# Data Mining in Datenbanken

Alexander Hinneburg
hinneburg@informatik.uni-halle.de
www.informatik.uni-halle.de/~hinneburg

# Vorlesungsplan

- 17.10. Einleitung
- 24.10. Ein- und Ausgabe
- 31.10. Reformationstag
- 7.11. Einfache Regeln, Naïve Bayes
- 14.11. Entscheidungsbäume, und Regeln
- 21.11. Assoziationsregeln
- 28.11. Lineare Modelle
- 5.12. Instanzbasiertes Lernen, Clustering
- 12.12. Evaluation I
- 19.12. Evaluation II
- 9.1. Entscheidungsbäume, Klassifikationsregeln
- 16.1. Lineare Modelle, Numerische Vorhersage
- 23.1. Clustering
- 30.1. Attribut-Selektion, Diskretisierung, Transformationen
- 6.2. Kombination von Modellen, Lernen von nicht-klassifizierten Beispielen

# Veranstaltungsmodalitäten

- Vorlesung
  - Mischung aus Folien und Tafel
  - Ian Witten, Eibe Frank: Data Mining, 2. Aufl., Morgan Kaufmann, 2005
- Übung
  - Theorieaufgaben
  - Praktische Aufgaben,  EKA Data Mining Suite (http://www.cs.waikato.ac.nz/~ml/weka/), Oracle
  - Präsentation der Ergebnisse, 3-5 Folien
- Schein
  - 50% der Übungspunkte
  - Klausur Ende des Semesters
  - Wirtschaftsinformatiker: 2+2 Vorlesung, 6 Leistungspunkte
- Blockseminar
  - 2-3 Tage, Anfang April 2006 (vor Vorlesungsbeginn)

# Einleitung

- Daten versus Wissen
- Data Mining und Machine Learning
- Strukturierte Beschreibungen
  - Regeln: Klassifikation und Assoziation
  - Entscheidungsbäume
- Daten
  - Wetter, Kontaktlinsen, CPU Performanz, Gewerkschaftsverhandlungen, Sojabohnen Klassifikation
- Echte Anwendungen
  - Kreditanwendungen, Satellitenaufnahmen, Lastvorhersage, Maschinenfehlerdiagnose, Warenkorbanalyse
- Verallgemeinerung als Suche
- Data Mining und Ethik

# Daten versus Wissen

- Gesellschaft erzeugt riesige Datenmengen
  - Quellen: Wirtschaft, Wissenschaft, Medizin, Geo-Anwendungen, Sport, Umwelt, …
- Potentiell wertvolle Ressource
- Rohdaten sind nutzlos: Methoden um Informationen automatisch daraus zu gewinnen
  - Daten: gespeicherte Fakten
  - Informationen: in den Daten inhärente Muster

# Data Mining Aufgabe

- Extrahiere
  - implizite,
  - vorher unbekannte,
  - potentiell nützliche

  Informationen aus Daten
- Bedarf: Programme, die Muster und Regelmäßigkeiten in Daten detektieren
- Interessante Muster $\Rightarrow$ gute Vorhersagen
  - Problem 1: meisten Muster nicht interessant
  - Problem 2: Muster sind unscharf, unecht
  - Problem 3: Daten fehlen, sind fehlerbehaftet

# Data Mining Methoden

- *Algorithmen zum Gewinnen (Lernen) struktureller Beschreibungen aus Beispielen*
- strukturelle Beschreibungen repräsentieren Muster explizit
  - Vorhersage in neuen Situationen möglich
  - Zum Verstehen und Erklären wie Vorhersage entstand (*meist wichtiger*)
- Methoden stammen aus Künstlicher Intelligenz, Statistik, Datenbanken, Visualisierung, Theoretische Informatik

