CSE8803: Classification Methods

2018.06.08
딥메니아 파트2-1

Introduction

Big data big pictures

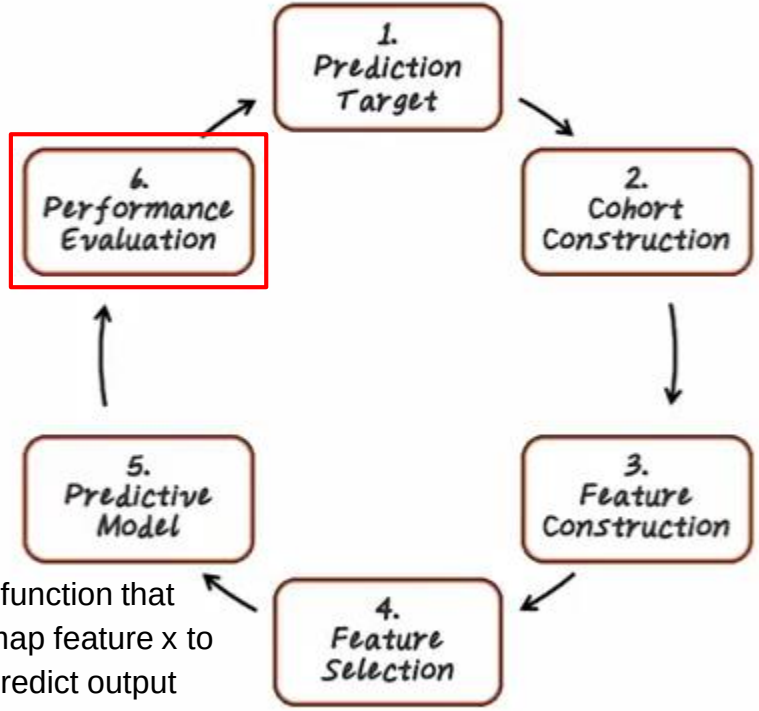
1. Big data systems
2. Scalable mission learning algorithms
3. Healthcare applications
    a. Predictive model: using historical data to view the model for predicting future outcome

       ex) predict which treatment is likely to work for an epilepsy patient

    a. Computational phenotyping is about turning messy electronic health records into meaningful clinical concepts.
    b. Patient similarity: uses health data to identify groups of patients sharing similar characteristics.

1. Predictive Model Review

PREDICTIVE MODELING PIPELINE



: function that map feature x to predict output target y

PREDICTIVE MODELS

Target Error

$$y = f(x) + e$$

Features

REGRESSION

- Target y is continuous
- Performance Metrics
  - Mean absolute error
  - Mean squared error
  - R^2

CLASSIFICATION

- Target y is binary
- Performance Metrics
  - True/False positive rate
  - Positive predictive values
  - F1
  - Area under the ROC curve
  - ...

2. Confusion Matrix

PERFORMANCE METRICS: CONFUSION MATRIX

= Contingency table

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

PERFORMANCE METRICS: CONFUSION MATRIX

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

Ground Truth	
Condition Positive	Condition Negative

PERFORMANCE METRICS: CONFUSION MATRIX

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

True Positive

PERFORMANCE METRICS: CONFUSION MATRIX

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

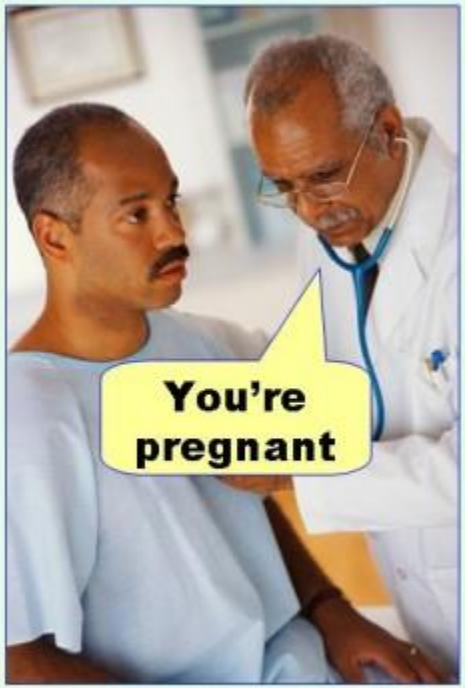
False Positive
(Type I error)

PERFORMANCE METRICS: CONFUSION MATRIX

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

False Negative
(Type II error)

Type I error
(false positive)



Type II error
(false negative)

PERFORMANCE METRICS: CONFUSION MATRIX

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

True Negative

CONFUSION MATRIX QUIZ

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
	[ ]	[ ]	935
Prediction	Prediction Outcome Positive 155	True Positive	False Positive
		[ ]	100
	Prediction Outcome Negative	False Negative	True Negative
	[ ]	10	[ ]

Please fill in the missing numbers.

