Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Problemanalye und Datenvorverarbeitung

Tobias Scheffer

Überblick

- Analyse von Lernproblemen
  - Verständnis der Anforderungen
  - Entwicklung eines Lösung
  - Entwicklung einer Evaluierungsstrategie
- Datenvorverarbeitung
  - Datenintegration
  - Merkmalsrepräsentation
  - Fehlende Werte
  - Merkmalsselektion

Problemanalyse

- Ingenieurmäßige Herangehensweise an Probleme
- Verständnis der Anforderungen
  - Zielstellung, Qualitätsmetrik
  - Eigenschaften der Daten, des Daten erzeugenden Prozesses
  - Anwendungsspezifische Anforderungen
  - Einordnung in Taxonomie von Paradigmen
  - Passen Annahmen, die bekannten Verfahren zugrunde liegen, zu Anforderungen?
- Entwicklung einer Lösung
- Entwicklung einer Evaluierungs- und Teststrategie

Verständnis der Anforderungen

- Unterschiedliche Kulturen in Industriezweigen
- Z.B. Automobilindustrie
  - 10- bis 20-seitige, ausgearbeitete Software Requirements Specifications sind die Regel
- Regelfall
  - Anwender haben ungefähre Vorstellung davon, welche Eigenschaften eine gute Lösung hätte
  - Genaue Zielstellung, Anforderungen müssen in Interviews ermittelt werden.

Fallbeispiel Email Service Provider

- Problem: Email-Spam lässt Festplatten volllaufen und erschöpft Server-Kapazitäten
- Server und Netzwerkspeicher massive Kostenfaktoren
- Rechtliche Anforderung: zur Übermittlung entgegengenommene Nachrichten dürfen nicht gelöscht werden

Fallbeispiel Email Service Provider

- Einzelne Email-Kampagnen erzeugen gigantische Datenmengen, verteilt über Botnetz-Knoten

Fallbeispiel Email Service Provider

- Administratoren bemerken große Kampagnen, schreiben regulären Ausdruck der Emails matcht
- Email-Server lehnt matchende Emails in der SMTP-Session ab, nimmt Nachrichten nicht entgegen.
- Problem: Kampagnen müssen rechtzeitig bemerkt werden, Handlung der Admins notwendig (Urlaub? Wochenende?)
- Falls Nachrichten nicht ankommen, steigt die Zahl der Beschwerdeanrufe im Call Center

Fallbeispiel Email Service Provider

- Anforderungen an automatisierte Lösung?
- Evaluierungsmetrik?
- Modellierung als Lernproblem?
  - Art des Lernproblems?
  - Modellraum?
  - Verlustfunktion? Regularisierer?

Taxonomie von Lernproblemen

- Überwacht: Daten enthalten Werte für vorherzusagende Variable
  - Klassifikation: Kategorielle Variable
  - Regression: Kontinuierliche Variable
  - Ordinale Regression: Endliche, geordnete Menge
  - Rankings: Reihenfolge von Elementen
  - Strukturierte Vorhersage: Sequenz, Baum, Graph, …
  - Empfehlungen: Nutzer-Produkt-Matrix

Taxonomie von Lernproblemen

- Unüberwacht: Strukturelle Eigenschaften der Trainingsdaten aufdecken
  - Clusteranalyse
  - Entdecken neuer Attribute
  - Anomaliedetektion
- Steuerungen / Reinforcement-Lernen: Steuerung eines dynamischen Systems
- Viele weitere Modelle
  - Überwachtes Clustern
  - Halbüberwachtes Lernen
  - …

Verfügbarkeit der Daten

- Batch-Lernen: Alle Trainingsdaten verfügbar
- Online-Lernen: Daten fallen einzeln, sequenziell an; Modell ändert sich inkrementell

Eigenschaften der Daten

- Umfang der Daten
  - Sehr wenige?
  - So viele, dass sie nur verteilt gespeichert und verarbeitet werden können?
- Anzahl der Attribute
  - Viele? Wenige?
  - Sparse (Viele Attribute, aber bei jeder Instanz sind nur wenige von null verschieden)?
- Qualität
  - Fehlende Werte?
  - Falsche Werte durch Messfehler?

Eigenschaften der Daten

- Verteilungseigenschaften: repräsentativ?
  - Klassenverhältnis ausgeglichen? Eine Klasse extrem selten?
  - Klassenverhältnis in Trainingsdaten so wie bei der Anwendung? Klassen unterrepräsentiert?
  - Verteilung $p(\boldsymbol{x})$ der Instanzen so wie bei der Anwendung? („Lernen unter Covariate Shift)
  - Werte des Zielattributs aus „echter“ Verteilung oder aus einer Hilfsverteilung (Laborexperimente, Simulationsdaten)
  - Daten aktuell? Verändert sich der Prozess zeitlich?

