Machine Learning, Data Mining, and Knowledge Discovery: An Introduction

AHPCRC Workshop - 8/16/11 - Dr. Martin

Based on slides by Gregory Piatetsky-Shapiro from Kdnuggets

http://www.kdnuggets.com/data_mining_course/

Outline

- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- Data Mining with Weka

2

Machine Learning / Data Mining Application areas

- Science
  - astronomy, bioinformatics, drug discovery, …
- Business
  - CRM (Customer Relationship management), fraud detection, e-commerce, manufacturing, sports/entertainment, telecom, targeted marketing, health care, …
- Web:
  - search engines, advertising, web and text mining, recommender systems, spam filtering …
- Government
  - surveillance, crime detection, profiling tax cheaters, …

3

Business: Data Mining for Customer Modeling

- Customer Tasks:
  - attrition prediction
  - targeted marketing:
    - cross-sell, customer acquisition
  - credit-risk
  - fraud detection
- Industries
  - banking, telecom, retail sales, …

4

Customer Attrition: Case Study

- Situation: Attrition rate at for mobile phone customers is around 25-30% a year!
- With this in mind, what is our task?
  - Assume we have customer information for the past N months.

5

Customer Attrition: Case Study

Task:

- Predict who is likely to attrite next month.
- Estimate customer value and what is the cost-effective offer to be made to this customer.

6

Customer Attrition Results

- Verizon Wireless built a customer data warehouse
- Identified potential attriters
- Developed multiple, regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

(Reported in 2003)

7

e-commerce

- A person buys a book (product) at Amazon.com

What is the task?

8

Successful e-commerce – Case Study

- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful

9

Unsuccessful e-commerce case study (KDD-Cup 2000)

- Data: clickstream and purchase data from Gazelle.com, legwear and legcare e-tailer
- Q: Characterize visitors who spend more than $12 on an average order at the site
- Dataset of 3,465 purchases, 1,831 customers
- Very interesting analysis by Cup participants
  - thousands of hours - $X,000,000 (Millions) of consulting
- Total sales -- $Y,000
- Obituary: Gazelle.com out of business, Aug 2000
- Google "kdd cup 2000 gazelle"

10

Genomic Microarrays – Case Study

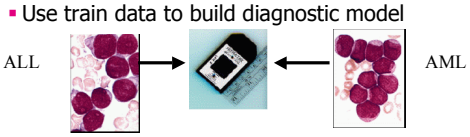
Given microarray data for a number of samples (patients), can we

- Accurately diagnose the disease?
- Predict outcome for given treatment?
- Recommend best treatment?

11

Example: ALL/AML data

- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model

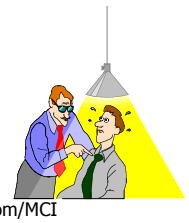


Results on test data:
33/34 correct, 1 error may be mislabeled

12

Security and Fraud Detection - Case Study

- Credit Card Fraud Detection
- Detection of Money laundering
  - FAIS (US Treasury)
- Securities Fraud
  - NASDAQ KDD system
- Phone fraud
  - AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terrorism detection at Salt Lake Olympics 2002



13

Data Mining and Privacy

- in 2006, NSA (National Security Agency) was reported to be mining years of call info, to identify terrorism networks
- Social network analysis has a potential to find networks
- Invasion of privacy – do you mind if your call information is in a gov database?
- What if NSA program finds one real suspect for 1,000 false leads ? 1,000,000 false leads?

14

Problems Suitable for Data-Mining

- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!

Privacy considerations important if personal data is involved

15

Outline

- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- Data Mining with Weka

16

Knowledge Discovery Definition

Knowledge Discovery in Data is the

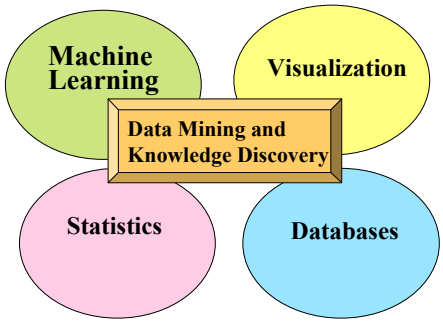
non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

17

Related Fields



18

Outline

- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- Data Mining with Weka

19

Finding patterns

- Goal: programs that detect patterns and regularities in the data
- Strong patterns ⇒ good predictions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

20

Machine learning techniques

- *Algorithms for acquiring structural descriptions from examples*
- Structural descriptions represent patterns explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived (*may be even more important*)
- Methods originate from artificial intelligence, statistics, and research on databases

witten&eibe 21

Can machines really learn?