CONFUSION MATRIX QUIZ

		Ground Truth	
	TOTAL POPULATION 1000	Condition Positive 65	Condition Negative 935
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100
	Prediction Outcome Negative 845	False Negative 10	True Negative 835

Please fill in the missing numbers.

3. Accuracy Metrics

PERFORMANCE METRICS: ACCURACY

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative
Accuracy = $\frac{\text{True positive} + \text{True negative}}{\text{Total population}}$		True Positive Rate = $\frac{\text{True positive}}{\text{Condition positive}}$	False Positive Rate = $\frac{\text{False Positive}}{\text{Condition negative}}$
		False Negative Rate = $\frac{\text{False negative}}{\text{Condition positive}}$	True Negative Rate = $\frac{\text{True negative}}{\text{Condition negative}}$

True Positive Rate (Sensitivity, Recall) = $\frac{\text{True positive}}{\text{Condition positive}}$

Sensitivity (민감도): 실제로 질병이 있는 케이스에서 질병이 있다고 판정하는 분율

PERFORMANCE METRICS: ACCURACY

		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative
Accuracy = $\frac{\text{True positive} + \text{True negative}}{\text{Total population}}$		True Positive Rate = $\frac{\text{True positive}}{\text{Condition positive}}$	False Positive Rate = $\frac{\text{False Positive}}{\text{Condition negative}}$
		False Negative Rate = $\frac{\text{False negative}}{\text{Condition positive}}$	True Negative Rate = $\frac{\text{True negative}}{\text{Condition negative}}$

True Negative Rate (Specificity) = $\frac{\text{True negative}}{\text{Condition negative}}$

Specificity (특이도): 실제로 질병이 없는 케이스에서 질병이 없다고 판정하는 분율

ACCURACY METRICS QUIZ

		Ground Truth	
	TOTAL POPULATION **1000**	Condition Positive 65	Condition Negative 935
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100
	Prediction Outcome Negative 845	False Negative 10	True Negative 835
Accuracy **89%**		True Positive Rate ☐	False Positive Rate ☐
		False Negative Rate ☐	True Negative Rate ☐

Please fill in the missing numbers.

ACCURACY METRICS QUIZ

		Ground Truth	
	TOTAL POPULATION **1000**	Condition Positive 65	Condition Negative 935
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100
	Prediction Outcome Negative 845	False Negative 10	True Negative 835
$\frac{55+835}{1000}$ **Accuracy** **89%**		True Positive Rate $\frac{55}{65}$ **85%**	False Positive Rate $\frac{100}{935}$ **11%**
		False Negative Rate $\frac{10}{65}$ **15%**	True Negative Rate $\frac{835}{935}$ **89%**

Please fill in the missing numbers.

4. Predictive Metrics

PERFORMANCE METRICS: PREDICTIVE

		Ground Truth		
	TOTAL POPULATION	Condition Positive	Condition Negative	Prevalence = Condition Positive / Total population
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)	Positive Predictive Value = True Positive / Prediction outcome positive
				False Discovery Rate = False Positive / Prediction outcome positive
	Prediction Outcome Negative	False Negative (Type II error)	True Negative	False Omission Rate = False negative / Prediction outcome negative
				Negative Predictive Value = True negative / Prediction outcome negative
Accuracy = (True positive + True negative) / Total population		True Positive Rate = True positive / Condition positive	False Positive Rate = False Positive / Condition negative	
		False Negative Rate = False negative / Condition positive	True Negative Rate = True negative / Condition negative	

Positive Predictive Value (양성예측도): 검사에서 질병이 있다고 판정한 케이스에서 실제로 질병이 있는 분율

Negative Predictive Value (음성예측도): 검사에서 질병이 없다고 판정한 케이스에서 실제로 질병이 없는 분율

PREDICTIVE METRICS QUIZ

		Ground Truth			
	TOTAL POPULATION **1000**	Condition Positive 65	Condition Negative 935	**Prevalence**	
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100	Positive Predictive Value	False Discovery Rate
	Prediction Outcome Negative 845	False Negative 10	True Negative 835	False Omission Rate 1%	Negative Predictive Value 99%
Accuracy **89%**		True Positive Rate 85%	False Positive Rate 11%		
		False Negative Rate 15%	True Negative Rate 89%		

Please fill in the missing numbers.

