CSDS503 / COMP552 – Advanced Machine Learning

Faizad Ullah

Evaluation of Classifiers

Evaluation of Classifiers

- Loss Function:
  - To optimize the model's parameters, measures the difference between predicted and expected outputs of the model.

- Evaluation Metrics:
  - A Performance / Evaluation metrics is used to evaluate the model after training or during training.

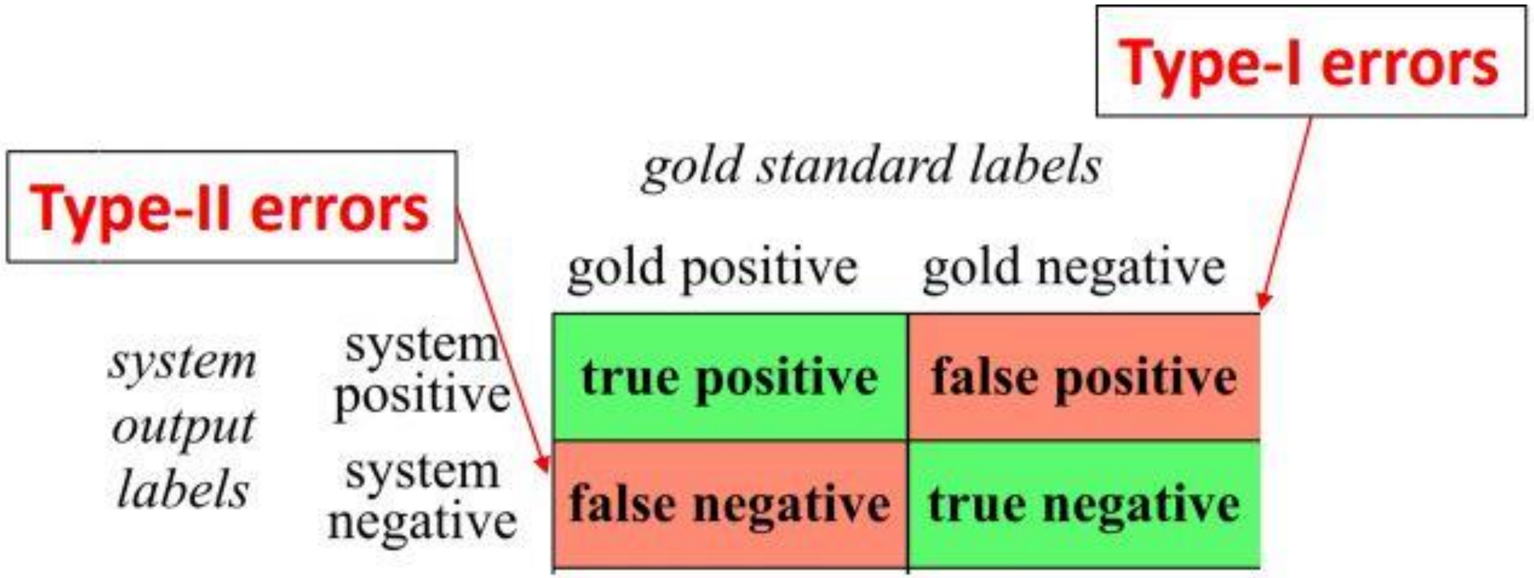
The Output of a Classifier

#	Height (inches)	Weight (kgs)	B.P. Sys	B.P. Dia	Heart disease			
	$\vec{x}$				y	$h(\vec{x})$		
1	62	70	120	80	No	No	True Negative	
2	72	90	110	70	No	Yes	False Positive	Type-I Error
3	74	80	130	70	No	No	True Negative	
4	65	120	150	90	Yes	Yes	True Positive	
5	67	100	140	85	Yes	No	False Negative	Type-II Error
6	64	110	130	90	No	Yes	False Positive	Type-I Error
7	69	150	170	100	Yes	Yes	True Positive	
8	75	127	160	95	Yes	No	False Negative	Type-II Error
9	66	66	135	90	Yes	Yes	True Positive	

$\boldsymbol{y}$: → Gold labels / Ground truth

$\boldsymbol{h}(\boldsymbol{x})$: → Predicted labels

Performance Measures


Type-I errors
Type-II errors
gold standard labels
gold positive
gold negative
system output labels
system positive
system negative
true positive
false positive
false negative
true negative