# Kontaktlinsen Daten

| Age | Spectacle prescription | Astigmatism | Tear production rate | Recommended lenses |
|---|---|---|---|---|
| Young | Myope | No | Reduced | None |
| Young | Myope | No | Normal | Soft |
| Young | Myope | Yes | Reduced | None |
| Young | Myope | Yes | Normal | Hard |
| Young | Hypermetrope | No | Reduced | None |
| Young | Hypermetrope | No | Normal | Soft |
| Young | Hypermetrope | Yes | Reduced | None |
| Young | Hypermetrope | Yes | Normal | hard |
| Pre-presbyopic | Myope | No | Reduced | None |
| Pre-presbyopic | Myope | No | Normal | Soft |
| Pre-presbyopic | Myope | Yes | Reduced | None |
| Pre-presbyopic | Myope | Yes | Normal | Hard |
| Pre-presbyopic | Hypermetrope | No | Reduced | None |
| Pre-presbyopic | Hypermetrope | No | Normal | Soft |
| Pre-presbyopic | Hypermetrope | Yes | Reduced | None |
| Pre-presbyopic | Hypermetrope | Yes | Normal | None |
| Presbyopic | Myope | No | Reduced | None |
| Presbyopic | Myope | No | Normal | None |
| Presbyopic | Myope | Yes | Reduced | None |
| Presbyopic | Myope | Yes | Normal | Hard |
| Presbyopic | Hypermetrope | No | Reduced | None |
| Presbyopic | Hypermetrope | No | Normal | Soft |
| Presbyopic | Hypermetrope | Yes | Reduced | None |
| Presbyopic | Hypermetrope | Yes | Normal | None |

# Können Maschinen wirklich lernen?

- Definitionen für “lernen” aus dem Lexikon:

  **Wissenserwerb durch Studium, Erfahrung oder Belehrung**

  **Kenntnis erlangen durch sich informieren oder durch Beobachtung**

  **Auswendig lernen**

  **Informiert sein, etwas ermitteln, Instruktionen erhalten,**

- Operationale Definition:

  **Etwas lernt, wenn es sein Verhalten oder Eigenschaften so ändert, dass es in Zukunft besser funktioniert.**

- Impliziert Lernen eine Intention?

# Das Wetter Problem

- Umstände um Tennis zu spielen

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | Normal | False | Yes |
| Rainy | Cool | Normal | True | No |
| … | … | … | … | … |

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

# Wetterdaten mit gemischten Attributen

- Einige Attribute haben nummerische Werte

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | 85 | 85 | False | No |
| Sunny | 80 | 90 | True | No |
| Overcast | 83 | 86 | False | Yes |
| Rainy | 75 | 80 | False | Yes |
| … | … | … | … | … |

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

# Ross Quinlan

- Machine learning Forscher seit 1970
- University of Sydney, Australia

1986 “Induction of decision trees” *ML Journal*

1993 *C4.5: Programs for machine learning*. Morgan Kaufmann

199? Started

RULEQUEST RESEARCH
*data mining tools*

# Klassifikations & Assoziationsregeln

- Klassifikationsregel:
  Vorhersage für den Wert eines gegebenen Attributes

```
If outlook = sunny and humidity = high
   then play = no
```