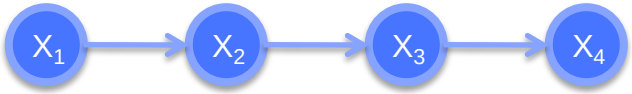
Eigenschaften der Daten

- Eine oder mehrere Datenquellen?
- Glaubwürdigkeit? Qualität? Konsistenz?
- Verfügbarkeit
  - Vorgegebener Datensatz?
  - Muss Protokoll zum Sammeln der Daten entwickelt werden?

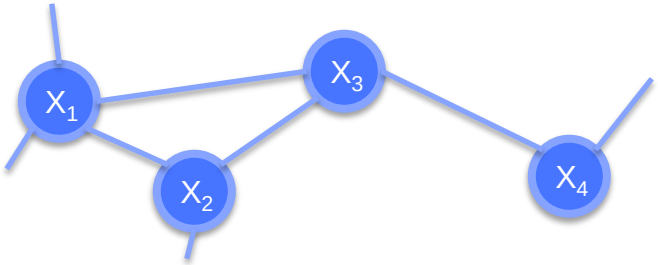
Eigenschaften der Daten

- Abhängigkeiten der Daten, z.B.
  - Unabhängige Einzelbeobachtungen

  - Sequenzen

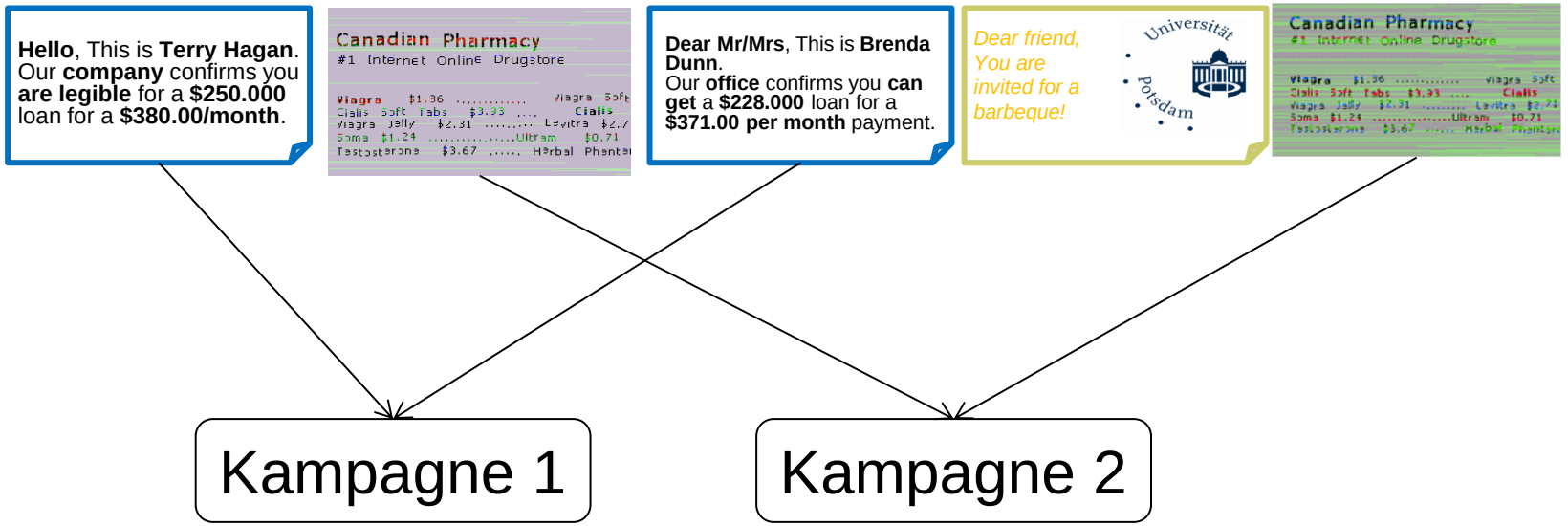


  - Vernetzte Daten



Fallbeispiel Email Service Provider

- Modellierung als zwei Lernprobleme
  1. Entdecken von Kampagnen



Fallbeispiel Email Service Provider

- Modellierung als zwei Lernprobleme
  1. Entdecken von Kampagnen
  2. Erzeugen eines regulären Ausdrucks für jede Kampagne

This is a [a-z]+to make a right move and receive your due
benefits… if you are qualified but are lacking that piece of paper. Get one from us in a
\S+(\S+){0,2} time.

