Data Mining with Weka

Baseline accuracy

Ian H. Witten

Baseline accuracy

Use diabetes dataset and default holdout

- Open file diabetes.arff
- Test option: Percentage split
- Try these classifiers:
  - *trees > J48* 76%
  - *bayes > NaiveBayes* 77%
  - *lazy > IBk* 73%
  - *rules > PART* 74%

  (we'll learn about them later)
- 768 instances (500 negative, 268 positive)
- Always guess "negative": 500/768 65%
- rules > ZeroR: most likely class!

Baseline accuracy

Sometimes baseline is best!

- Open supermarket.arff and blindly apply

rules > ZeroR	64%
trees > J48	63%
bayes > NaiveBayes	63%
lazy > IBk	38% (!!)
rules > PART	63%

- Attributes are not informative
- Don't just apply Weka to a dataset:
  you need to understand what's going on!

Baseline accuracy

- Consider whether differences are likely to be significant
- Always try a simple baseline,
  e.g. rules > ZeroR

- Look at the dataset
- Don't blindly apply Weka:
  try to understand what's going on!