notwendig zum Lernen ist zu bestimmen.

Beispiel: Das Perceptron

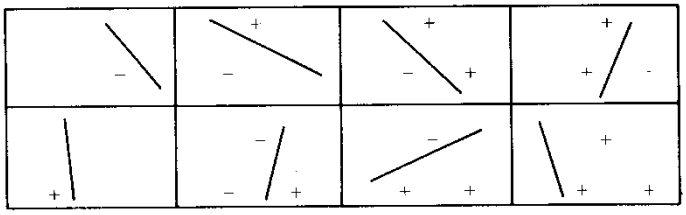




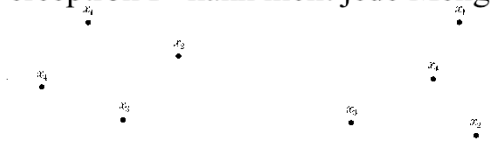
$$h_\omega(y) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} \alpha_i y_i \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

Beispiel: Die Kapazität des Perceptron

Das Perceptron P^2 kann jede Menge von 3 Punkten separieren



Das Perceptron P^2 kann nicht jede Menge von 4 Punkten separieren



Kapazitäten einiger einfacher Modellklassen

- Das Perceptron P^n hat die Kapazität: $VCdim(P^n) = n+1$
- Alle endliche Modellklassen M:
  $lg(||M||) / lg(||X||) \leq VCdim(M) \leq lg(||M||)$
- Monomials, d.h. konjunktive Begriffe über binären Attributen (Ein Vektor über {0,1,?}): $VCdim(B^n) = O(n)$
- Disjunktive Begriffe über binären Attributen
  $VCdim(B^n) = O(2^n)$

Lerntheorie: Die Menge von Beispielen ist polynomial von der Sicherheit und Qualität der Vorhersage und der Kapazität der Modellklasse abhängig.

Fazit Data Understanding

- Attribute werden unterschieden in Nomial, Ordinal und Skalar.
- Die Granularität der Beispiele muss zur Analyseaufgabe passen.
- Der Sampling Bias einer Beispielmenge ist entscheidend für die Nützlichkeit der Daten zur Analyse.
- Die Daten in Datenbanken müssen gemäss den Anforderungen der Analyseaufgabe aufbereitet werden.
- Zur Aufbereitung der Daten gehört eine adequate Denormalisierung.
- Die Theorie des Maschinellen Lernens (PAC-Learning) liefert Abschätzungen über die Menge der benötigten Daten.

Literatur

Daten fürs Data Mining:

- Pyle, D.: Data Preparation for Data Mining, Morgan Kaufmann, 1999.
- Weiss, S.; Indurkhya, N.: Predictive Data Mining - a practical guide, Morgan KaufmannPublishers Inc, 1998.