Information Theoretic Metrics for Multi-Class Predictor Evaluation

Sam Steingold, Michal Laclavík

Magnetic Media Online

NYC ML 2015-04-16

Table of Contents

Predictor

A *predictor* is a black box which spits out an estimate of an unknown parameter.

E.g.:

- Will it rain tomorrow?
- Will this person buy this product?
- Is this person a terrorist?
- Is this stock a good investment?

Examples

Perfect - always right

Mislabeled - always the opposite

Random - independent of the actual

- San Diego Weather Forecast:

  Actual : 3 days of rain per 365 days

  Predict : sunshine *always*!

- Coin flip

  Actual : true half the time

  Predict : true if coin lands Head

Why Evaluate Predictors?

- Which one is better?
- How much to *pay* for one?
  - You can always flip the coin yourself, so the *random* predictor is the least valuable!
- When to use this one and not that one?

Confusion Matrix + Cost

Predictor		Predicted		
		Sun	Rain	Hurricane
Actual	Sun	100	10	1
	Rain	5	20	6
	Hurricane	0	3	2

Costs		Predicted		
		Sun	Rain	Hurricane
Actual	Sun	0	1	3
	Rain	2	0	2
	Hurricane	10	5	0

Total cost (i.e., predictor value) $= 45$

Confusion/Costs Matrix

Probability/Costs		Predicted			
		Target		Non-target	
Actual	Bought	1%	\$1	9%	\$0
	Did not buy	9%	(\$0.1)	81%	\$0

Profitable

Expected value of one customer: $\$0.001 > 0$.

Worthless!

The **Predicted** and **Actual** are *independent*!

Table of Contents

What if the Cost is Unknown?

Total Population	N	Predicted	
		True(PT)	False(PF)
Actual	True(AT)	TP	FN(type2)
	False(AF)	FP(type1)	TN

Perfect : $FN = FP = 0$

Mislabeled : $TP = TN = 0$

Random (Predicted & Actual are independent) :

$$TP = \frac{PT \times AT}{N} \quad FN = \frac{PF \times AT}{N}$$
$$FP = \frac{PT \times AF}{N} \quad TN = \frac{PF \times AF}{N}$$

Metrics Based on the Confusion Matrix

8 partial measures

1. Positive predictive value (PPV, Precision): $\frac{TP}{PT}$
2. False discovery rate (FDR): $\frac{FP}{PT}$
3. False omission rate (FOR): $\frac{FN}{PF}$
4. Negative predictive value (NPV): $\frac{TN}{PF}$
5. True positive rate (TPR, Sensitivity, Recall): $\frac{TP}{AT}$
6. False positive rate (FPR, Fall-out): $\frac{FP}{AF}$
7. False negative rate (FNR): $\frac{FN}{AT}$
8. True negative rate (TNR, Specificity): $\frac{TN}{AF}$

Metrics Based on the Confusion Matrix

4 total measures

1. Accuracy: $\mathbb{P}(\text{Actual} = \text{Predicted})$.
2. F_1: the harmonic average of Precision and Recall
3. Matthew's Correlation Coefficient (MCC): AKA Pearson correlation coefficient.
4. Proficiency: the proportion of the information contained in the Actual distribution which is captured by the Predictor.

Metric Requirements

Meaning : the meaning of the metric should be transparent without resorting to averages of meaningful values

Discrimination :

- **Weak** : its value is 1 for the perfect predictor (and only for it)
- **Strong** : *additionally*, its value is 0 for a worthless (random with any base rate) predictor (and only for such a predictor)

Universality : the metric should be usable in any setting, whether binary or multi-class, classification (a unique class is assigned to each example) or categorization/community detection (an example can be placed into multiple categories or communities)

Accuracy

- $\mathbb{P}(\text{Actual} = \text{Predicted}) = \frac{\text{tp}+\text{tn}}{N}$
- Perfect: 1
- Mislabeled: 0
- Sun Diego Weather Forecast:
  - Accuracy $= 362/365 = 99.2\%$
  - The predictor is worthless!
- Does not detect a random predictor

F_1-Score

- The harmonic average of Precision and Recall: $\frac{2\times\mathrm{tp}}{2\times\mathrm{tp}+\mathrm{fp}+\mathrm{fn}}$
- Perfect: 1
- 0 if either Precision or Recall is 0
- Correctly handles SDWF (because Recall $= 0$)...
- ...But only if we label *rain* as True!
- Otherwise Recall $= 100\%$, Precision $= 99.2\%$, $F_1 = 99.6\%$
- F_1 is Asymmetric (Positive vs Negative)

