

CMSC726 Spring 2006: Evaluation, Part II

readings: [Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria](#), Rich Caruana and Alexandru Bucykesy-Mizil
sources: Rich Caruana, Cornell University

Supervised Learning Performance Criteria

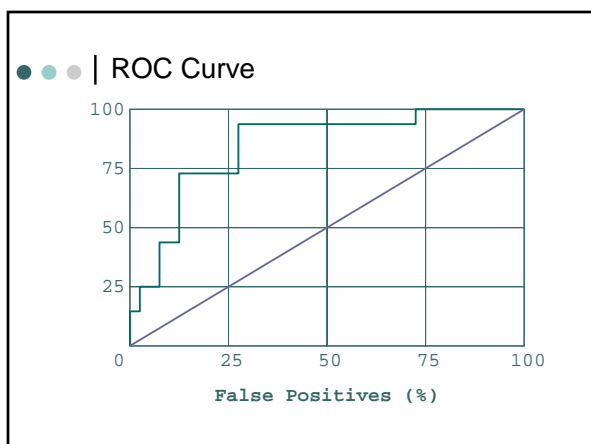
- Accuracy
- Lift
- F-Score
- Area under the ROC Curve
- Average Precision
- Precision/Recall Break-Even Point
- Squared Error
- Cross Entropy
- Probability Calibration

Types of Performance Criteria for Supervised Learning

- Interpret as threshold
 - classification accuracy
- Interpret as probabilities
 - (conditional) likelihood, squared error
- Interpret as ranking
 - ROC curves

Receiver Operator Characteristic (ROC) Curves

- Originally from signal detection
- Becoming very popular for ML
- Used in:
 - Two class problems
 - Where predictions are ordered in some way (e.g., neural network activation is often taken as an indication of how strong or weak a prediction is)
- Plotting an ROC curve:
 - Sort predictions (right) by their predicted strength
 - Start at the bottom left
 - For each positive example, go up $1/P$ units where P is the number of positive examples
 - For each negative example, go right $1/N$ units where N is the number of negative examples



ROC Properties

- Can visualize the tradeoff between coverage and accuracy (as we lower the threshold for prediction how many more true positives will we get in exchange for more false positives)
- Gives a better feel when comparing algorithms
 - Algorithms may do well in different portions of the curve
- A perfect curve would start in the bottom left, go to the top left, then over to the top right
 - A random prediction curve would be a line from the bottom left to the top right
- When comparing curves:
 - Can look to see if one curve dominates the other (is always better)
 - Can compare the area under the curve (very popular – some people even do t-tests on these numbers)

● ● ● | Lift

- Lift measures how much better a classifier is at predicting positives than a baseline classifier that randomly predicts positives (at the rate observed for positives in the data)

$$\text{LIFT} = \frac{\% \text{ of true positives about the threshold}}{\% \text{ of dataset about the threshold}}$$

● ● ● | Precision/Recall

- Precision: fraction of examples predicted as positive that are actually positive
- Recall: fraction of true positives that are predicted as positives
- Combining measures:
 - precision-recall F score: harmonic mean of the precision and recall at a given threshold
 - precision at recall level: set recall, measure precision
 - break even point: the precision at which the precision equals recall
 - average precision: average of the precisions at eleven evenly spaced recall levels.