Data Mining with Weka

Cross-validation results

Ian H. Witten

Cross-validation results

Is cross-validation really better than repeated holdout?

- Diabetes dataset
- Baseline accuracy (rules > ZeroR): 65.1%
- trees > J48
- 10-fold cross-validation 73.8%
- … with different random number seed

1	2	3	4	5	6	7	8	9	10
73.8	75.0	75.5	75.5	74.4	75.6	73.6	74.0	74.5	73.0

Cross-validation results

Sample mean $\bar{x} = \frac{\sum x_i}{n}$

Variance $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Standard deviation σ

holdout (10%)	cross-validation (10-fold)
75.3	73.8
77.9	75.0
80.5	75.5
74.0	75.5
71.4	74.4
70.1	75.6
79.2	73.6
71.4	74.0
80.5	74.5
67.5	73.0
$\bar{x} = 74.8$	$\bar{x} = 74.5$
$\sigma = 4.6$	$\sigma = 0.9$

Cross-validation results

- Why 10-fold? E.g. 20-fold: 75.1%
- Cross-validation really is better than repeated holdout
- It reduces the variance of the estimate