CSE4334/5334 Data Mining

Model Evaluation

(Put together from many sources)
(chapter 4.11)

Sharma Chakravarthy

Department of Computer Science and Engineering
University of Texas at Arlington
Fall 2019 (acknowledgement to Pang-Ning Tan, Michael Steinbach and Vipin Kumar, and Jiawei Han, Micheline Kamber and Jian Pei)

Model Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?
- Methods for Performance Evaluation
  - How to obtain reliable estimates?
- Methods for Model Comparison
  - How to compare the relative performance among competing models?

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes (efficiency) to classify or build models, scalability, etc.
- What is the accuracy of the model?
- How can we quantify it?
- Example: Suppose we have a model and a test data set
  - 1000 test data points; 550 actual yes and 450 actual no
  - Now, we run this test data set on the model and get its output
  - Suppose the model classifies/predicts 500 (out of 550) as yes correctly and 100 (out of 450) as no correctly!

What is the accuracy of the model?

Confusion matrix captures predicted info

- **True positives:** data points labeled as positive that are predicted as positive (500 out of 550)
- **True negatives:** data points labeled as negative that are predicted as negative (100 out of 450)
- ----------
- **False positives:** data points labeled as negative that are predicted as positive (350) also, false alarm!
- **False negatives:** data points labeled as positive that are predicted negative (50)

Balanced data …

Confusion matrix	ACTUAL CLASS		
		Class=Yes	Class=No
PREDICTED CLASS	Class=Yes	500 (TP)	350 (FP)
	Class=No	50 (FN)	100 (TN)

ideal	classifier!	Actual class	
		Yes	no
Predicted class	Yes	550 (TP)	0 (FP)
	no	0 (FN)	450 (TN)

- Most widely-used metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{500+100}{500+100+50+350} = 0.6$$

Why is accuracy not so good?
FP is high!
If FP reduces (to 100), accuracy goes up (to 0.85)!
When is accuracy 1?

1000 test data; 550 y, 450 n

Predicted 500 y correctly
Predicted 100 n correctly

Predicted 350 n as y incorrectly
Predicted 50 y as n incorrectly

Imbalance data …

	ACTUAL CLASS	
	Class=Yes	Class=No
PREDICTED CLASS Class=Yes	500 (TP)	350 (FP)
PREDICTED CLASS Class=No	1500 (FN)	100 (TN)

2450 test data; 2000 y, 450 n

Predicted 500 y correctly
Predicted 100 n correctly

Predicted 350 n as y incorrectly
Predicted 1500 y as n incorrectly

What contributed to low accuracy?

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{500 + 100}{500 + 100 + 1500 + 350} = 0.244$$

Another imbalance data …

	ACTUAL CLASS	
	Class=Yes	Class=No
PREDICTED CLASS Class=Yes	1950 (TP)	80 (FP)
PREDICTED CLASS Class=No	50 (FN)	20 (TN)

2100 data; 2000 y, 100 n

Predicted 1950 y correctly
Predicted 20 n correctly

Imbalance does not always mean Low accuracy.

- Most widely-used metric: $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

$$Accuracy = \frac{1950 + 20}{1950 + 20 + 50 + 80} = 0.94$$

Dependence of the representativeness of training set!

Depends on the generated model

Same imbalance data with a different approach!

2100 data; 2000 y, 100 n

Suppose I classify everything as y?
I can do this without even generating a model!

What is its accuracy? Confusion matrix?

That is,
Predicted 2000 y correctly
Predicted 0 n correctly

Smart model	ACTUAL CLASS	
	Class=Yes	Class=No
PREDICTED CLASS Class=Yes	2000 (TP)	100 (FP)
PREDICTED CLASS Class=No	0 (FN)	0 (TN)

What is the problem with this approach?

A single measure of accuracy does not seem to tell the whole story!

What is the solution?

$$Accuracy = \frac{2000}{2000+0+0+100} = 0.952$$

Evaluation

- Many a times there is a cost to misclassification?
  - ❖ You are classifying a patient as having a disease or not having a disease!
  - a) Cost of classifying a patient who does not have a disease as having one? (FP)
  - b) Cost of classifying a person with a disease as not having one? (FN)
- a) May be acceptable if there is a simple, inexpensive test to confirm that!
- On the other hand, b) is NOT acceptable. The cost is too high!
- Same goes for crime, legal matters, student grades, stock market, etc.
- How can we include cost in addition to accuracy?

9

Including Cost of Classification

500 test points
Actual: 300 y and 200 n

Cost Matrix	ACTUAL CLASS		
	C(i|j)	y	n
PREDICTED CLASS	y	-1	1
	n	100	0

Which classifier would you Choose and why?

low accuracy, low cost ?
high accuracy, high cost ?

M1 because low FN!

