Data Mining with Weka

Classification by regression

Ian H. Witten

Classification by regression

Can a regression scheme be used for classification? Yes!

Two-class problem

- Training: call the classes 0 and 1
- Prediction: set a threshold for predicting class 0 or 1

Multi-class problem: "multi-response linear regression"

- Training: perform a regression for each class
  - *Set output to 1 for training instances that belong to the class, 0 for instances that don't*
- Prediction: choose the class with the largest output

… or use "pairwise linear regression", which performs a regression for every pair of classes

Classification by regression

Investigate two-class classification by regression

- Open file diabetes.arff
- Use the NominalToBinary attribute filter to convert to numeric
  - *but first set Class: class (Nom) to No class,*
    because attribute filters do not operate on the class value
- Choose functions>LinearRegression
- Run
- Set Output predictions option

Classification by regression

More extensive investigation

Why are we doing this?

- It's an interesting idea
- Will lead to quite good performance
- Leads in to "Logistic regression" (next lesson), with excellent performance
- Learn some cool techniques with Weka

Strategy

- Add a new attribute ("classification") that gives the regression output
- Use OneR to optimize the split point for the two classes
  (first restore the class back to its original nominal value)

Classification by regression

- Supervised attribute filter AddClassification
  - *choose functions>LinearRegression as classifier*
  - *set outputClassification to true*
  - *Apply; adds new attribute called "classification"*
- Convert class attribute back to nominal
  - *unsupervised attribute filter NumericToNominal*
  - *set attributeIndices to 9*
  - *delete all the other attributes*
- Classify panel
  - *unset Output predictions option*
  - *change prediction from (Num) classification to (Nom) class*
- Select rules>OneR; run it
  - *rule is based on classification attribute, but it's complex*
- Change minBucketSize parameter from 6 to 100
  - *simpler rule (threshold 0.47) that performs quite well: 76.8%*

Classification by regression

- Extend linear regression to classification
  - *Easy with two classes*
  - *Else use multi-response linear regression, or pairwise linear regression*
- Also learned about
  - *Unsupervised attribute filter NominalToBinary, NumericToNominal*
  - *Supervised attribute filter AddClassification*
  - *Setting/unsetting the class*
  - *OneR's minBucketSize parameter*
- But we can do better: Logistic regression
  - *next lesson*