PREDICTIVE METRICS QUIZ

		Ground Truth		
	TOTAL POPULATION 1000	Condition Positive 65	Condition Negative 935	$\frac{65}{1000}$ **Prevalence** 7%
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100	Positive Predictive Value $\frac{55}{155}$ 35%
				False Discovery Rate $\frac{100}{155}$ 65%
	Prediction Outcome Negative 845	False Negative 10	True Negative 835	False Omission Rate $\frac{10}{845}$ 1%
				Negative Predictive Value $\frac{835}{845}$ 99%
$\frac{55+835}{1000}$ **Accuracy** 89%		True Positive Rate $\frac{55}{65}$ 85%	False Positive Rate $\frac{100}{935}$ 11%	
		False Negative Rate $\frac{10}{65}$ 15%	True Negative Rate $\frac{835}{935}$ 89%	

Please fill in the missing numbers.

5. F1 score

F_1 SCORE

: harmonic mean of those two measures, positive predictive value(precision) and the true positive rate(recall)

		Ground Truth		
	TOTAL POPULATION	Condition Positive	Condition Negative	**Prevalence** = $\frac{\text{Condition Positive}}{\text{Total population}}$
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)	Positive Predictive Value = $\frac{\text{True Positive}}{\text{Prediction outcome positive}}$	False Discovery Rate = $\frac{\text{False Positive}}{\text{Prediction outcome positive}}$
	Prediction Outcome Negative	False Negative (Type II error)	True Negative	False Omission Rate = $\frac{\text{False negative}}{\text{Prediction outcome negative}}$	Negative Predictive Value = $\frac{\text{True negative}}{\text{Prediction outcome negative}}$
Accuracy = $\frac{\text{True positive + True negative}}{\text{Total population}}$		True Positive Rate = $\frac{\text{True positive}}{\text{Condition positive}}$	False Positive Rate = $\frac{\text{False Positive}}{\text{Condition negative}}$	
		False Negative Rate = $\frac{\text{False negative}}{\text{Condition positive}}$	True Negative Rate = $\frac{\text{True negative}}{\text{Condition negative}}$	

$$F_1 = 2 \times \frac{PPV \times TPR}{PPV + TPR}$$

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

https://en.wikipedia.org/wiki/F1_score

F_1 QUIZ

		Ground Truth			
	TOTAL POPULATION **1000**	Condition Positive 65	Condition Negative 935	**Prevalence** **7%**	
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100	Positive Predictive Value 35%	False Discovery Rate 65%
	Prediction Outcome Negative 845	False Negative 10	True Negative 835	False Omission Rate 1%	Negative Predictive Value 99%
Accuracy **89%**		True Positive Rate 85%	False Positive Rate 11%		
		False Negative Rate 15%	True Negative Rate 89%		

Please calculate the F_1 score.

F_1 QUIZ

		Ground Truth			
	TOTAL POPULATION **1000**	Condition Positive 65	Condition Negative 935	**Prevalence** **7%**	
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100	Positive Predictive Value 35%	False Discovery Rate 65%
	Prediction Outcome Negative 845	False Negative 10	True Negative 835	False Omission Rate 1%	Negative Predictive Value 99%
Accuracy **89%**		True Positive Rate 85%	False Positive Rate 11%		
		False Negative Rate 15%	True Negative Rate 89%		

Please calculate the F_1 score.

0.5

$$2 \times \frac{(0.35 \times 0.85)}{(0.35+0.85)}$$

$$= 0.495833333$$

CLASSIFIER QUIZ

Which of these is the best classifier?

○A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

PPV = 0.69

F_1 = 0.66

Accuracy = 0.68

○B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

PPV = 0.50

F_1 = 0.61

Accuracy= 0.50

○C

TP=76	FP=12	88
FN=24	TN=88	112
100	100	200

PPV = 0.86 = $\frac{76}{88}$ TPR = 0.76 = $\frac{76}{100}$

F_1 = 0.81 = 2 X $\frac{(0.86 \times 0.76)}{(0.86+0.76)}$

Accuracy = 0.82 = $\frac{76+88}{200}$

CLASSIFIER QUIZ

Which of these is the best classifier?

higher performance matrix

A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

PPV = 0.69

F_1 = 0.66

Accuracy = 0.68

B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

PPV = 0.50

F_1 = 0.61

Accuracy= 0.50

C

TP=76	FP=12	88
FN=24	TN=88	112
100	100	200

PPV = 0.86

F_1 = 0.81

Accuracy = 0.82

CLASSIFIER QUIZ 2

Which of these is the best classifier?