- Definitions of "learning" from dictionary:
  - **To get knowledge of by study, experience, or being taught** } Difficult to measure
  - **To become aware by information or from observation** } Trivial for computers
  - **To commit to memory** } Trivial for computers
  - **To be informed of, ascertain; to receive instruction** } Trivial for computers
- Operational definition:
  - **Things learn when they change their behavior in a way that makes them perform better in the future.** } Does a slipper learn?
- Does learning imply intention?

witten&eibe 22

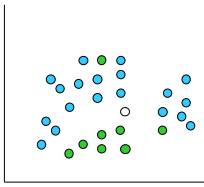
Major Data Mining Tasks

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection**: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationships
- …

23

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

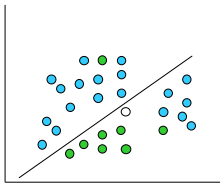


Many approaches:
Regression,
Decision Trees,
Bayesian,
Neural Networks,
…

Given a set of points from classes ● ●
what is the class of new point ○?

24

Classification: Linear Regression

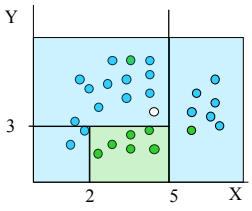


- Linear Regression

  $w_0 + w_1 x + w_2 y >= 0$
- Regression computes w_i from data to minimize squared error to 'fit' the data
- Not flexible enough

25

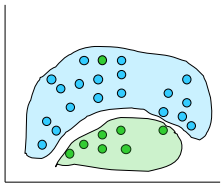
Classification: Decision Trees



if X > 5 then blue
else if Y > 3 then blue
else if X > 2 then green
else blue

26

Classification: Neural Nets



- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

27

Built in Data Sets

- Weka comes with some built in data sets
- Described in chapter 1
- We'll start with the Weather Problem
  - Toy (very small)
  - Data is entirely fictitious

28

But First…

- Components of the input:
  - Concepts: kinds of things that can be learned
    - Aim: intelligible and operational concept description
  - Instances: the individual, independent examples of a concept
    - Note: more complicated forms of input are possible
  - Attributes: measuring aspects of an instance
    - We will focus on nominal and numeric ones

29

What's in an attribute?

- Each instance is described by a fixed predefined set of features, its "attributes"
- But: number of attributes may vary in practice
  - Possible solution: "irrelevant value" flag
- Related problem: existence of an attribute may depend of value of another one
- Possible attribute types ("levels of measurement"):
  - *Nominal, ordinal, interval* and *ratio*

witten&eibe

30

What's a concept?

- Data Mining Tasks (Styles of learning):
  - Classification learning:
    predicting a discrete class
  - Association learning:
    detecting associations between features
  - Clustering:
    grouping similar instances into clusters
  - Numeric prediction:
    predicting a numeric quantity
- Concept: thing to be learned
- Concept description: output of learning scheme

The weather problem

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	mild	normal	false	yes
rainy	mild	normal	true	no
overcast	mild	normal	true	yes
sunny	mild	high	false	no
sunny	mild	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

Given past data, Can you come up with the rules for Play/Not Play ?

What is the game?

The weather problem

- Given this data, what are the rules for play/not play?

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
...	...	...	...	...

The weather problem

- Conditions for playing

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
...	...	...	...	...

```
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
```


Weather data with mixed attributes

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Weather data with mixed attributes

- How will the rules change when some attributes have numeric values?

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
...	...	...	...	...

Weather data with mixed attributes

- Rules with mixed attributes

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
```


Some fun with WEKA

- Open WEKA preferably in Linux
- We need to find the data file
  - find . -name *arff -ls
  - May want to copy into an easier place to get to
  - gunzip *.gz
  - Take a look at the file format

The ARFF format

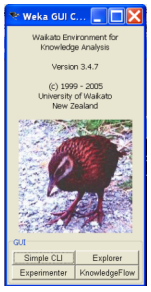
```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```


- Open Weka Explorer
- Open file…
- Choose weather.arff
  - Note that if you have a file in .csv format
    - E.g. from Excel
    - It can be opened and will be automatically converted to .arff format

Weka



Classifying Weather Data

- Click on Classify
  - Choose bayes -> NaïveBayesSimple
  - Choose trees -> J48
  - Try some more

Keep Exploring

- Try the iris data set
- Does it work better?

43