Introduction to Artificial Intelligence

Acknowledgement

- The slides of this lecture have been taken from the lecture slides of CS307 – "Introduction to Artificial Intelligence" and CSE652 – "Knowledge Discovery and Data mining" by Dr. Sajjad Haider.

How to Estimated Classification Accuracy or Error Rates

- Partition: Training-and-testing
  - use two independent data sets, e.g., training set (2/3), test set(1/3)
  - used for data set with large number of exmples
- Cross-validation
  - divide the data set into *k* subsamples
  - use *k-1* subsamples as training data and one sub-sample as test data—*k*-fold cross-validation
  - for data set with moderate size

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

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

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

Metrics for Performance Evaluation…

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	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 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10
- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

	PREDICTED CLASS		
	C(i|j)	**Class=Yes**	**Class=No**
ACTUAL CLASS	**Class=Yes**	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Cost Matrix (Cont'd)

	PREDICTED CLASS		
		True	False
ACTUAL CLASS	True	10	5
	False	1	14

	PREDICTED CLASS		
		True	False
ACTUAL CLASS	True	10	3
	False	3	14

	PREDICTED CLASS		
		True	False
ACTUAL CLASS	True	10	6
	False	0	14

All three confusion matrices have the same accuracy value, i.e., **24 / 30**

What if the cost of misclassification is not the same for both type of errors?

Cost Matrix (Cont'd)

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	5x5
	False	1	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	3x5
	False	3	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	6x5
	False	0	14

Suppose the cost of misclassifying True as False is 5 while the cost of misclassifying False as True is 1.

Accuracy values are:
24/50, 24/42, 24/54

Cost Matrix (Cont'd)

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	5x4
	False	1	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	3x4
	False	3	14

		PREDICTED CLASS	
		True	False
ACTUAL CLASS	True	10	6x4
	False	0	14

Suppose the cost of misclassifying True as False is **4** while the cost of misclassifying False as True is 1.

Accuracy values are:
24/45, 24/39, 24/48

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F-measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

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

Recall and Precision

Actual	Prediction
T	T
T	F
F	T
F	F
F	T
T	T
T	T
T	F
F	T
T	T

Recall and Precision

Actual	Prediction
T	T
T	F
F	T
F	F
F	T
T	T
T	T
T	F
F	T
T	T

- Recall = 4 / 6

Recall and Precision

Actual	Prediction
T	T
T	F
F	T
F	F
F	T
T	T
T	T
T	F
F	T
T	T

- Recall = 4 / 6
- Precision = 4 / 7
- F-Measure = 8 / 13