- Assoziationsregel:
  Vorhersage des Wertes eines beliebigen Attributes o. Kombination

```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```

# Wetterdaten (komplett)

| | | | | |
|---|---|---|---|---|
| • sunny | hot | high | FALSE | no |
| • sunny | hot | high | TRUE | no |
| • overcast | hot | high | FALSE | yes |
| • rainy | mild | high | FALSE | yes |
| • rainy | cool | normal | FALSE | yes |
| • rainy | cool | normal | TRUE | no |
| • overcast | cool | normal | TRUE | yes |
| • sunny | mild | high | FALSE | no |
| • sunny | cool | normal | FALSE | yes |
| • rainy | mild | normal | FALSE | yes |
| • sunny | mild | normal | TRUE | yes |
| • overcast | mild | high | TRUE | yes |
| • overcast | hot | normal | FALSE | yes |
| • rainy | mild | high | TRUE | no |

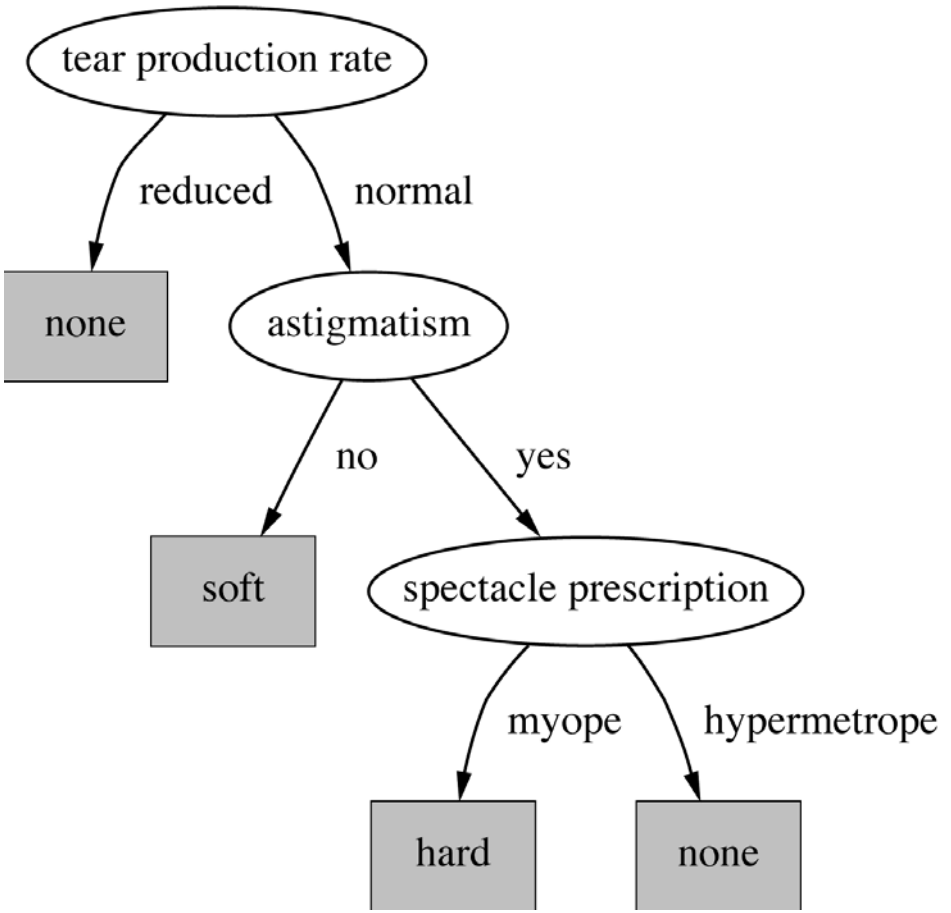
# Kontaktlinsen Daten

| Age | Spectacle prescription | Astigmatism | Tear production rate | Recommended lenses |
|---|---|---|---|---|
| Young | Myope | No | Reduced | None |
| Young | Myope | No | Normal | Soft |
| Young | Myope | Yes | Reduced | None |
| Young | Myope | Yes | Normal | Hard |
| Young | Hypermetrope | No | Reduced | None |
| Young | Hypermetrope | No | Normal | Soft |
| Young | Hypermetrope | Yes | Reduced | None |
| Young | Hypermetrope | Yes | Normal | hard |
| Pre-presbyopic | Myope | No | Reduced | None |
| Pre-presbyopic | Myope | No | Normal | Soft |
| Pre-presbyopic | Myope | Yes | Reduced | None |
| Pre-presbyopic | Myope | Yes | Normal | Hard |
| Pre-presbyopic | Hypermetrope | No | Reduced | None |
| Pre-presbyopic | Hypermetrope | No | Normal | Soft |
| Pre-presbyopic | Hypermetrope | Yes | Reduced | None |
| Pre-presbyopic | Hypermetrope | Yes | Normal | None |
| Presbyopic | Myope | No | Reduced | None |
| Presbyopic | Myope | No | Normal | None |
| Presbyopic | Myope | Yes | Reduced | None |
| Presbyopic | Myope | Yes | Normal | Hard |
| Presbyopic | Hypermetrope | No | Reduced | None |
| Presbyopic | Hypermetrope | No | Normal | Soft |
| Presbyopic | Hypermetrope | Yes | Reduced | None |
| Presbyopic | Hypermetrope | Yes | Normal | None |

# Komplete und korrekte Regelmenge