C can be easily improved by reversing the prediction

○ A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

PPV = 0.69

F_1 = 0.66

ACC = 0.68

○ B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

PPV = 0.50

F_1 = 0.61

ACC = 0.50

✓ C

TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

PPV = 0.21

F_1 = 0.22

ACC = 0.18

5. ROC

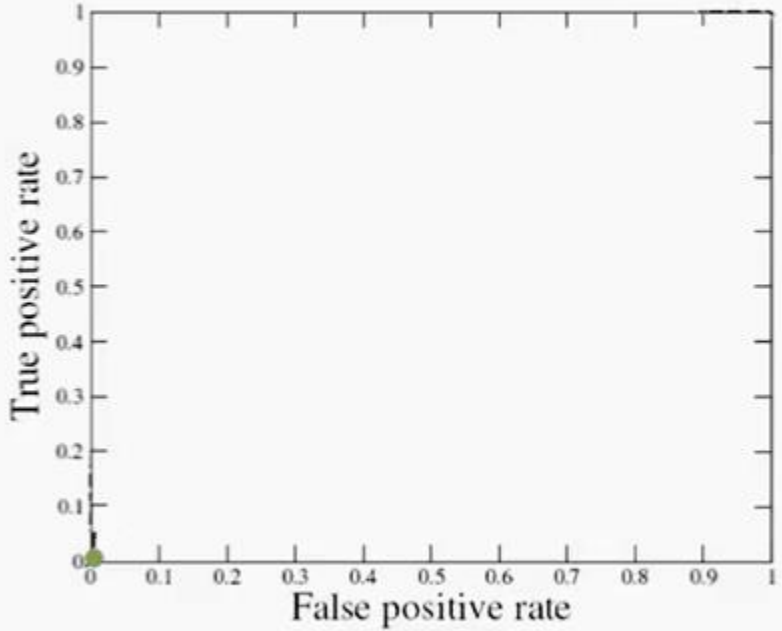
Prediction score of binary classification will be between 0 and 1

The value closer to 1 means prediction outcome positive, and the value closer to 0 means prediction outcome negative.

The ROC curve illustrates overall performance of a classifier.

Since AUC (area under this ROC curve) doesn't depend on the choice of the threshold, it becomes the most popular performance metric for classification problems.

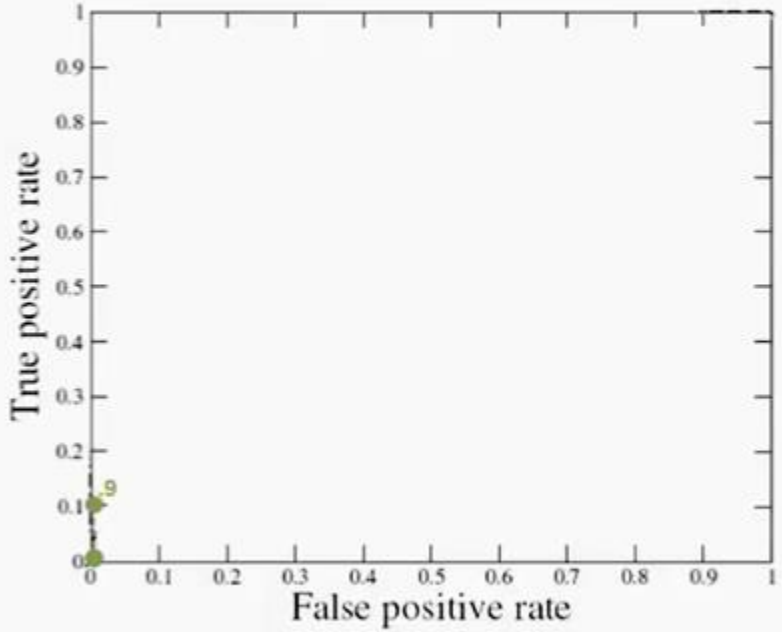
RECEIVER OPERATING CHARACTERISTIC (ROC)



#p = 10, #n=10

Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	p	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

