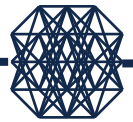
3. Machine Learning Basics

Prof. Jimeng Sun

ILLINOIS

Evaluation metrics



- Metrics for classification tasks
- Metrics for regression tasks
- Metrics for clustering tasks
- Evaluation strategy

Performance Metrics: Classification

		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

ILLINOIS

Performance Metrics: Classification

		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

Prediction	Prediction Outcome Positive
	Prediction Outcome Negative

ILLINOIS

Performance Metrics: Classification

		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

ILLINOIS

Performance Metrics: Classification

		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

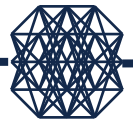
Performance Metrics: Classification

		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)

ILLINOIS

Performance Metrics: Classification



		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)

ILLINOIS

Performance Metrics: Classification

		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

Performance Metrics: Classification

		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

ILLINOIS

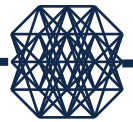
Confusion Matrix Quiz

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

Please fill in the missing numbers.

ILLINOIS

Confusion Matrix Quiz

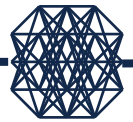


		Ground Truth	
	TOTAL POPULATION **1000**	Condition Positive **65**	Condition Negative 935
Prediction	Prediction Outcome Positive 155	True Positive **55**	False Positive 100
	Prediction Outcome Negative **845**	False Negative 10	True Negative **835**

Please fill in the missing numbers.

ILLINOIS

Performance Metrics: Classification



		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 = $\frac{\text{True negative}}{\text{Condition negative}}$

ILLINOIS

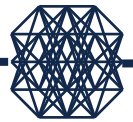
Performance Metrics: Classification

		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) = $\frac{\text{True positive}}{\text{Condition positive}}$

ILLINOIS

PERFORMANCE METRICS: CLASSIFICATION



		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 = $\frac{\text{True negative}}{\text{Condition negative}}$

False Negative Rate

$$= \frac{\text{False negative}}{\text{Condition positive}}$$

ILLINOIS

Performance Metrics: Classification

		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 = $\frac{\text{True negative}}{\text{Condition negative}}$

False Positive Rate = $\frac{\text{False Positive}}{\text{Condition negative}}$

ILLINOIS

Performance Metrics: Classification

		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) = $\frac{\text{True negative}}{\text{Condition negative}}$

ILLINOIS

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.

ILLINOIS

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 **85%**	False Positive Rate **11%**
		False Negative Rate **15%**	True Negative Rate **89%**

Please fill in the missing numbers.

ILLINOIS

Performance Metrics: Classification

		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 $= \frac{\text{True positive} + \text{True negative}}{\text{Total population}}$		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 $= \frac{\text{True negative}}{\text{Condition negative}}$	

Prevalence $= \frac{\text{Condition Positive}}{\text{Total population}}$

ILLINOIS

Performance Metrics: Classification

		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 = $\frac{\text{True positive} + \text{True negative}}{\text{Total population}}$		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 = $\frac{\text{True negative}}{\text{Condition negative}}$	

Positive Predictive Value (Precision)

$$= \frac{\text{True Positive}}{\text{Test outcome positive}}$$

ILLINOIS

Performance Metrics: Classification

		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 $= \frac{\text{True Positive}}{\text{Prediction outcome positive}}$
	Prediction Outcome Negative	False Negative (Type II error)	True Negative	False Omission Rate $= \frac{\text{False negative}}{\text{Prediction outcome negative}}$
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 $= \frac{\text{True negative}}{\text{Condition negative}}$	

False Discovery Rate $= \frac{\text{False Positive}}{\text{Prediction outcome positive}}$

False Discovery Rate $= \frac{\text{False Positive}}{\text{Prediction outcome positive}}$

ILLINOIS

Performance Metrics: Classification

		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 $= \frac{\text{True Positive}}{\text{Prediction outcome positive}}$
	Prediction Outcome Negative	False Negative (Type II error)	True Negative	False Omission Rate $= \frac{\text{False negative}}{\text{Prediction outcome negative}}$
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 $= \frac{\text{True negative}}{\text{Condition negative}}$	

		False Discovery Rate $= \frac{\text{False Positive}}{\text{Prediction outcome positive}}$
		Negative Predictive Value $= \frac{\text{True negative}}{\text{Prediction outcome negative}}$

Negative Predictive Value $= \frac{\text{True negative}}{\text{Prediction outcome negative}}$

ILLINOIS

Performance Metrics: Classification

		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 $= \frac{\text{True positive} + \text{True negative}}{\text{Total population}}$		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 $= \frac{\text{True negative}}{\text{Condition negative}}$		

False Omission Rate $= \frac{\text{False negative}}{\text{Prediction outcome negative}}$

ILLINOIS

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	Negative Predictive Value
Accuracy **89%**		True Positive Rate 85%	False Positive Rate 11%		
		False Negative Rate 15%	True Negative Rate 89%		

Please fill in the missing numbers.