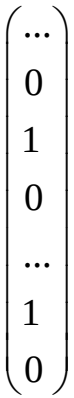
\S+(\S+){1,19} to start improving your life!

~C[A-Z]*FOR A FREE CONSULTATION~

(1-407-245-7320 You must |1-407-245-7320 You should |1-407-245-7320 Please)leave us a (|voice)message with your (name and |)phone number with country code if outside USA and
\S+(\S+){0,4}ll (|contact |get back to)you \S+(\S+){0,4}.

Fallbeispiel: Entdecken von Kampagnen

- Unüberwachtes Lernen: Clusteranalyse
- Online-Verarbeitung des Datenstromes
- Optimierungskriterium
  - Wahrscheinlichste Aufteilung in Cluster
- Instanzen: Header- und Wort-Attribute



Email-Header-Attribute

Alternative

Beneficiary

Friend

…

Sterling

Zoo

Dear Beneficiary,

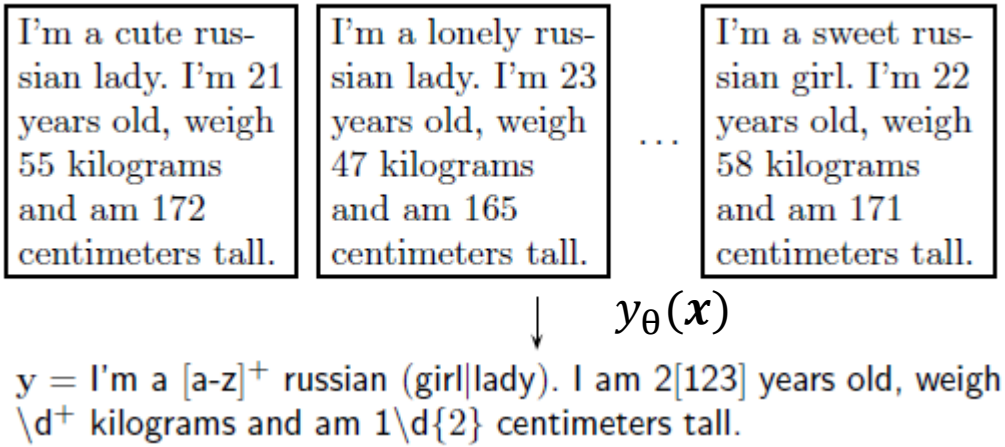
your Email address has been picked online in this years MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…

Fallbeispiel: Entdecken von Kampagnen

- Offline-Evaluierung:
  - Speichere Alle Emails in begrenztem Zeitraum
  - Teile von Hand in Cluster ein
  - Metrik: Übereinstimmung gefundener und von Hand festgelegter Cluster
  - False-Positive-Rate, Fals-Negative-Rate
- Online-Evaluierung, Tests
  - Finde Cluster im operativen Betrieb
  - Lege Administratoren zum Blockieren vor
  - Unvollständige Kampagne? Keine einheitliche Kampagne?

Fallbeispiel: Finde regulären Ausdruck

- Instanzen $\boldsymbol{x}$: Emails (Menge von Strings)
- Zielattribut $\boldsymbol{y}$: Regulärer Ausdruck



Fallbeispiel: Finde regulären Ausdruck

- Trainingsdaten $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$: Menge von Strings und zugehöriger regulärer Ausdruck eines Postmasters

This is a nice way to make a right move and receive your due
This is a best chance to make a right move and receive your
This is a good chance to make a right move and receive your
This is a good chance to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a fraction of the time. If you want to get better - you must Contact us 24 hours a day and 7 days a week! to start improving your life! ~CALL US FOR A FREE CONSULTATION~ 1-407-245-7320 You should leave us a message with your phone number with country code if outside USA and name and we'll contact you asap.

This is a [a-z]+to make a right move and receive your due benefits... if you are qualified but are lacking that piece of paper. Get one from us in a
\S+(\S+){0,2} time.

\S+(\S+){1,19} to start improving your life!