Accuracy

$$Accuracy\ \frac{tp + tn}{tp + fp + tn + fn}$$

Correct predictions over all predictions

A Real Example 1

- You want to know the people's sentiments about yourself, Ali.
- You build a system that detects tweets about you.
  - The positive class is "tweets about you", the negative class is all "other tweets".
- Imagine that you looked at a million tweets.
- 100 of them are "tweets about you", 999,900, are "other tweets".
- You created a classifier that stupidly classified every tweet as "not about you"
- Make a confusion matrix: (tp, fp, fn, tn)
  - 0 true positives
  - 0 false positives
  - 100 false negatives
  - 999,900 true negatives
- Accuracy = 999,900/1,000,000 or 99.99%!

A Real Example 1

- Accuracy is not a good metric when the goal is to discover a rare event, or at least not completely balanced in frequency.

- Class imbalance is a very common situation in the world!

		Gold Labels			
		Gold Positive	**Gold Negative**		
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)	$\frac{tp}{tp+fp}$	“Precision” aka "Positive Predictive Value”
	Predicted Negative	False Negatives (fn)	True Negatives (tn)	$\frac{tn}{fn+tn}$	“Negative Predictive Value”
		$\frac{tp}{tp+fn}$ “Recall” aka "Sensitivity" aka "True Positive Rate“	$\frac{tn}{fp+tn}$ "Specificity" aka "True Negative Rate“		
		$\frac{fn}{tp+fn}$ 1 - Sensitivity = "False Negative Rate“ aka “False Rejection Rate"	$\frac{fp}{fp+tn}$ 1 - Specificity = "False Positive Rate“ aka “False Acceptance Rate"		

$$Accuracy = \frac{tp+tn}{tp+fp+tn+fn}$$

- Accuracy: $\frac{My\ Correct\ Answers}{All\ Questions} = \frac{\text{tp + tn}}{\text{tp+tn+fp+fn}}$
  - What fraction of time am I correct in my classification
- Precision $\frac{True\ Positives}{My\ Positives} = \frac{\text{tp}}{\text{tp+fp}}$
  - How much should you trust me when I say that something tests positive
  - What fraction of my positives are true positives
- Recall = Sensitivity $\frac{True\ Positives}{Real\ Positives} = \frac{\text{tp}}{\text{tp+fn}}$
  - How much of the reality has been covered by my positive output?
  - What fraction of the true positives is captured by my positives?
- Specificity $\frac{True\ Negatives}{Real\ Negatives} = \frac{\text{tn}}{\text{tn+fp}}$
  - How much of the reality has been covered by my negative output?
  - What fraction of the true negatives is captured by my negatives?

A Real Example 2

- You are shown a set of 21 coins: 10 gold and 11 copper. Your task to accept all gold coins and reject all copper ones.

- You accept 7 coins as being gold (these are your positives)
  - 5 of these are actually gold (these are your true positives, tp)
  - 2 of these are copper (these are your false positives, fp)
  - You falsely rejected 5 gold ones (false negatives, fn)
  - You correctly rejected 9 copper ones (true negatives, tn)

A Real Example 2

	Actual Gold	Actual Copper
Predicted Gold	5	2
Predicted Copper	5	9

Accuracy = 14/21

Precision = 5/7

Recall = 5/10

Specificity = 9/11

Realistic Extremes

- You accept only one coin and that is gold
  - Your precision is very high (1/1) but recall is very low (1/10)

	Ac Gld	Ac Cop
Pr Gld	1	0
Pr Cop	9	11

- You return all 21 coins
  - Your recall is very high (10/10) but precision is very low (10/21)

	Ac Gld	Ac Cop
Pr Gld	10	11
Pr Cop	0	0

- Only one out of the 21 coins is gold. And you reject everything.
  - Your accuracy is very high (20/21 = 0.95) but precision/recall are 0.

	Ac Gld	Ac Cop
Pr Gld	0	0
Pr Cop	1	20

- So, what do we do now?
- A combined measure?

Issues with Precision and Recall

- One possible way may be to combine both.
- But, how to combine Precision and Recall?
- Average?

