Weka and Machine Learning

Abram Hindle

2013-02-15 Fri

Outline

1. Weka
2. Machine Learning
3. Learners
4. Weka Stuff
5. Evaluation/Validation

Licensing

Intro

- AI that learns from data
  - Learn what spam looks likes to filter it out
- Classify data into types
  - Learning spam
- Cluster data by similarity
  - Finding messages that are similar to spam
- Find important and distinct properties of the data.
  - Viagra is a spam keyword!

Kinds of ML

- Supervised
  - we give it classified examples and hope it can classify more
- Unsupervised
  - labels unknown, let the algorithm find them
- Semi Supervised learning
  - labelled and unlablled.
- Reinforcement Learning
  - policies to reward the learner

Kinds of Learners

- Tree Based
  - C4.5 (J48)
  - Random Forest
  - Decision Tree
- Rule Learners
  - Ripper (jRip)
- Support Vector Machines
  - SVM/LibSCM
- Bayesian Nets

Weka Makes some disinction

- Bayes
- Functions
- Lazy
- Meta
- Misc
- Rules
- Trees

Learners operate on different classes and values

- Some learners are boolean (True/False)
- Some learners are class (A/B/C/..)
- Some learners learn counts (1,2,3,..)
- Some learners learn real functions ($Y = b + ax$)

ZeroR Learner

- The smartest monkey
- Always chooses the class with the largest number of entities
- Good as a base line.
- You have to beat ZeroR.

C4.5/J48

- Produces a decision tree
- The model is code and interpretable
- Sometimes trees are too big.
- each branch is a conditional
- each leaf is a class

JRip/Ripper

- learns and prunes a small set of rules
- copy & paste into code

Naive Bayes

- Asks the question what is the probability of this value belonging to this class?
- multiplies all of these probabilities together

Logistic Regression

- We've already discussed this
- Regression used for true false

K-NN

- nearest neighbor
- use euclidean distance to find the

ARFF Files

- Class should be the last element of the data
- Like CSV but with a type header
- String, Bool, Char, Class, Int, Float types
    - note different types for different types of jobs

Accuracy

- How many classifications were correct?
- If 90% of your data is 1 class you want better than 90% accuracy
- Bad for class imbalance

Kappa

- Cohen's Kappa
- like correlation
- agreement between classifier and actual data
- Very good for class imbalance

Precision

- How many of your classifications are right

Recall

- How much of the class did you find
- Might depend on the class
- You can have high precision for a class and have low recall

F-Measure

- Combination of Precision and Recall
- Geometric mean
- Can tune to one or the other

TP/FP Rate

- True Positives
- True Negatives
- Actual accuracy for all classes

ROC Area

- Area under the Receiver Operating Characteristic Curve
- We plot True Positive versus True Negative
- sensitivity (TPR) versus specificity (TNR)
- AUC ROC 0.5 - garbage
- AUC ROC 0.7 - good