```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope
    and astigmatic = no  then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no
    and tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age young and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age = pre-presbyopic
    and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
```

# Entscheidungsbaum für dieses Problem



# Klassifizierung von Iris-Blumen

| | Sepal length | Sepal width | Petal length | Petal width | Type |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | Iris setosa |
| … | | | | | |
| 51 | 7.0 | 3.2 | 4.7 | 1.4 | Iris versicolor |
| 52 | 6.4 | 3.2 | 4.5 | 1.5 | Iris versicolor |
| … | | | | | |
| 101 | 6.3 | 3.3 | 6.0 | 2.5 | Iris virginica |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 | Iris virginica |
| … | | | | | |

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

# Vorhersage CPU Performanz

- Beispiel: 209 verschiedene Computer Konfig.

| | Cycle time (ns) | Main memory (Kb) | | Cache (Kb) | Channels | | Performance |
|---|---|---|---|---|---|---|---|
| | MYCT | MMIN | MMAX | CACH | CHMIN | CHMAX | PRP |
| 1 | 125 | 256 | 6000 | 256 | 16 | 128 | 198 |
| 2 | 29 | 8000 | 32000 | 32 | 8 | 32 | 269 |
| … | | | | | | | |
| 208 | 480 | 512 | 8000 | 32 | 0 | 0 | 67 |
| 209 | 480 | 1000 | 4000 | 0 | 0 | 0 | 45 |

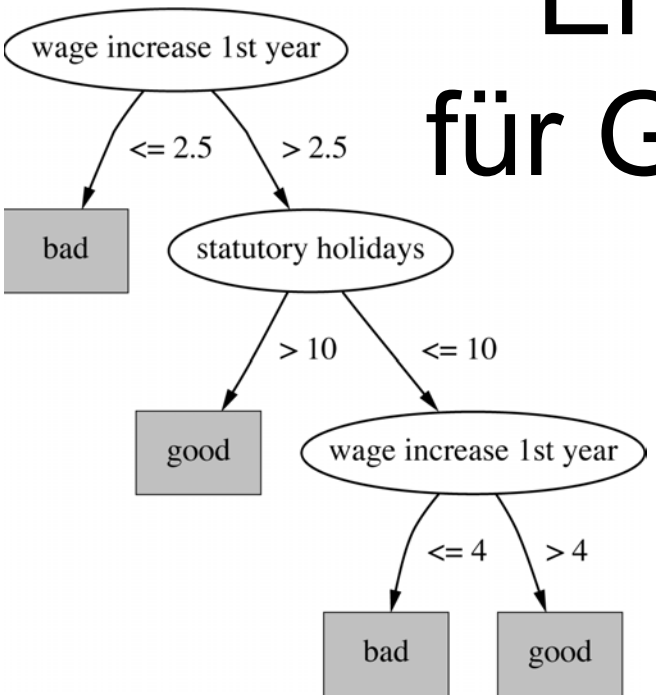
- Lineare Regression Funktion

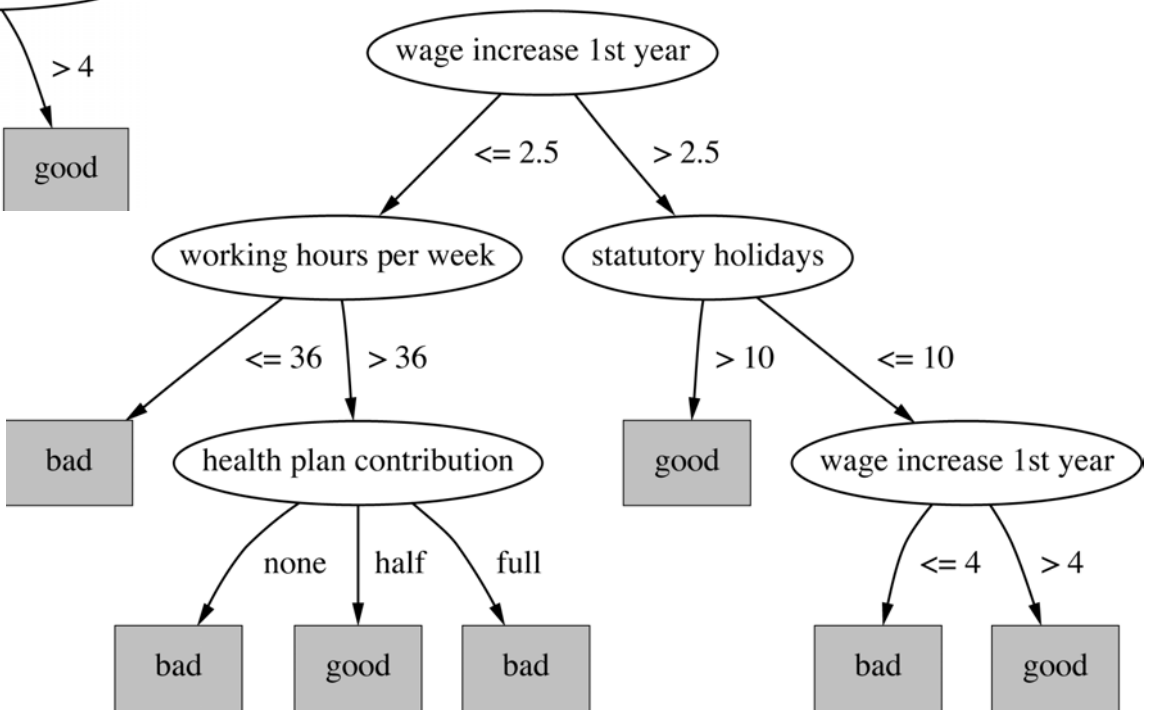
```
PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
      + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