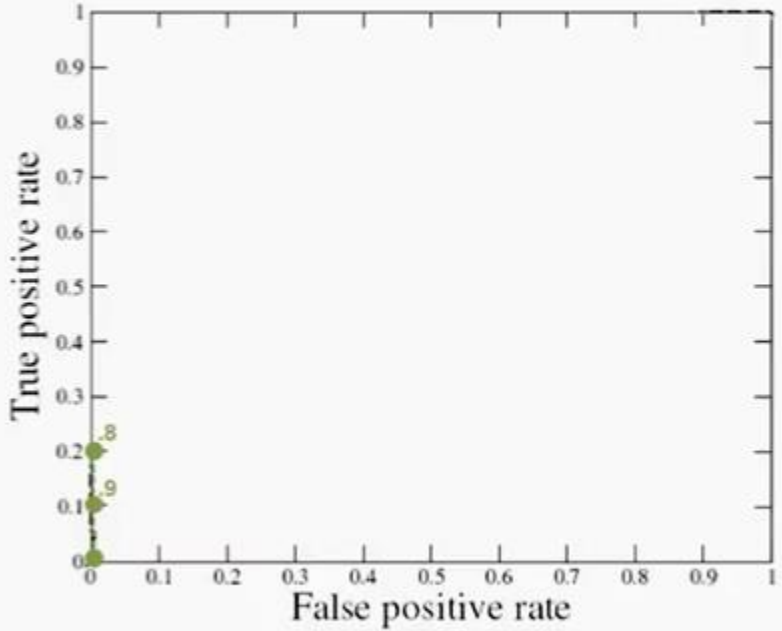
RECEIVER OPERATING CHARACTERISTIC (ROC)



#p = 10, #n=10

Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	p	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

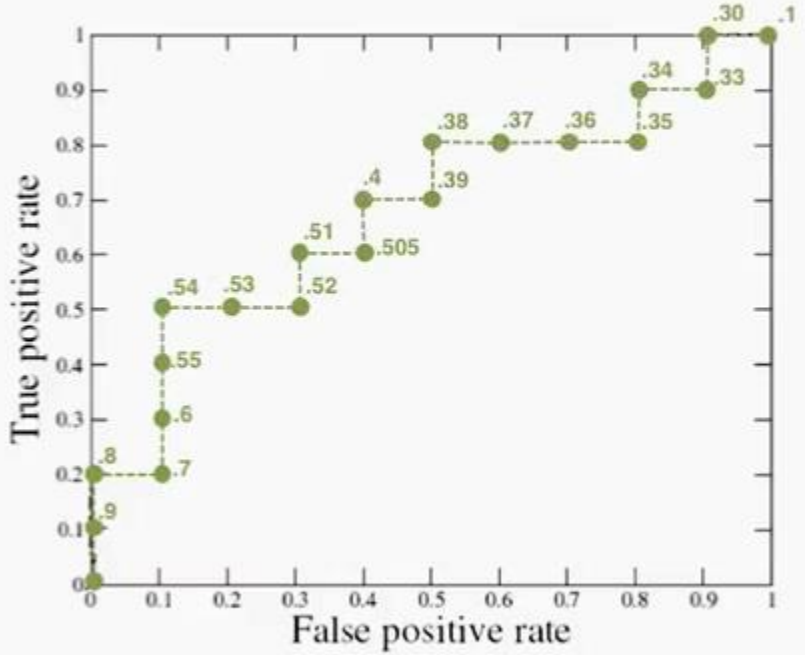
RECEIVER OPERATING CHARACTERISTIC (ROC)



#p = 10, #n=10

Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	p	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

RECEIVER OPERATING CHARACTERISTIC (ROC)