~C[A-Z]*FOR A FREE CONSULTATION~

Fallbeispiel: Finde regulären Ausdruck

- Art des Lernproblems:

Fallbeispiel: Finde regulären Ausdruck

- Art des Lernproblems:
  - Trainingsdaten enthalten gewünschte reguläre Ausdrücke: $\Rightarrow$ überwachtes Lernen
  - Zielvariable enthält reguläre Ausdrücke; diskret, strukturiert $\Rightarrow$ Strukturvorhersage (structured Output)

Fallbeispiel: Finde regulären Ausdruck

- Verlustfunktion $\ell(y_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i)$: Unterschiedlichkeit der beiden regulären Ausdrücke

Fallbeispiel: Finde regulären Ausdruck

- Verlustfunktion $\ell(y_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i)$: Unterschiedlichkeit der beiden regulären Ausdrücke
  - Anteil unterschiedlicher Knoten der Syntaxbäume
- Regularisierung: L_2

Fallbeispiel: Gesamtevaluierung

Fallbeispiel: Gesamtevaluierung

- Online-Evaluierung
  - Kampagnen und reguläre Ausdrücke werden Postmastern im operativen Betrieb vorgeschlagen
  - Akzeptanz-, Änderungsrate, Rate verworfener Kampagnen
  - Beschwerderate im Call Center

Überblick

- Analyse von Lernproblemen
  - Verständnis der Anforderungen
  - Entwicklung eines Lösung
  - Entwicklung einer Evaluierungsstrategie
- Datenvorverarbeitung
  - Datenintegration
  - Merkmalsrepräsentation
  - Fehlende Werte
  - Merkmalsselektion

Datenintegration

- Mehrere Datenquellen konsistent zusammenführen, z.B. in Data Warehouse
- Integration unterschiedlicher Datenformate
- Schema-Integration: gleiche/ähnliche Attribute in unterschiedlichen Quellen
- Daten-Konflikte (z.B. Umrechnung von Einheiten).
- Redundante Informationen erkennen (z.B. Dubletten).

Merkmalsrepräsentation

- Attribute werden transformiert, an Struktur des Modells angepasst
- Z.B.: lineare Modelle berechnen Vektorprodukt aus Attribut- und Parametervektor:
  - Alle Attribute müssen numerisch sein
  - Kategorielle Attribute („rot“, „grün“), Attribute ohne Ordnung und Texte müssen umgewandelt werden

Farbe	Größe	Produktkategorie
rot	23	173
blau	23	173
grün	45	36

→

Farbe rot	Farbe blau	Farbe grün	Größe	Produktkategorie 1	…	Produktkategorie 173
1	0	0	23	0		1
0	1	0	23	0		1
0	0	1	45	0		0

Merkmalsrepräsentation

- Texte: TF- oder TFIDF-Repräsentation
- Termfrequenz-Vektor: Eine Dimension pro Wort in Wörterbuch
- Wert: Anzahl Vorkommen des Wortes im Text

$$\mathbf{x} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 1 \\ 0 \end{pmatrix}$$

Alternative
Beneficiary
Friend
…
Sterling
Zoo

Email

Dear Beneficiary,

your Email address has been picked online in this years MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…

Merkmalsrepräsentation

- Texte: TF- oder TFIDF-Repräsentation
- Häufig vorkommende Worte („und“, „oder“, „ist“) häufig für Bedeutung des Textes nicht relevant
- Idee: Häufige Wörter heruntergewichten
- Inverse Dokumentenfrequenz

$$IDF(wort_i) = \log \frac{\#\text{Dokumente}}{\#\text{Dokumente , in denen Wort}_i \text{ vorkommt}}$$

Merkmalsrepräsentation

- TFIDF-Vektoren

$$TFIDF(\mathrm{x}) = \frac{1}{|\mathrm{x}|}\begin{pmatrix} TF(Wort_1)\cdot IDF(Wort_1) \\ \vdots \\ TF(Wort_n)\cdot IDF(Wort_n) \end{pmatrix}$$

Merkmalsrepräsentation

- Texte: N-Gramm-Vektoren
- In TFIDF-Repräsentation geht die Reihenfolge der Wörter verloren
- N-Gramm-Merkmale: Ein Attribut pro k-Tupel aufeinanderfolgender Wörter (für alle $k \leq N$)

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \\ \dots \\ 1 \end{pmatrix}$$

Dear
Dear Beneficiary
Dear Beneficiary your

has been picked

Thousand pounds Sterling

Email

Dear Beneficiary,

your Email address has been picked online in this years MICROSOFT CONSUMER AWARD as a Winner of One Hundred and Fifty Five Thousand Pounds Sterling…

Merkmalsrepräsentation

- Attribute können sehr unterschiedliche Wertebereiche haben
- Es kann sinnvoll sein, die Wertebereiche zu normalisieren