Model M_1 Count	ACTUAL CLASS		
		y	n
PREDICTED CLASS	y	250	150
	n	50	50

Accuracy = 60%
Cost = -250+150++5000 =4900

Model M_2 Count	ACTUAL CLASS		
		y	n
PREDICTED CLASS	y	200	50
	n	100	150

Accuracy = 70%
Cost = -200+50+10000+0 = 8950

Cost vs Accuracy

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

Accuracy is proportional to cost if

1. C(Yes|Yes)=C(No|No) = p
2. C(Yes|No)=C(No|Yes) = q

Cost	ACTUAL CLASS		
		Class=Yes	Class=No
PREDICTED CLASS	Class=Yes	p	q
	Class=No	q	p

N = a + b + c + d
Accuracy = (a + d)/N
Cost = p (a + d) + q (b + c)
= p (a + d) + q (N – a – d)
= q N – (q – p)(a + d)
= N [q – (q-p) × Accuracy]

Recall and Precision

Recall $= \frac{TP}{TP+FN}$ What proportion of positive values Does the model capture? Also termed sensitivity

Precision $= \frac{TP}{TP+FP}$ What proportion of positive values Are correct?

Weighted Accuracy $= \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$

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

- Recall is biased towards C(Yes|Yes) & C(Yes|No)
- Precision is biased towards C(Yes|Yes) & C(No|Yes)
- Can we reduce precision and recall to a single number? How should we combine them?
  - Arithmetic Mean, geometric mean, or harmonic mean?

Recall and Precision

Recall $= \frac{TP}{TP+FN}$ What proportion of relevant positive values That are retrieved?

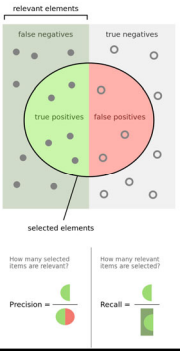
= P(retrieved | relevant)

Precision $= \frac{TP}{TP+FP}$ What proportion of retrieved Positive values that are relevant?

= P(relevant | retrieved)

Count		ACTUAL CLASS	
		Class=Yes (relevant)	Class=No (not relevant)
PREDICTED CLASS	Class=Yes (retrieved)	TP (a)	FP (c)
	Class=No (not retrieved)	FN (b)	TN (d)

Pictorially (wiki)



Others

- Sensitivity, recall, or True Positive Rate (TPR)
  - TPR = TP/(TP+FN)
- Specificity, selectivity, or true negative rate (TNR)
  - TNR = TN/(TN+FP)
- Precision or positive predictive value (PPV)
  - PPV = TP/(TP+FP)
- Miss rate or false negative rate (FNR)
  - FNR = FN/(FN+TP)
- Fall-out or false positive rate (FPR)
  - FPR = FP/(FP+TN)

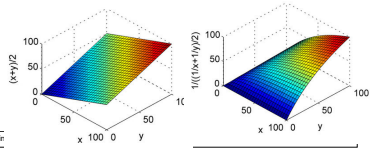
F1 Score

Recall $= \frac{TP}{TP+FN}$ Note: recall is 1 (max) when FN is 0

Precision $= \frac{TP}{TP+FP}$ Note: precision is 1 (max) when FP is 0

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

- What we want is for both (recall and precision) to be high! When both are 1, the model is perfect!
- Arithmetic Mean will work well at end points (0, 0) and (1, 1) as 0 and 1 correctly!
- We also want it to be low (0), when one is 0 (low) and the other is high (1)
  Mean is not appropriate for this!
- We want both numbers to contribute equally!
- Harmonic mean actually does this!
- looks for harmony in numbers



F1 Score or Harmonic mean of recall and precision

$F1 = 2 * \frac{precision * recall}{precision + recall}$

Look up the actual formula!

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

- F1 score for (0, 0) is 0
- F1 score for (1, 1) is 1
- F1 score for (0, 1) is 0 (not 0.5)
- F1 score for (1. 0) is 0 (not 0.5)
- F1 score gives equal weight to both measures. 1 in F1 indicates equal unit weight.
  - Can be customized as Fβ
- What is ROC curve?

HW: for the examples discussed earlier, Compute recall, precision, and F1 score And compare with accuracy

A combined measure: *F (general formula)*

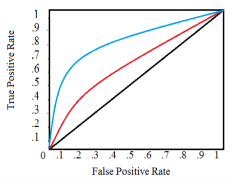
- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
  - i.e., with β = 1 or α = ½
- Harmonic mean is a conservative average
  - See CJ van Rijsbergen, *Information Retrieval*

18

ROC curve (receiver operating characteristic)



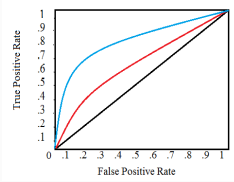
How do you plot it?

ROC is a graphical plot that illustrates the Diagnostic ability of a binary classifier system as its discrimination threshold is varied!

- Black line is considered a random classifier!
- Red and blue are two different models!
- True positive rate (TPR) (sensitivity) is the recall and
- False positive rate (FPR) is the probability of false alarm. Both are calculated from the confusion matrix!
- FPR is FP/(FP + TN)
- Area under the ROC curve is important.
  - The larger the area, the better is the model! Blue model is better than black and random models! (why?)
  - TPR increases better/faster than FPR!

19

ROC curve area



How does the best model Curve look like?

Y-axis is easy to understand! How to understand the x-axis And the area?

- If the TPR increases in proportion to FPR (note that they are independent!), you get the diagonal (base line or minimum expectation!)
- But if the recall (or TPR) increases at a higher rate (faster) than the FPR, the model is better!
- Hence the area under the curve (or AUC) is important evaluate a model!
- Blue model is better than the red model for that reason

20

Thank You !!!

For more information visit:
http://itlab.uta.edu