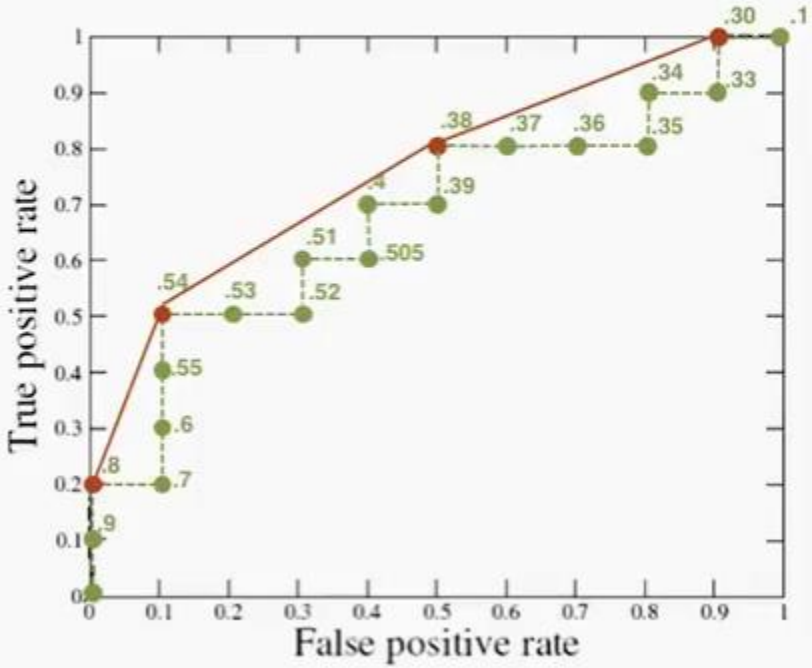


#p = 10, #n=10

Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	p	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

CLASSIFICATION METRIC: ROC QUIZ

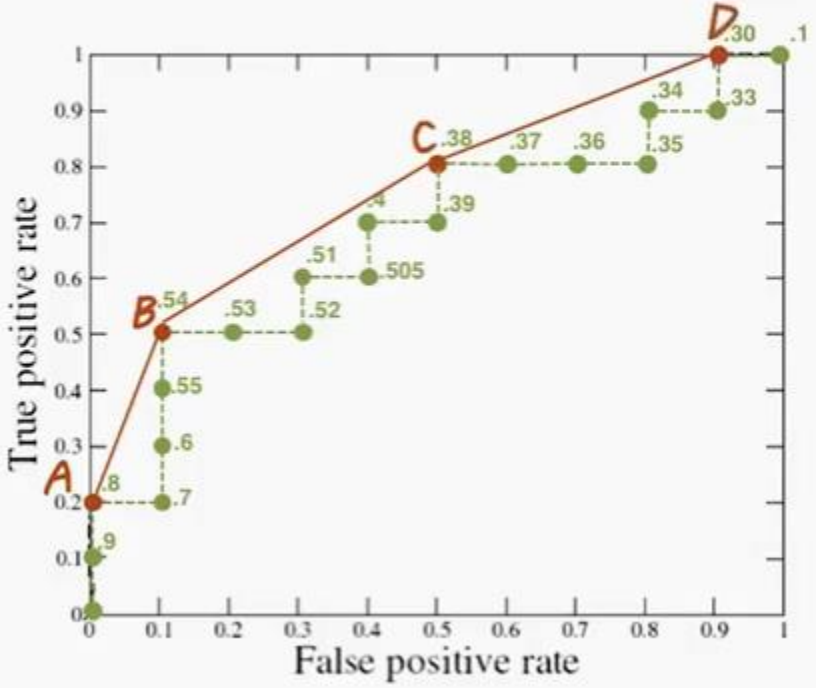
Which of the following would be a good threshold for this classifier?



Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	p	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

CLASSIFICATION METRIC: ROC QUIZ

Which of the following would be a good threshold for this classifier?



- ○ A

Low false positive rate

- ○ B
- ○ C

High true positive rate

- ○ D

Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	p	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

Answer: the optimal classification threshold may vary different according to preference

6. Regression Metrics MAE, MSE

REGRESSION METRICS: MAE, MSE

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i}|y_i - \hat{y}_i|$$

Difference between the prediction and the ground truth value

MAE is more robust against the outliers.

It's harder to work with because this absolute value is not differentiable.

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i}(y_i - \hat{y}_i)^2$$

The most popular regression metrics

Average of the squared error between the prediction and the ground truth value

MSE is easier to work with because the derivative of the square term is linear.

MSE will greatly affected by outliers because of the square term.

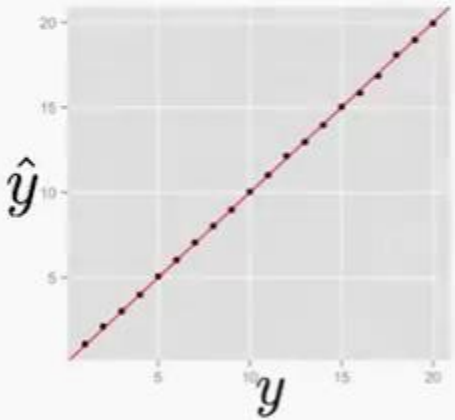
REGRESSION METRICS: MAE, MSE

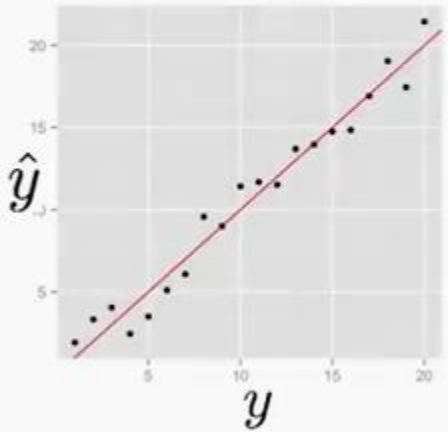
Mean Absolute Error (MAE)