# Gewerkschaftsverhandlungen

| Attribute | Type | 1 | 2 | 3 | … | 40 |
|---|---|---|---|---|---|---|
| Duration | (Number of years) | 1 | 2 | 3 | | 2 |
| Wage increase first year | Percentage | 2% | 4% | 4.3% | | 4.5 |
| Wage increase second year | Percentage | ? | 5% | 4.4% | | 4.0 |
| Wage increase third year | Percentage | ? | ? | ? | | ? |
| Cost of living adjustment | {none,tcf,tc} | none | tcf | ? | | none |
| Working hours per week | (Number of hours) | 28 | 35 | 38 | | 40 |
| Pension | {none,ret-allw, empl-cntr} | none | ? | ? | | ? |
| Standby pay | Percentage | ? | 13% | ? | | ? |
| Shift-work supplement | Percentage | ? | 5% | 4% | | 4 |
| Education allowance | {yes,no} | yes | ? | ? | | ? |
| Statutory holidays | (Number of days) | 11 | 15 | 12 | | 12 |
| Vacation | {below-avg,avg,gen} | avg | gen | gen | | avg |
| Long-term disability assistance | {yes,no} | no | ? | ? | | yes |
| Dental plan contribution | {none,half,full} | none | ? | full | | full |
| Bereavement assistance | {yes,no} | no | ? | ? | | yes |
| Health plan contribution | {none,half,full} | none | ? | full | | half |
| Acceptability of contract | {good,bad} | bad | good | good | | good |

# Entscheidungsbäume für Gewerkschaftsdaten



# Sojabohnen Klassifikation

| | Attribute | Number of values | Sample value |
|---|---|---|---|
| *Environment* | Time of occurrence | 7 | July |
| | Precipitation | 3 | Above normal |
| … | | | |
| *Seed* | Condition | 2 | Normal |
| | Mold growth | 2 | Absent |
| … | | | |
| *Fruit* | Condition of fruit pods | 4 | Normal |
| | Fruit spots | 5 | ? |
| *Leaves* | Condition | 2 | Abnormal |
| | Leaf spot size | 3 | ? |
| … | | | |
| *Stem* | Condition | 2 | Abnormal |
| | Stem lodging | 2 | Yes |
| … | | | |
| *Roots* | Condition | 3 | Normal |
| *Diagnosis* | | 19 | Diaporthe stem canker |

# Bedeutung von Hintergrundwissen

```
If leaf condition is normal
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

```
If leaf malformation is absent
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

Aber hier, “leaf condition is normal” impliziert “leaf malformation is absent”!