Arithmetic Mean

$$AM = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n}$$

For 2 values: $$AM = \frac{a_1 + a_2}{2}$$

Geometric Mean

$$GM = \sqrt[n]{a_1 . a_2 . a_3 \dots a_n}$$

For 2 values: $$GM = \sqrt[2]{a_1 a_2}$$

Harmonic Mean

$$HM = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \cdots + \frac{1}{a_n}}$$

For 2 values: $$HM = \frac{2}{\frac{1}{a_1} + \frac{1}{a_2}} = \frac{2a_1 a_2}{a_1 + a_2}$$

x0	x1	x2	x3	x4	x5	x6	x7	x8	AM	GM	HM
1	2	3	4	5	6	7	8	9	5.00	4.15	3.18
2	4	8	16	32	64	128	256	512	113.56	32.00	9.02
5	5	5	5	5	5	5	5	5	5.00	5.00	5.00
5	5	5	5	5	5	5	5	**10**	5.56	5.40	5.29
5	5	5	5	5	5	5	5	**100**	15.56	6.97	5.59
5	5	5	5	5	5	5	5	**1000**	115.56	9.01	5.62
5	5	5	5	5	5	5	5	**10000**	1115.56	11.63	5.62
5	5	5	5	5	5	5	5	**100000**	11115.56	15.03	5.62
5	5	5	5	5	5	5	100000	**100000**	22226.11	45.16	6.43
5	5	5	5	5	100000	100000	100000	**100000**	44447.22	407.89	9.00
5	100000	100000	100000	100000	100000	100000	100000	**100000**	88889.44	33274.21	44.98
100000	100000	100000	100000	100000	100000	100000	100000	**100000**	100000.0	100000.0	100000.0

F-1-MEASURE

- The harmonic mean of P and R:
  - Is high when both P and R are high.
  - Is low when even one of P and R is low.
- A combined measure that assesses the P/R tradeoff is the F-measure (weighted harmonic mean of precision and recall)

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

Precision	Recall	F-1
0	1	0
0.1	0.9	0.18
0.2	0.8	0.32
0.3	0.7	0.42
0.4	0.6	0.48
0.5	0.5	0.5
0.6	0.4	0.48
0.7	0.3	0.42
0.8	0.2	0.32
0.9	0.1	0.18
1	0	0
1	1	1
0.5	1	0.666667
1	0.5	0.666667
0.1	1	0.181818
1	0.1	0.181818

More than 2 Classes

More than two classes

- Lots of classification tasks in language processing have more than two classes:
  - Sentiment analysis (positive, negative, neutral),
  - Part-of-speech tagging (|POS tags|)
  - Emotion detection (|emotions|)

More than two classes

- Any-of or multi-label classification
  - An instance can belong to one or more than one class.
- One-of or multinomial classification
  - Classes are mutually exclusive: each instance in exactly one class

Evaluation

- one-of email categorization decision (urgent, normal, spam)

		gold labels			
		urgent	normal	spam	
system output	urgent	8	10	1	**precision**u= $\frac{8}{8+10+1}$
	normal	5	60	50	**precision**n= $\frac{60}{5+60+50}$
	spam	3	30	200	**precision**s= $\frac{200}{3+30+200}$
		recallu = $\frac{8}{8+5+3}$	**recall**n = $\frac{60}{10+60+30}$	**recall**s = $\frac{200}{1+50+200}$	

Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
  1. Macro-averaging: Compute performance for each class, then average.
  2. Micro-averaging: Collect decisions for all classes, compute contingency table, evaluate

gold labels

system output	urgent	normal	spam	
urgent	8	10	1	$\textbf{precision}_u = \frac{8}{8+10+1}$
normal	5	60	50	$\textbf{precision}_n = \frac{60}{5+60+50}$
spam	3	30	200	$\textbf{precision}_s = \frac{200}{3+30+200}$
	$\textbf{recall}_u = \frac{8}{8+5+3}$	$\textbf{recall}_n = \frac{60}{10+60+30}$	$\textbf{recall}_s = \frac{200}{1+50+200}$	

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

Micro Averaging

Evaluation

- A micro-average is dominated by the more frequent class (in this case spam)
  - The counts are pooled
- The macro-average better reflects the statistics of the smaller classes
  - More appropriate when performance on all the classes is equally important.

References

- Jurafsky and Martin, SLP3,
- https://web.stanford.edu/~jurafsky/slp3/