MODEL COMPARISON

SYRACUSE UNIVERSITY
School of Information Studies

BASELINES FOR MODEL EVALUATION

If your classification model reached 80% accuracy, is it "good enough"?

Two common baselines for comparison

- **Random guess**: If there are two categories, a model based on random guess would result in 50% accuracy.
- **Majority vote**: If the data set is skewed, a trivial model would assign all test data to the larger category.
  - In the Titanic training data set, the majority vote model would result in 549/891 = 62% accuracy.

Your model is expected to outperform the common baselines.

MAJORITY VOTE BASELINE

	Predictions			
Classes	buy_computer = yes	buy_computer = no	Total	Recall(%)
buy_computer = yes	7,000	0	7,000	1
buy_computer = no	3,000	0	3,000	0
Total	10,000	0	10,000	
Precision (%)	.70	n/a		

FAIR COMPARISON

When comparing the performance of two models, e.g., an unpruned tree vs. a pruned tree, make sure the comparison is fair, meaning, the test data should be exactly the same.

Common mistakes:

- Run hold-out test on one model but cross-validation on another model.
- Set up different numbers of folds for the two models when using cross-validation.
- Set up different split ratio for the two models when using hold-out test.

OTHER ASPECTS OF EVALUATION

When comparing two classification models, predictive capability (as measured by accuracy, precision, recall, etc.) is only one aspect to examine.

Other aspects:

- Speed
- Robustness
- Scalability
- Model interpretability

OTHER ASPECTS OF EVALUATION

Speed

- Time to construct model (training time)
- Time to use the model (classification/prediction time)

Robustness

- Handling noise and missing values

Scalability

- The data set size keeps increasing

Interpretability

- Understanding the insight provided by the model

IS THE MODEL GOOD ENOUGH?

There is always room for improvement for nontrivial prediction tasks.

Evaluation from system perspective

Evaluation from user perspective