# Echte Anwendungen

- Das Ergebnis des Data Mining
  —oder die DM Methode selbst—
  wird in praktischen Anwendungen eingesetzt
  - Bearbeiten von Kreditanträgen
  - Durchsuchen von Satellitenbilden nach Ölteppichen
  - Stromverbrauchvorhersage
  - Diagnose von Maschinenfehlern
  - Marketing und Verkauf
  - Automatische Klassifikation von Himmelsobjekten
  - Automatische Ergänzung von Formularen
  - Text Retrieval und Mining

# Kreditanträge (American Express)

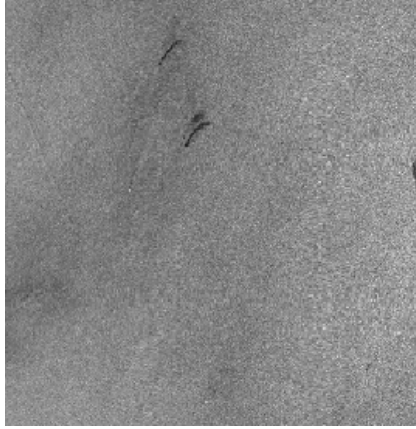
- Gegeben: Fragebogen mit finanziellen und persönliche Informationen
- Frage: soll der Antrag bewilligt werden?
- Einfache statist. Methode für 90% der Fälle
- Grenzfälle an Sachbearbeiter weitergereicht
- Aber: 50% der akzeptierten Grenzfälle konnten nicht zurückzahlen!
- Lösung: alle Grenzfälle zurückweisen?
  - Nein! Grenzfälle sind die aktivsten Kunden

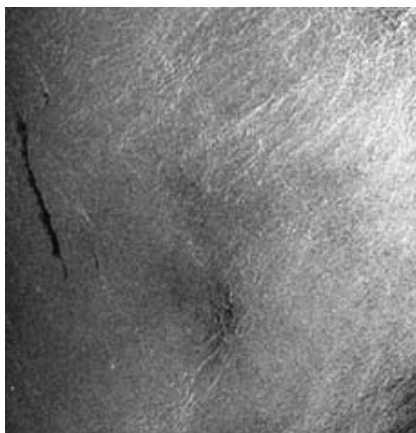
# Data Mining Einstieg

- 1000 Trainingsbeispiele von Grenzfällen
- 20 Attribute:
  - Alter
  - Jahre beim derzeitigen Arbeitgeber
  - Jahre bei derzeitiger Adresse
  - Jahre bei der Bank
  - andere Kredit Karten besessen ,…
- Gelernte Regeln: korrekt in 70% der Fälle
  - Sachbearbeiter nur 50%
- Regeln konnten genutzt werden, um die Entscheidung dem Kunden zu erklären

# Satellitenbilder

- Gegeben: Radar Satellitenbilder von Küstengewässern
- Problem: finde Ölteppiche in den Bildern
- Ölteppiche tauchen als dunkle Region mit sich ändernder Größe und Form auf
- Schwierigkeiten: ähnliche dunkle Regionen  können durch Wetterphenomene verursacht werden (z.B. Höhenwinde)
- Teure Bearbeitung erfordert spezialisiertes Personal





# Data Mining Einstieg

- Extrahiere dunkle Regionen aus normalisierten Bildern
- Attribute:
  - Größe der Region
  - Form und Fläche
  - Intensität
  - Schärfe und Zerklüftung der Grenzen
  - Nähe zu anderen Regionen
  - Infos über Hintergrund
- Bedingungen:
  - Wenig Trainingsbeispiele—Ölteppiche sind selten!
  - Ungleich verteilte Daten: meisten dunklen Regionen sind keine Ölteppiche
  - Anforderung: einstellbare Rate der Fehlalarme

# Stromverbrauchs-vorhersage

- Stromanbieter müssen den Verbrauch vorhersagen und bestellen
- Vorhersage von Min/Max Last für jede Stunde
  $\Rightarrow$ signifikante Einsparungen
- Gegeben: manuell konstruiertes Lastmodell mit Annahmen über “normale” klimatische Bedingungen
- Problem: anpassen der Wetterbedingungen
- Statisches Modell besteht aus:
  - Basislast des Jahres
  - periodische Lastveränderungen übers Jahr
  - Effekt der Ferien

# Data Mining Einstieg

- Vorhersage korrigiert mittels der “ähnlichsten” Tage
- Attribute:
  - Temperatur
  - Luftfeutigkeit
  - Windgeschwindigkeit
  - Wolkenbedeckung
  - plus Unterschied zwischen aktueller Last und vorhergesagter Last
- Durchschnittsunterschied der drei “ähnlichsten” Tage zum Modell hinzugefügt
- Lineare Regressionskoeffizienten bilden Attributgewichte in der Ähnlichkeitsfunktion