Matthews correlation coefficient

- AKA Phi coefficient, Pearson correlation coefficient:

$$\frac{\mathrm{tp} \times \mathrm{tn} - \mathrm{fp} \times \mathrm{fn}}{\sqrt{(\mathrm{tp} + \mathrm{fp}) \times (\mathrm{tp} + \mathrm{fn}) \times (\mathrm{fp} + \mathrm{tn}) \times (\mathrm{fn} + \mathrm{tn})}}$$

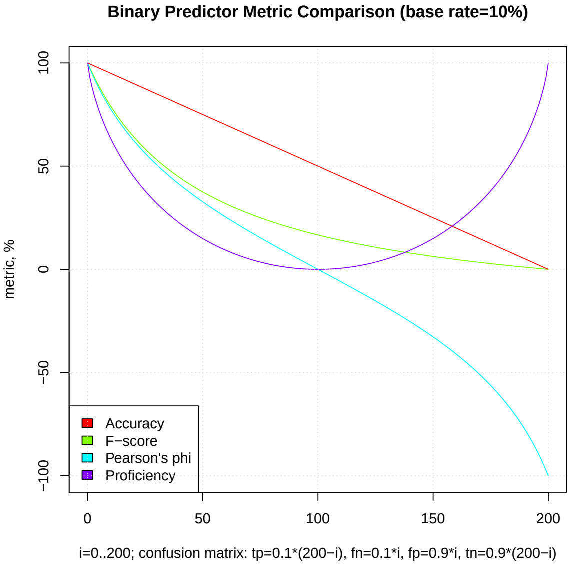
- Range: $[-1; 1]$
- Perfect: 1
- Mislabeled: -1
- Random: 0
- Handles San Diego Weather Forecast
- Hard to generalize to non-binary classifiers.

Uncertainty coefficient

- AKA Proficiency: $\alpha = \frac{I(Predicted;Actual)}{H(Actual)}$
- Range: $[0;1]$
- Measures the percentage of bits of information contained in the Actual which is captured by the Predictor.
- 1 for *both* Perfect *and* Mislabeled predictors
- 0 for the random predictor
- Handles San Diego Weather Forecast and all the possible quirks – *the best*.
- Easily generalizes to any number of categories.

Comparison

i=0	TP=20	FN=0
	FP=0	TN=180
i=50	TP=15	FN=5
	FP=45	TN=135
i=100	TP=10	FN=10
	FP=90	TN=90
i=150	TP=5	FN=15
	FP=135	TN=45
i=200	TP=0	FN=20
	FP=180	TN=0



2 Against 2 – take 1

$$A = \begin{bmatrix} \mathrm{tp} = 2 & \mathrm{fn} = 3 \\ \mathrm{fp} = 0 & \mathrm{tn} = 45 \end{bmatrix}; B = \begin{bmatrix} \mathrm{tp} = 5 & \mathrm{fn} = 0 \\ \mathrm{fp} = 7 & \mathrm{tn} = 38 \end{bmatrix}$$

	A	B
Proficiency	30.96%	**49.86%**
Pearson's ϕ	**61.24%**	59.32%
Accuracy	**94.00%**	86.00%
F_1-score	57.14%	**58.82%**

2 Against 2 – take 2

$$A = \begin{bmatrix} \text{tp} = 3 & \text{fn} = 2 \\ \text{fp} = 2 & \text{tn} = 43 \end{bmatrix}; B = \begin{bmatrix} \text{tp} = 5 & \text{fn} = 0 \\ \text{fp} = 7 & \text{tn} = 38 \end{bmatrix}$$

	A	B
Proficiency	28.96%	**49.86%**
Pearson's ϕ	55.56%	**59.32%**
Accuracy	**92.00%**	86.00%
F_1-score	**60.00%**	58.82%

Proficiency – The Odd One Out

$$A = \begin{bmatrix} \text{tp} = 3 & \text{fn} = 2 \\ \text{fp} = 1 & \text{tn} = 44 \end{bmatrix}; B = \begin{bmatrix} \text{tp} = 5 & \text{fn} = 0 \\ \text{fp} = 6 & \text{tn} = 39 \end{bmatrix}$$

	A	B
Proficiency	35.55%	**53.37%**
Pearson's ϕ	**63.89%**	62.76%
Accuracy	**94.00%**	88.00%
F_1-score	**66.67%**	62.50%

