Data Mining with Weka

Overfitting

Ian H. Witten

Overfitting

- Any machine learning method may "overfit" the training data …
  … by producing a classifier that fits the training data too tightly
- Works well on training data but not on independent test data
- Remember the "User classifier"? Imagine tediously putting a tiny circle around every single training data point
- Overfitting is a general problem
- … we illustrate it with OneR

Overfitting

Numeric attributes

Outlook	Temp	Humidity	Wind	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

Attribute	Rules	Errors	Total errors
Temp	85 → No	0/1	0/14
	80 → Yes	0/1	
	83 → Yes	0/1	
	75 → No	0/1	
	…	…	

- OneR has a parameter that limits the complexity of such rules
- How exactly does it work? Not so important …

Overfitting

Experiment with OneR

- Open file weather.numeric.arff
- Choose OneR rule learner (rules>OneR)
- Resulting rule is based on outlook attribute, so remove outlook
- Rule is based on humidity attribute

```
humidity:    < 82.5   -> yes
             >= 82.5  -> no
(10/14 instances correct)
```

Overfitting

Experiment with diabetes dataset

- Open file diabetes.arff
- Choose ZeroR rule learner (rules>ZeroR)
- Use cross-validation: 65.1%
- Choose OneR rule learner (rules>OneR)
- Use cross-validation: 72.1%
- Look at the rule (plas = plasma glucose concentration)
- Change minBucketSize parameter to 1:54.9%
- Evaluate on training set: 86.6%
- Look at rule again

Overfitting

- Overfitting is a general phenomenon that plagues all ML methods
- One reason why you must never evaluate on the training set
- Overfitting can occur more generally
- E.g try many ML methods, choose the best for your data
  - you cannot expect to get the same performance on new test data
- Divide data into training, test, validation sets?