# Diagnose von Maschinenfehlern


- Diagnose: klassische Domäne von Expertensystemen
- Gegeben: Fourier Analyse von Vibrationen gemessen an verschiedenen Haltepunkten
- Frage: welcher Fehler ist gegeben?
- Preventive Wartung von elektromechanischen Motoren und Generatoren
- Information ist sehr verrauscht
- Bisher: Diagnose durch Expertenregeln

# Data Mining Einstieg

- Verfügbar: 600 Fehler mit Expertendiagnose
- ~300 nichtbefriedigend, Rest genutzt zum Trainieren
- Attribute erweitert durch Zwischenkonzepte, die kausales Hintergrundwissen enthalten
- Experte nicht zufrieden mit initialen Regeln, weil sie sich nicht in Beziehung zum Hintergrundwissen setzen ließen
- Weiteres Hintergrundwissen ergab komplexere Regeln, die besser waren
- Gelernte  Regeln waren besser als Expertenregeln

# Marketing und Verkauf I

- Firmen speichern detailierte Daten über Verkauf und Marketing
- Anwendungen:
  - Kunden Loyalität:
    finde Kunden, die wechseln wollen durch Analyse ihres bisherigen Verhaltens
    (e.g. Banken/Telekommunikationsfirmen)
  - Sonderangebote:
    finde profitabele Kunden
    (z.B. zuverlässige Kreditkartenkunden, die in den Ferien mehr Geld brauchen)

# Marketing und Verkauf II

- Warenkorbanalyse
  - Assoziationsmethoden finden Produktgruppen, die oft zusammen gekauft werden (verwendet um Kassendaten zu untersuchen)
- Zeitliche Analyse von Kaufverhalten
- Identifiziere interessierte Kunden
  - Fokusiere Einführungswerbesendungen (Gezielte Aktionen sind billiger als das Gießkannenprinzip)

# Data Mining, Maschinelles Lernen und Statistik

- Geschichtliche Unterschiede (stark vereinfacht)
  - Statistik: testen von Hypothesen, kleine Datenmengen
  - Machine learning: finde die richtigen Hypothesen, meist Klassifikation
  - Data Mining: Anwendungsgetrieben, große Datenbanken
- Aber: viele Gemeinsamkeiten
  - Entscheidungsbäume (C4.5 und CART)
  - Nächste Nachbar Methoden
  - Associationsregeln und formale Begriffsanalyse
- Heute: Perspektiven konvergieren
  - Die meisten ML Algorithmen verwenden statistische Techniken
  - Verschiedene „Communities“, verschiedene Begriffe, gemeinsame Probleme

# Statistiker

## Sir Ronald Aylmer Fisher

* 17 Feb 1890 London, England
+ 29 July 1962 Adelaide, Australia

*Numerous distinguished contributions to developing the theory and application of statistics for making quantitative a vast field of biology*

- Leo Breiman
- Entwickelte Entscheidungsbäume
- *1984 Classification and Regression Trees*. Wadsworth.

# Problemverallgemeinerung als Suche

- Induktives Lernen: finde eine Konzept-Beschreibung, die auf die Daten paßt
- Beispiel: Regeln als Beschreibungssprache
  - Riesiger, aber endlicher Suchraum
- Einfache Lösung:
  - durchsuche den ganzen Konzeptraum
  - verwerfe Beschreibungen, die nicht auf die Beispiele passen
  - übrigbleibende Beschreibungen enthalten das Zielkonzept

# Durchsuchen des Konzeptraumes

- Suchraum für das Wetterproblem
  - 4 x 4 x 3 x 3 x 2 = 288 mögliche Kombinationen
  - mit 14 Regeln $\Rightarrow$ $2.7 \times 10^{34}$ mögliche Regelmengen
- Lösung: Kandidaten-Eliminations Algorithmus
- Andere praktische Probleme:
  - Mehr als eine Beschreibung bleibt übrig
  - Keine Beschreibung bleibt übrig
    - Sprache kann Zielkonzept nicht ausdrücken
    - *oder D*aten enthalten Rauschen