Accuracy – The Odd One Out

$$A = \begin{bmatrix} \text{tp} = 1 & \text{fn} = 4 \\ \text{fp} = 0 & \text{tn} = 45 \end{bmatrix}; B = \begin{bmatrix} \text{tp} = 5 & \text{fn} = 0 \\ \text{fp} = 13 & \text{tn} = 32 \end{bmatrix}$$

	A	B
Proficiency	14.77%	**34.57%**
Pearson's ϕ	42.86%	**44.44%**
Accuracy	**92.00%**	74.00%
F_1-score	33.33%	**43.48%**

F_1-score – The Odd One Out

$$A = \begin{bmatrix} \text{tp} = 1 & \text{fn} = 4 \\ \text{fp} = 0 & \text{tn} = 45 \end{bmatrix}; B = \begin{bmatrix} \text{tp} = 2 & \text{fn} = 3 \\ \text{fp} = 2 & \text{tn} = 43 \end{bmatrix}$$

	A	B
Proficiency	**14.77%**	14.71%
Pearson's ϕ	**42.86%**	39.32%
Accuracy	**92.00%**	90.00%
F_1-score	33.33%	**44.44%**

Predictor Re-Labeling

For a predictor P, let $1 - P$ be the re-labeled predictor, i.e., when P predicts 1, $1 - P$ predicts 0 and vice versa.

Then

$$\begin{aligned}\text{Accuracy}(1-P) &= 1 - \text{Accuracy}(P)\\ \phi(1-P) &= -\phi(P)\\ \alpha(1-P) &= \alpha(P)\end{aligned}$$

No similar simple relationship exists for F_1.

Table of Contents

Multi-Class Prediction

- Examples:
  - Character recognition
    - Mislabeling is bad
  - Group detection
    - Mislabeling is fine
- Metrics:
  - Accuracy $= \mathbb{P}(\text{Actual} = \text{Predicted})$
  - No Recall, Precision, F_1!

Pearson's ϕ

Define

$$\phi^2 = \frac{\chi^2}{N} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$O_{ij} = \mathbb{P}(\text{Predicted} = i \ \& \ \text{Actual} = j)$$
$$E_{ij} = \mathbb{P}(\text{Predicted} = i) \times \mathbb{P}(\text{Actual} = j)$$

- 0 for a worthless (independent) predictor
- Perfect predictor: depends on the data

Proficiency

Same as before!

$$\alpha = \frac{I(\text{Predicted}; \text{Actual})}{H(\text{Actual})}$$

$$H(A) = -\sum_{i=1}^{N} \mathbb{P}(A = i) \log_2 \mathbb{P}(A = i)$$

$$I(P; A) = \sum_{i=1}^{N} \sum_{j=1}^{N} O_{ij} \log_2 \frac{O_{ij}}{E_{ij}}$$

- 0 for the worthless predictor
- 1 for the perfect (and mis-labeled!) predictor

ϕ vs α

ϕ is to Chi-squared test
same as
α is to Likelihood-ratio test

NeymanPearson lemma

Likelihood-ratio test is the most powerful test.

This Metric is Old! Why is it Ignored?

Tradition : My teacher used it

Inertia : I used it previously

Cost : Log is more computationally expensive than ratios

- Not anymore!

Intuition : Information Theory is hard

- Intuition is learned: start Information Theory in High School!

Mislabeled = Perfect : Can be confusing or outright undesirable

- Use the Hungarian algorithm

Table of Contents

Multi-Label Categorization

- Examples:
  - Text Categorization
    - Mislabeling is bad
    - But may indicate problems with taxonomy
  - Community Detection
    - Mislabeling is fine
- Metrics:
  - No Accuracy: cannot handle partial matches
  - Precision & Recall work again!

Precision & Recall

$$\text{Recall} = \frac{\sum_i \#\{\text{objects correctly classified as } c_i\}}{\sum_i \#\{\text{objects actually in } c_i\}}$$
$$= \frac{\sum_i \#\{o_j \mid c_i \in \text{Actual}(o_j) \cap \text{Predicted}(o_j)\}}{\sum_i \#\{o_j \mid c_i \in \text{Actual}(o_j)\}}$$
$$= \frac{\sum_j \#[\text{Actual}(o_j) \cap \text{Predicted}(o_j)]}{\sum_j \#\text{Actual}(o_j)}$$

$$\text{Precision} = \frac{\sum_i \#\{\text{objects correctly classified as } c_i\}}{\sum_i \#\{\text{objects classified as } c_i\}}$$
$$= \frac{\sum_i \#\{o_j \mid c_i \in \text{Actual}(o_j) \cap \text{Predicted}(o_j)\}}{\sum_i \#\{o_j \mid c_i \in \text{Predicted}(o_j)\}}$$
$$= \frac{\sum_j \#[\text{Actual}(o_j) \cap \text{Predicted}(o_j)]}{\sum_j \#\text{Predicted}(o_j)}$$

Precision & Recall – ?!

