Data Mining with Weka

Pitfalls and pratfalls

Ian H. Witten

Pitfalls and pratfalls

Pitfall: A hidden or unsuspected danger or difficulty

Pratfall: A stupid and humiliating action

Pitfalls and pratfalls

Be skeptical

- In data mining, it's very easy to cheat
  - *whether consciously or unconsciously*
- For reliable tests, use a completely fresh sample of data that has never been seen before

Overfitting has many faces

- Don't test on the training set (of course!)
- Data that has been used for development (in any way) is tainted
- Leave some evaluation data aside for the *very* end

Pitfalls and pratfalls

Missing values

“Missing” means what …

- ❖ Unknown?
- ❖ Unrecorded?
- ❖ Irrelevant?

Should you: 1. Omit instances where the attribute value is missing?
or 2. Treat "missing" as a separate possible value?

Is there significance in the fact that a value is missing?

Most learning algorithms deal with missing values
– but they may make different assumptions about them

Pitfalls and pratfalls

OneR and J48 deal with missing values in different ways

- Load weather-nominal.arff
- OneR gets 43%, J48 gets 50% (using 10-fold cross-validation)
- Change the outlook value to unknown on the first four no instances
- OneR gets 93%, J48 still gets 50%
- Look at OneR's rules: it uses "?" as a fourth value for outlook

Pitfalls and pratfalls

No free lunch

- 2-class problem with 100 binary attributes
- Say you know a million instances, and their classes (training set)
- You don't know the classes of $2^{100} - 10^6$ examples!

  (that's 99.9999…% of the data set)
- How could you possibly figure them out?

In order to generalize, every learner must embody some knowledge or assumptions beyond the data it's given

A learning algorithm implicitly provides a set of assumptions

There can be no "universal" best algorithm (no free lunch)

Data mining is an experimental science

Pitfalls and pratfalls

- Be skeptical
- Overfitting has many faces
- Missing values – different assumptions
- No "universal" best learning algorithm
- Data mining is an experimental science
- It's very easy to be misled