# Der Versionsraum

- Raum der konsistenten Konzeptbeschreibungen
- Komplett durch zwei Mengen bestimmt
  - *L*: spezifischste Beschreibungen, die alle pos. und keine neg. Beispiele bestimmen
  - *G*: allgemeinsten Beschreibungen, die keine neg. und alle pos. Beispiele bestimmen
- Nur *L* und *G* müssen aktualisiert werden
- Nachteile
  - immer noch sehr berechnung intensiv
  - lößt nicht die anderen prakt. Probleme

# Versionsraum Beispiel

- Gegeben: rot oder grün, Kühe oder Hühner

*L*={} *G*={<*, *>}

<grün,Kuh>: positiv

*L*={<grün, Kuh>} *G*={<*, *>}

<rot,Huhn>: negativ

*L*={<grün, Kuh>} *G*={<grün,*>,<*,Kuh>}

<grün, Huhn>: positiv

*L*={<grün, *>} *G*={<grün, *>}

# Candidate-elimination algorithm

```
Initialize L and G
For each example e:
  If e is positive:
    Delete all elements from G that do not cover e
    For each element r in L that does not cover e:
      Replace r by all of its most specific generalizations
        that 1. cover e and
             2. are more specific than some element in G
    Remove elements from L that
      are more general than some other element in L
  If e is negative:
    Delete all elements from L that cover e
    For each element r in G that covers e:
      Replace r by all of its most general specializations
        that 1. do not cover e and
             2. are more general than some element in L
    Remove elements from G that
      are more specific than some other element in G
```

# Bias (Befangenheit)

- Wichtige Entscheidungen in Lernsystemen:
  - Konzeptbeschreibungssprache
  - Reihenfolge in der der Suchraum durchmustert wird
  - Methoden um overfitting (auswendiglernen) der Trainingsdaten zu vermeiden
- Diese Entsch. bestimmen den “Bias” der Suche:
  - Sprach Bias
  - Such Bias
  - Overfitting-Vermeidungs Bias

# Sprach Bias

- Wichtige Frage:
  - ist die Sprache universal
    oder beschränkt sie was gelernt werden kann?
- Universale Sprache kann beliebige Teilmengen der Beispiele beschreiben
- Falls Sprache logisches *oder* (“Disjunktion”) enthält, ist sie universal
- Beispiel: Regelmengen
- Hintergrundwissen kann genutzt werden, um Teile des Suchraums von vornherein auszuschließen

# Such Bias

- Suchheuristik
  - “Greedy” Suche: macht immer den aktuell bestmöglichen Schritt
  - “Beam search”: hält immer ein paar Alternativen vor
  - …
- Richtung der Suche
  - *Allgemein-hinzu-Spezifisch*
    - z.B. spezialisieren einer Rgel durch Hinzufügen von Bedingungen
  - *Spezifisch-hinzu-Allgemein*
    - z.B. Verallgemeinern eines Beispiels zu einer Regel

# Overfitting-Vermeidungs Bias

- Kann als Art Such Bias verstanden werden
- Verändertes Evaluationskriterium
  - z.B. balanciert Einfachheit und Anzahl der Fehler
- Veränderte Suchstrategie
  - z.B. pruning (Vereinfachen einer Beschreibung)
    - Pre-pruning: hält bei einer einfachen Beschreibung an, bevor die Suche mit einer komplexen B. fortgesetzt wird
    - Post-pruning: erzeugt zuerst komplexe Beschreibungen, die dann vereinfacht werden

# Data mining und Ethik I

- Ethische Fragestellungen tauchen in praktischen Anwendungen auf
- Data mining oft benutzt zum Unterscheiden
  - z.B. Kreditanträge: Verwendung von Informationen (z.B. Geschlecht, Religion, Herkunft) ist unethisch
- Ethische Situation hängt von der Anwendung ab
  - z.B. dieselbe Information ist ok in medizinischen Anwendungen
- Attribute können problematische Informationen enthalten
  - z.B. Postleitzahl kann mit Herkunft korrelieren

# Data mining und Ethik II

- Wichtige Fragen:
  - Wer hat Zugang zu den Daten?
  - Für welchen Zweck werden die Daten gesammelt?
  - Welche Schlußfolgerung dürfen daraus gezogen werden?
- Ergebnisse müssen Vorteile enthalten, für wen?
- Rein statistische Argumente sind niemals ausreichend!
- Werden die Ergebnisse sinnvoll und gut verwendet?