MODEL EVALUATION METRICS

SYRACUSE UNIVERSITY
School of Information Studies

METRICS FOR MODEL PERFORMANCE

Accuracy is the most common measure, but it has limitations, especially on skewed data set.

Data set with similar number of examples in each category is “balanced,” otherwise “unbalanced” or “skewed.”

Titanic training data set is skewed, with more negative examples than positive ones.

- 549 “0”: Did not survive
- 342 “1”: Survived

PROBLEM WITH ACCURACY MEASURE

We need to learn some fundamental concepts first:

Confusion matrix for two classes (can be extended to multiple classes)

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

ACCURACY DEFINITION BASED ON CONFUSION MATRIX

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

LIMITATION OF ACCURACY

Consider a two-class problem:

- Number of Class 0 examples = 9,990
- Number of Class 1 examples = 10

If a model predicts every test example as “0,” the model’s accuracy is 9,990/10,000 = 99.9 %.

- Accuracy is misleading because the trivial model does not detect any Class 1 example.

TWO TYPES OF ERROR

Market analysis: To predict if a student is going to buy new computer or not.

Prediction result in a confusion matrix:

	Predictions		
Classes	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	6,000	1,000	7,000
buy_computer = no	500	2,500	3,000
Total	6,500	3,500	10,000

False negative: Missed customers

False positive: Wrong targets

WHICH TYPE OF ERROR MATTERS MORE?

For a company, one type of error might be more costly than the other.

E.g., one would rather send out more coupons than miss a potential buyer.

E.g., one would rather tolerate some junk mail in inbox than risk misclassify a regular mail to junk.

The accuracy measure does not differentiate these two types of errors, but precision and recall would do.

PRECISION AND RECALL

Concepts borrowed from the information retrieval field

Define precision and recall on each category

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

PRECISION

$$\text{Precision}_{\text{class=yes}} = \frac{a}{a+c} = \frac{TP}{TP+FP}$$

Meaning: Among all positive predictions, how many are correct?

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

RECALL

$$\text{Recall}_{\text{class=yes}} = \frac{a}{a+b} = \frac{TP}{TP+FN}$$

Meaning: Among all positive examples, how many are correctly predicted?

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

EXAMPLE: CALCULATE PRECISION AND RECALL

	Predictions			
Classes	buy_computer = yes	buy_computer = no	Total	Recall(%)
buy_computer = yes	6,000	1,000	7,000	6,000/7,000
buy_computer = no	500	2,500	3,000	2,500/3,500
Total	6,500	3,500	10,000	
Precision (%)	6,000/6,500	2,500/3,500		

F-MEASURE

An ideal model would achieve high precision and recall on all categories.

But in reality, precision and recall are like the two sides of a seesaw: If one goes up, the other might go down.

F-measure is a weighted average of precision and recall.

$$\mathrm{F}_{\mathrm{class=yes}} = \frac{2 \cdot \ precision \cdot \ recall}{precision + recall}$$