Machine Learning and Data Mining An Introduction with WEKA

AHPCRC Workshop - 8/18/10 - Dr. Martin

Based on slides by Gregory Piatetsky-Shapiro from Kdnuggets

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

Some review

- What are we doing?
- Data Mining
- And a really brief intro to machine learning

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

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

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

  To commit to memory

  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.

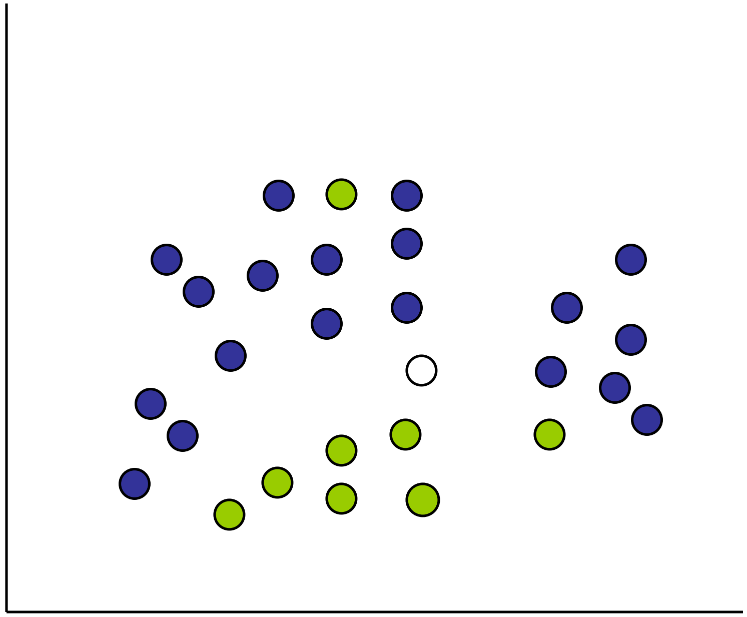
  } Does a slipper learn?

- Does learning imply intention?

witten&eibe

Classification

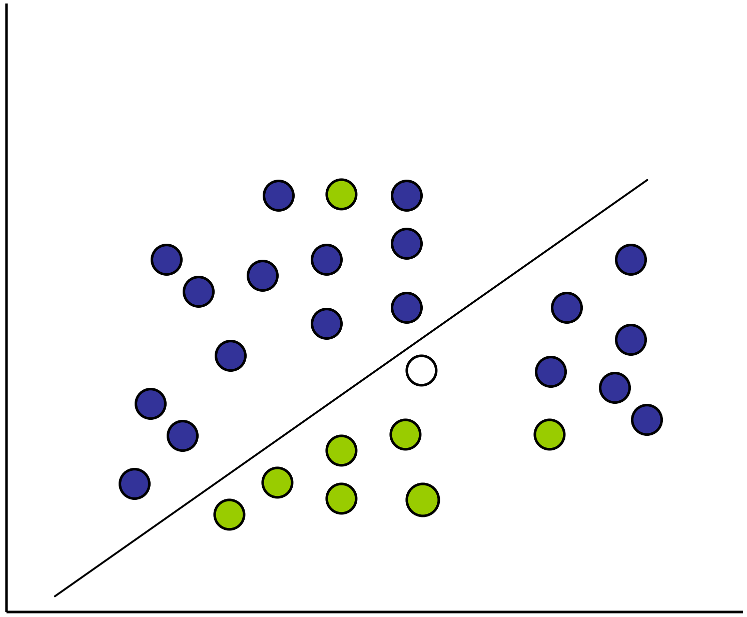
Learn a method for predicting the instance class from pre-labeled (classified) instances



