Performance Metrics and Clustering Analysis

Lesson 7

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Lecture Overview

- Performance Metrics of Models
- ROC / AUC
- Imbalanced Data
- Clustering Analysis
- K-means Clustering

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Performance Metrics

Performance Metrics in Classification

		Actual	
		Positive	Negative
Predicted	Pos	n_{11}	n_{12}
	Neg	n_{21}	n_{22}

- Type I error (False positive):
$$\Pr(\hat{y} = Pos \mid y = Neg) = n_{12}/(n_{12} + n_{22}) = n_{12}/n_{\bullet 2}$$
- Type II error(False negative):
$$\Pr(\hat{y} = Neg \mid y = Pos) = n_{21}/(n_{11} + n_{21}) = n_{21}/n_{\bullet 1}$$

- Accuracy: $(n_{11} + n_{22})/(n_{11} + n_{22} + n_{12} + n_{21})$
- Sensitivity (True Positive Rate, Recall): Among the (n_{11}+n_{21}) true positive cases, the percentage that is predicted as positive:
$$\Pr(\hat{y} = Pos \mid y = Pos) = n_{11}/(n_{11} + n_{21}) = n_{11}/n_{\bullet 1} = 1 - Type\ II\ \text{Error}$$
- Precision: Among the (n_{11}+n_{12}) predicted positive cases, the percentage that is actually positive:
$$\Pr(y = Pos \mid \hat{y} = Pos) = n_{11}/(n_{11} + n_{12}) = n_{11}/n_{1\bullet}$$

F-Score

- Sometimes we want a single number that informs us of the quality of the solution. A popular way to combine precision (P) and recall (R) into a single number is by taking their harmonic mean. This is known as the balanced f-measure:

$$F = \frac{2 \times P \times R}{P + R}$$

Note: Precision, recall, f-measure depend on what class is considered.

Example:

- Binary data set, flip positive and negative → different precision and recall
- Precision on flipped task != recall on original task
  - Or vice versa

Best advice:

- If less sure, report 2 sets of precision/recall/f-measure which vary based on the class to spot.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Example of Flipping Positive and Negative, and the Impacts on Performance Metrics

		Actual	
		Positive	Negative
Predicted	Pos	n_{11}	n_{12}
	Neg	n_{21}	n_{22}

- **Accuracy**: $(n_{11}+n_{22})/(n_{11}+n_{22}+n_{12}+n_{21})$
- **Sensitivity** (True Positive Rate, Recall): Among the (n_{11}+n_{21}) true positive cases, the percentage that is predicted as positive:

$$\Pr(\hat{y}=Pos \mid y=Pos)=n_{11}/(n_{11}+n_{21})$$

- **Precision**: Among the (n_{11}+n_{12}) predicted positive cases, the percentage that is actually positive:

$$\Pr(y=Pos \mid \hat{y}=Pos)=n_{11}/(n_{11}+n_{12})$$

		Actual	
		Negative	Positive
Predicted	Neg	n_{11}	n_{12}
	Pos	n_{21}	n_{22}

- **Recall:**

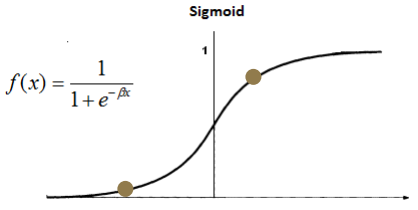
$$\Pr(\hat{y}=Pos \mid y=Pos)=n_{22}/(n_{12}+n_{22})$$

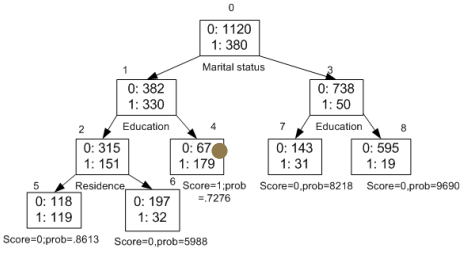
- **Precision:**

$$\Pr(y=Pos \mid \hat{y}=Pos)=n_{22}/(n_{21}+n_{22})$$

Accuracy, Recall, and Precision All Depend on Threshold

- Usually, classification model always outputs probability that an observation belongs to a class





- Whether we assign label 0 or 1 depends on whether the probability is greater than a threshold
  - In many algorithms, 0.5 is the default threshold

Let's See How the Threshold Impacts the Performance Metrics

Education	Age	Employer Sector	Prob(Salary>65k)	Actual	Threshold
High-School	50	Government	0.38	0	0.5
Bachelor	45	Private	0.69	1	
Associate	30	Non-profit	0.61	0	
Bachelor	36	Private	0.73	1	
Master	42	Private	0.82	1	
PhD	48	Government	0.7	1	
Master	25	Private	0.56	1	
Associate	20	Non-profit	0.48	0	
Bachelor	37	Private	0.92	0	
PhD	51	Government	0.79	1	

		Actual	
		Positive	Negative
Predict	Positive	6	2
	Negative	0	2
Recall=	100		
Precision=	75		

Education	Age	Employer Sector	Prob(Salary>65k)	Actual	Threshold
High-School	50	Government	0.38	0	0.8
Bachelor	45	Private	0.69	1	
Associate	30	Non-profit	0.61	0	
Bachelor	36	Private	0.73	1	
Master	42	Private	0.82	1	
PhD	48	Government	0.7	1	
Master	25	Private	0.56	1	
Associate	20	Non-profit	0.48	0	
Bachelor	37	Private	0.92	0	
PhD	51	Government	0.79	1	

		Actual	
		Positive	Negative
Predict	Positive	1	1
	Negative	5	3
Recall=	16.66667		
Precision=	50		

- Different threshold, you may get different performance metrics
- Is there a performance metrics that is independent with the threshold?

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Performance Metrics for Classification Problems
- –Recall, Precision - % of actual positive,
- –F-Score – combination of precision and recall
- –Type I errors – false positive
- –Type II errors – false negative

>Dependencies of Recall, Precision, F-Score, and Accuracy on Probability Threshold