Merkmalsrepräsentation

- Feature-Normalisierung:
  - Min/Max-Normalisierung: $x^{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} (x^{new}_{\max} - x^{new}_{\min}) + x^{new}_{\min}$
  - Z-Score-Normalisierung: $x^{new} = \frac{x - \mu_x}{\sigma_x}$
  - Dezimal-Skalierung: $x^{new} = |x| \cdot 10^a \quad a = \max_x \{ i \in \mathbb{Z} \mid |x| \cdot 10^i < 1 \}$
  - Logarithmische Skalierung: $x^{new} = \log_a x$

Merkmalsrepräsentation

- Es kann sinnvoll sein, komplexe Merkmale zu konstruieren, wenn solche Merkmalskombinationen nicht im Modellraum liegen
- Feature-Konstruktion:
  - Kombination elementarer Feature, z.B. $(x_i, x_j) \rightarrow \left(x_j, \sqrt{x_i x_j}, x_i + x_j \right)$
  - Mapping elementarer Feature, z.B. $x_i \rightarrow \left(x_i, \log x_i, x_i^2 \right)$

Attribute mit fehlenden Werten

- Ursache für fehlende Daten:
  - Zufälliges Fehlen (z.B. Speicherfehler, Fehlfunktion eines Messinstruments).
  - Systematisches Fehlen (z.B. Wert zu einem früheren Zeitpunkt unwichtig/unbekannt).
  - Daten-Integration (z.B. gelöschte Werte aufgrund von Inkonsistenzen).
  - Daten-Aggregation (z.B. aus Datenschutzgründen).
  - …

Attribute mit fehlenden Werten

- Betreffende Instanzen/Attribute löschen.
  - Verringert Anzahl verfügbarer Daten, nur sinnvoll wenn nur wenige Werte fehlen
- Erweiterung des Wertebereichs (z.B. „missing“) und/oder der Attributmenge (z.B. binäres Attribut „Attribut_XY_bekannt“).
- Fehlende Werte aus Daten schätzen:
  - Mean/Median Imputation (evtl. Klassenabhängig).
  - Inferenz aus den Daten (z.B. mittels EM-Algorithmus).
- Fehlende Werte nicht behandeln (späteres Lern-/Analyseverfahren berücksichtigt fehlende Werte).

Attribute mit fehlerhaften Werten

- Identifizierung fehlerhafter Werte:
  - Binning: äquidistante Diskretisierung in *Bins*
    $\Rightarrow$ Bins mit einem/wenigen Instanzen enthalten evtl. Ausreißer.
  - Clustering: Suche nach Regionen mit hoher Datendichte (*Cluster*)
    $\Rightarrow$ Cluster mit einem/wenigen Instanzen enthalten evtl. Ausreißer.
  - Active Learning/Labeling: Widerspruch zwischen Daten und Modell
    $\Rightarrow$ auffällige Instanzen werden Menschen zur Kontrolle vorgeschlagen.
- Behandlung fehlerhafter Werte:
  - Glättung numerischer Werte (z.B. Regression, Moving Average).
  - Diskretisierung (z.B. Alter $\Rightarrow$ {kücken, bivi, uhu}).
  - Als fehlend behandeln.

Merkmalsselektion

- Auswahl einer Teilmenge der Attribute kann zu besseren Ergebnissen führen
- Dimensionsreduktion
- Viele Merkmalsselektionsverfahren
  - Z.B. Hauptkomponentenanalyse
  - Forward- / Backward-Selection
- Evaluierung mit Training-/Test oder Cross-Validation:
  - Trainiere Modell mit unterschiedlichen Teilmengen auf Trainingsmenge
  - Evaluiere auf Testmenge

Merkmalsselektion

- Merkmalsselektion für lineare Modelle, z.B.
  - Trainiere lineares Modell
  - Lass Attribute mit kleinsten Gewichten weg
  - Trainiere neues Modell, evaluiere auf Testmenge
  - …

Problemanalyse, Datenvorverarbeitung

- Maschinelles Lernen (wie Informatik generell) ist Ingenieurwissenschaft
  - Problemanalyse, Erfassung von Anforderungen
  - Abbildung auf bekannte Paradigmen, Rückgriff auf Stand der Technik
  - Ableitung von Lösung und Evaluierungsstrategie
- Datenvorverarbeitung häufig Unterschied zwischen guter Lösung und Lösung die nicht funktioniert
  - Datenbintegration
  - Konstruktion guter Attribute
  - Fehlende, fehlerhafte Werte
  - Merkmalsselektion