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

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

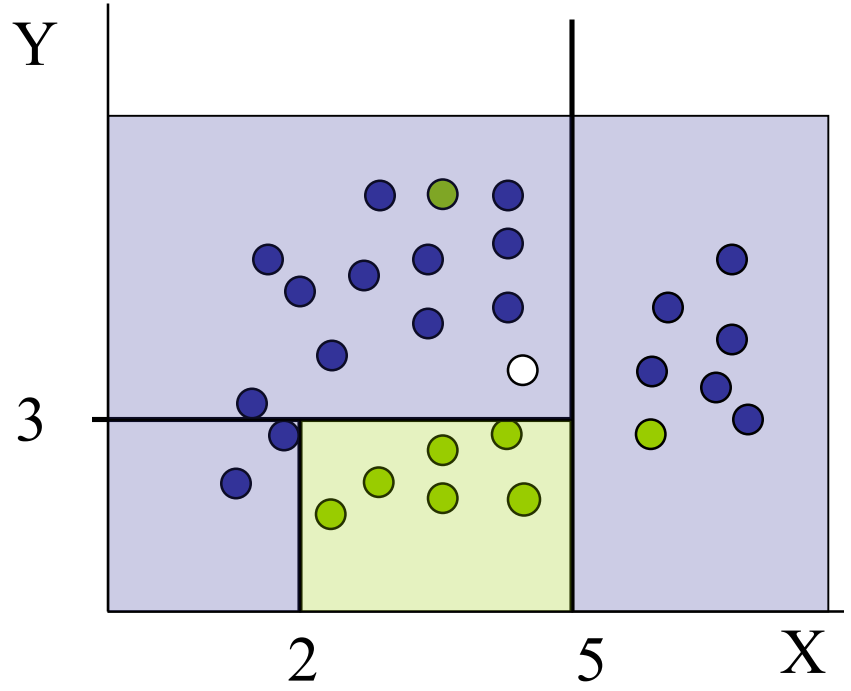
Classification: Linear Regression



- Linear Regression

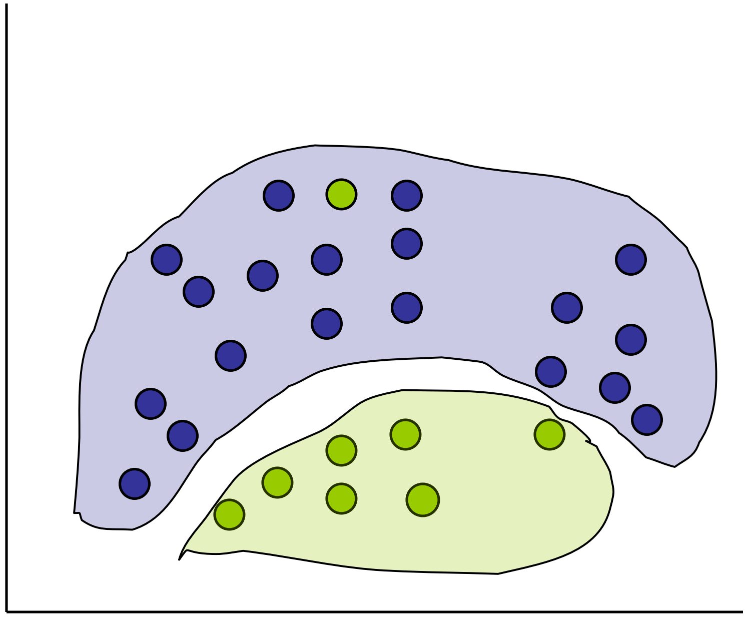
  $w_0 + w_1 x + w_2 y >= 0$
- Regression computes w_i from data to minimize squared error to ‘fit’ the data
- Not flexible enough

Classification: Decision Trees



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

Classification: Neural Nets



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

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

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

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*

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

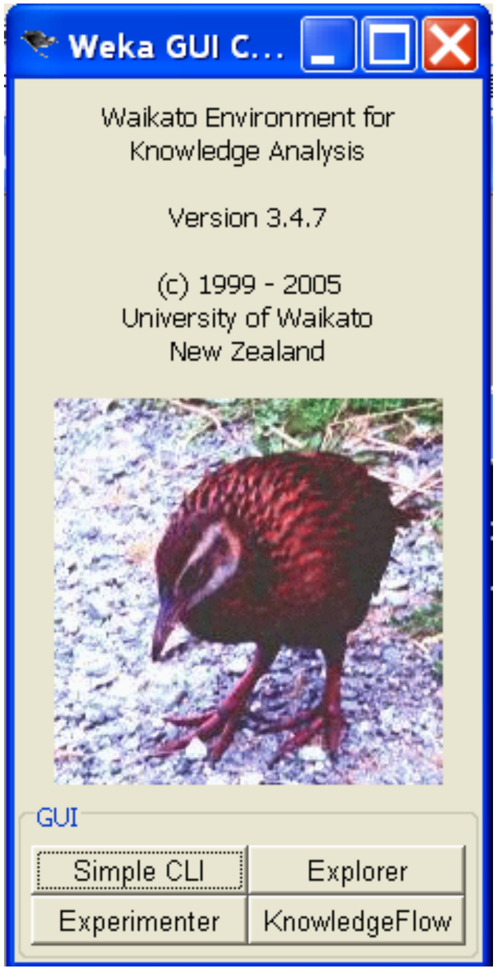
```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@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



Waikato Environment for
Knowledge Analysis

Version 3.4.7

(c) 1999 - 2005
University of Waikato
New Zealand

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?