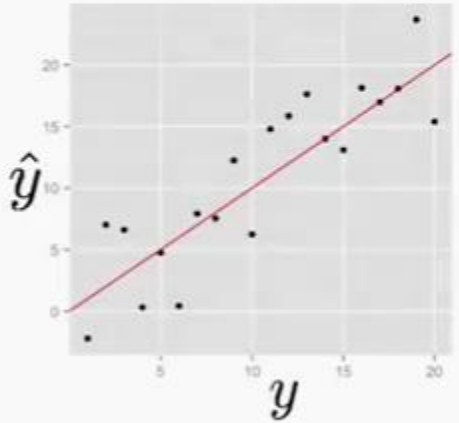
$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$







Y: prediction

X: ground truth value

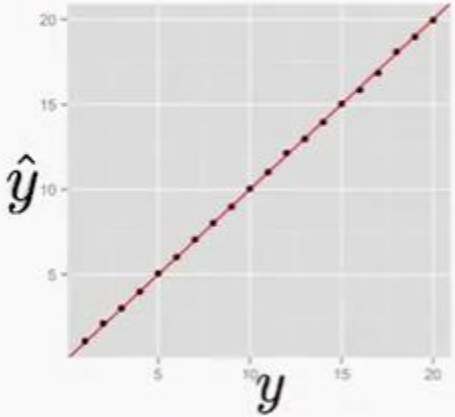
REGRESSION METRICS: MAE, MSE

Mean Absolute Error (MAE)

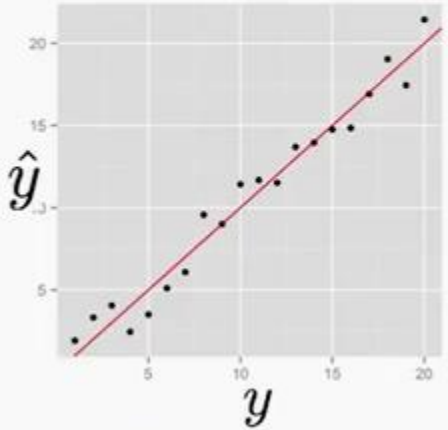
$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

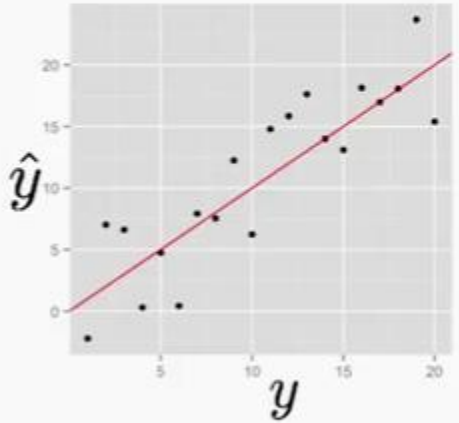
$$MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$



MAE = 0.0837
MSE = 0.0129



MAE = 0.7804
MSE = 1.1883



MAE = 3.4328
MSE = 18.6435

MSE grows a lot faster!

7. R2

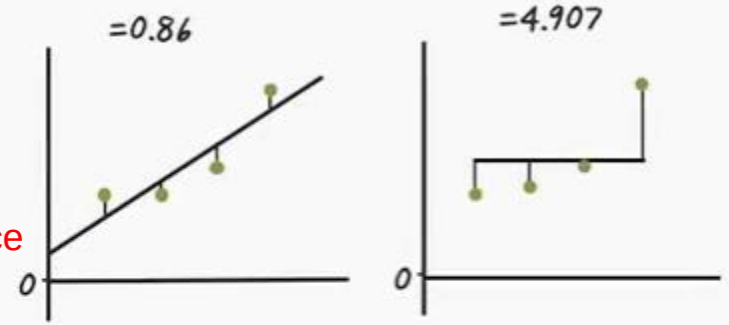
REGRESSION METRICS: R^2

Coefficient of determination R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

