Measuring Prediction Performance

CS4780/5780 – Introduction to Machine Learning

Thorsten Joachims
Cornell University

Why is 0/1 Error not Enough?

- Reason 1: Some applications have asymmetric cost of errors.
  - Example: Spam Filtering

$$\Delta(y', y) = \begin{cases} 1 & if\ y = Ham, y' = Spam \\ 10 & if\ y = Spam, y' = Ham \\ 0 & else \end{cases}$$

Why is 0/1 Error not Enough?

- Reason 2: Some applications have class imbalance.
  - Example: Search Engine
  - Feature vector: $\vec{\mathrm{x}} = \phi(query, document)$
  - Label: $y \in \{1, -1\}$ indicating relevance to query

  → Baseline classifier that always predicts -1 (i.e. not relevant) has 99.99999% accuracy.

Contingency Table

Counts	$y = +1$	$y = -1$
$h(\vec{x}) = +1$	True Positives (TP)	False Positives (FP)
$h(\vec{x}) = -1$	False Negatives (FN)	True Negatives (TN)

$$m = TP + FP + FN + TN$$

Performance measures

- Error Rate: $\frac{FP+FN}{m}$
- Accuracy: $\frac{TP+TN}{m}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- Weighted Loss: $\frac{\lambda_{TP}*TP+\lambda_{FP}*FP+\lambda_{FN}*FN+\lambda_{TN}*TN}{m}$

Classification vs. Ranking

Most rules output score, not just classification.

- SVM: $\vec{w} \cdot \vec{x} + b$
- Tree: Leaf purity
- K-NN: Weighted vote
- Naïve Bayes: P(Y|X)

→ Sort by score.

Example:

- ErrorRate=2/8 (same all "always -1")
- Recall=2/2 (0 for "always -1")
- Precision=2/4 (NaN for "always -1")

Test Example	Score	True Label
7	3.5	+1
3	2.1	-1
4	0.7	+1
5	0.1	-1
1	-0.4	-1
2	-0.9	-1
6	-2.3	-1
8	-5.1	-1

Evaluating Rankings: DCG

Discounted Cumulative Gain (DCG)

- Evaluate utility of a search-engine ranking r to the user.

$$DCG(r) = \sum_{(\vec{x}_i, y_i) \in S} \frac{1[y_i = 1]}{\log_2(rank(i) + 1)}$$

Test Example	Score	True Label
7	3.5	+1
3	2.1	-1
4	0.7	+1
5	0.1	-1
1	-0.4	-1
2	-0.9	-1
6	-2.3	-1
8	-5.1	-1

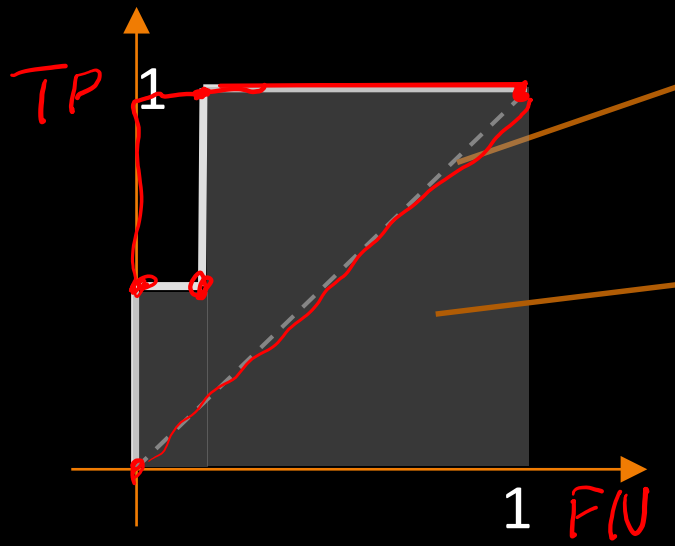
Example

- $DCG(r) = \frac{1}{\log_2(1+1)} + \frac{1}{\log_2(3+1)}$

Evaluating Rankings: ROC Area

Receiver Operating Characteristic (ROC)

- Sweep threshold from high to low and plot $(\frac{TP}{m_+}, \frac{FN}{m_-})$



Random ordering

ROC Area

Test Example	Score	True Label
7	3.5	+1
3	2.1	-1
4	0.7	+1
5	0.1	-1
1	-0.4	-1
2	-0.9	-1
6	-2.3	-1
8	-5.1	-1

Summary

- Error rate is only one among many performance measures
- Error rate is typically not informative for unbalanced classes
- Performance measure should be chosen to meaningful for the application