- The above is the “macro” Precision & Recall (and F_1)
- Can also define “micro” Precision & Recall (and F_1)
- There is some confusion as to which is which

Side Note

Single label per object $\Longrightarrow$

$$\text{Precision} = \text{Recall} = \text{Accuracy} = F_1$$

Proficiency: Definition

Introduce binary random variables:

$$\mathrm{Ac}_i := c_i \in \mathrm{Actual}$$
$$\mathrm{Pc}_i := c_i \in \mathrm{Predicted}$$

Define:

$$\alpha = \frac{I(\prod_i \mathrm{Pc}_i; \prod_i \mathrm{Ac}_i)}{H(\prod_i \mathrm{Ac}_i)}$$

Problem: cannot compute!

- KDD Cup 2005 Taxonomy: 67 categories
- Cartesian product: $267 > 1020 \gg 800k$ examples

Proficiency: Estimate

Numerator : Assume that Ac_i is independent of everything but Pc_i (similar to Nave Bayes).

Denominator : Use $H(A \times B) \geq H(A) + H(B)$

Define:

$$\alpha = \frac{\sum_i I(\mathrm{Pc}_i; \mathrm{Ac}_i)}{\sum_i H(\mathrm{Ac}_i)} = \frac{\sum_i H(\mathrm{Ac}_i)\alpha(\mathrm{Pc}_i, \mathrm{Ac}_i)}{\sum_i H(\mathrm{Ac}_i)}$$

where

$$\alpha(\mathrm{Pc}_i, \mathrm{Ac}_i) = \frac{I(\mathrm{Pc}_i; \mathrm{Ac}_i)}{H(\mathrm{Ac}_i)}$$

Proficiency: Permuted

Recover re-labeling invariance

: Let $M(c)$ be the optimal assignment with the cost matrix being the pairwise mutual informations.
Define Permuted Proficiency metric:

$$\alpha' = \frac{\sum_i I(M(\mathrm{Pc}_i); \mathrm{Ac}_i)}{\sum_i H(\mathrm{Ac}_i)} = \frac{\sum_i H(\mathrm{Ac}_i)\alpha(M(\mathrm{Pc}_i), \mathrm{Ac}_i)}{\sum_i H(\mathrm{Ac}_i)}$$

M is optimal implies $\alpha \leq \alpha'$
(equality iff the optimal assignment is the identity.)

Proficiency: Properties

Meaning : (an estimate of) the share of the information contained in the actual distribution recovered by the classifier.

Strong Discrimination : yes!

Universality : the independence assumption above weakens the claim that the metric has the same meaning across all domains and data sets.

Example: KDD Cup 2005

- 800 queries
- 67 categories
- 3 human labelers

Actual	labeler 1	labeler 2	labeler 3
Predicted	labeler 2	labeler 3	labeler 1
Precision	63.48%	36.50%	58.66%
Recall	41.41%	58.62%	55.99%
α	24.73%	28.06%	33.26%
α'	25.02%	28.62%	33.51%
Reassigned	9	12	11

Each Human Against Dice

Pit each of the three human labelers against the random labeler with the same category probability distribution:

	Labeler 1	Labeler 2	Labeler 3
F_1	14.3%	7.7%	19.2%
examples/category	3.7 ± 1.1	2.4 ± 0.9	3.8 ± 1.1
categories/example	44 ± 56	28 ± 31	48 ± 71

Academic Setting

Consider a typical University department:

Every professor serves on 9 administrative committees out of 10 available.

Worthless Predictor

Assign each professor to 9 random committees.

Performance

- Precision = Recall = 90%
- Proficiency: $\alpha = 0$

Numeric Stability

- Think of the data as an infinite stream of observations, and view the actually available data as a sample.
- How would the metrics change if the sample is different?
- All metrics have approximately the same variability (standard deviation):
  - $\approx 1\%$ for 800 observations of KDD Cup 2005
  - $\approx 0.5\%$ for 10,000 observations in the Magnetic data set

Table of Contents

Summary

- If you know the costs use the expected value.
- If you know what you want (Recall/Precision &c) use it.
- If you want a general metric, use Proficiency instead of F_1.

Implementation

Python code in
`https://github.com/Magnetic/proficiency-metric`

Contributions of implementations in other languages are welcome!