MSE

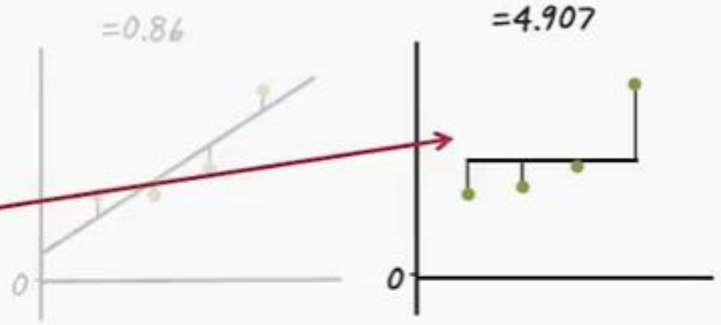
Variance



REGRESSION METRICS: R^2

Coefficient of determination R^2

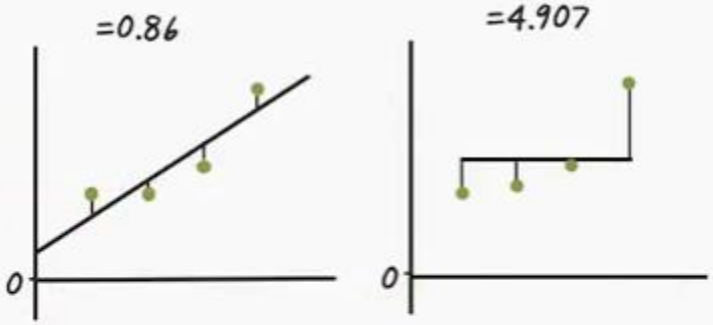
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$



REGRESSION METRICS: R^2

Coefficient of determination R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$



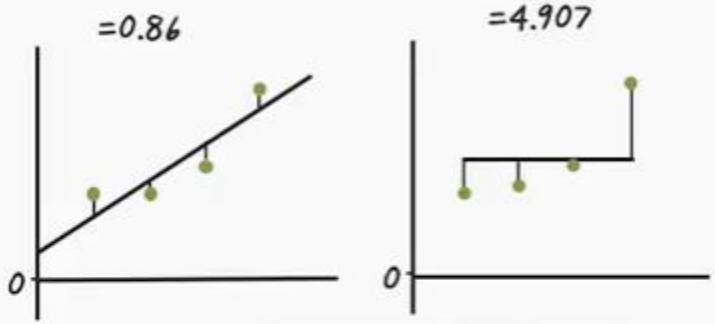
=0.82

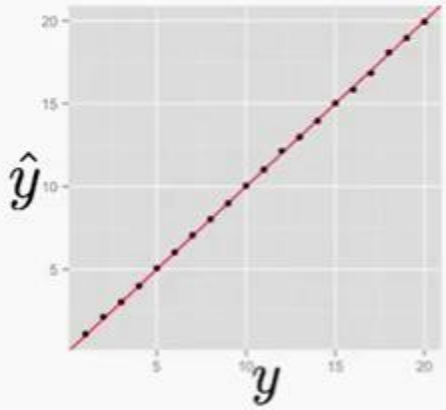
1: The regression fits perfectly for the data.
0: The line doesn't fit data at all.

REGRESSION METRICS: R^2

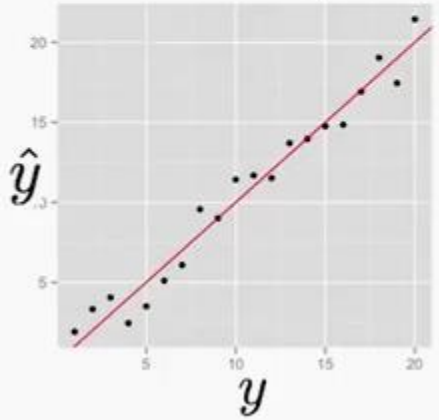
Coefficient of determination R^2

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

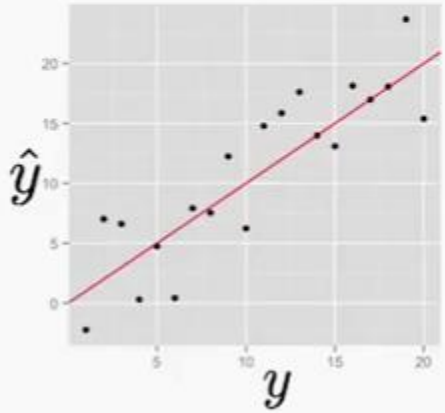




R^2 = 0.9997



R^2 = 0.7803



R^2 = 0.7404