ILLINOIS

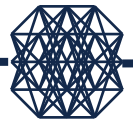
Predictive Metrics Quiz

		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 fill in the missing numbers.

ILLINOIS

F_1 Score

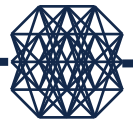


		Ground Truth		
	TOTAL POPULATION	Condition Positive	Condition Negative	**Prevalence** $= \frac{\text{Condition Positive}}{\text{Total population}}$
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)	Positive Predictive Value $= \frac{\text{True Positive}}{\text{Prediction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

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

ILLINOIS

F_1 Quiz

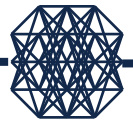


		Ground Truth			
	TOTAL POPULATION **1000**	Condition Positive 65	Condition Negative 935	**Prevalence** **7%**	
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100	Positive Predictive Value 35%	False Discovery Rate 65%
	Prediction Outcome Negative 845	False Negative 10	True 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.

ILLINOIS

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

ILLINOIS

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
F_1 = 0.81
Accuracy = 0.82

ILLINOIS

Classifier Quiz 2

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
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

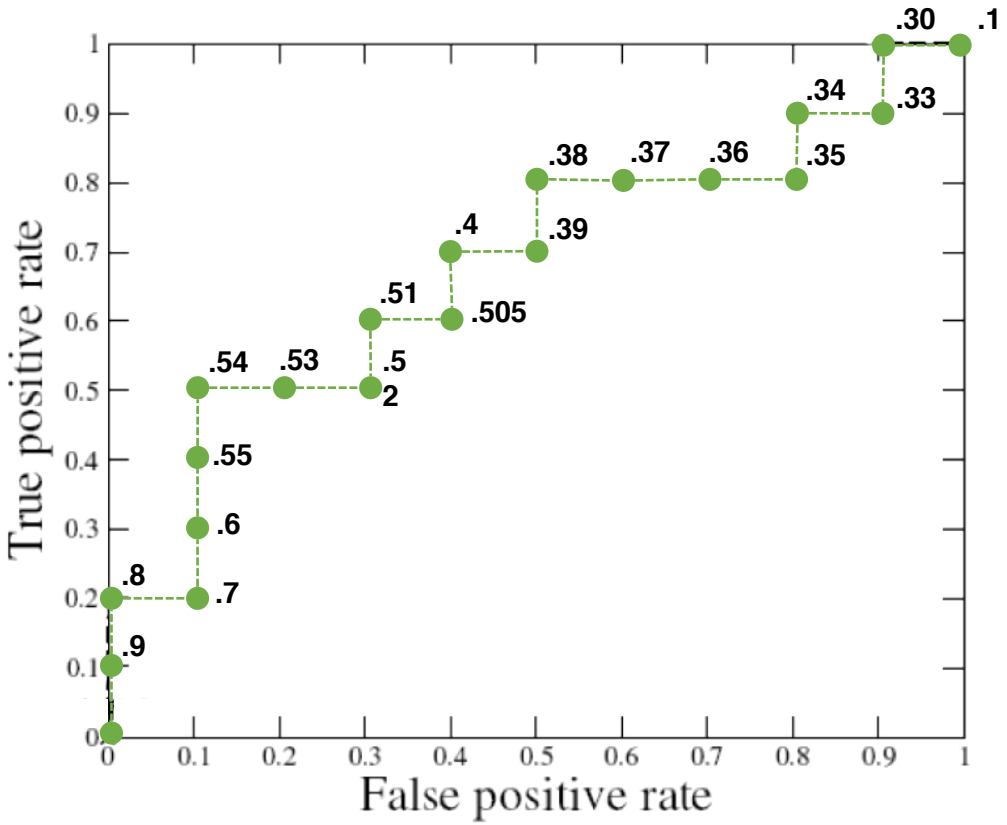
✓ C

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

PPV = 0.21 → 0.86
F_1 = 0.22 → 0.81
ACC = 0.18 → 0.82

ILLINOIS

Receiver Operating Characteristic (ROC)



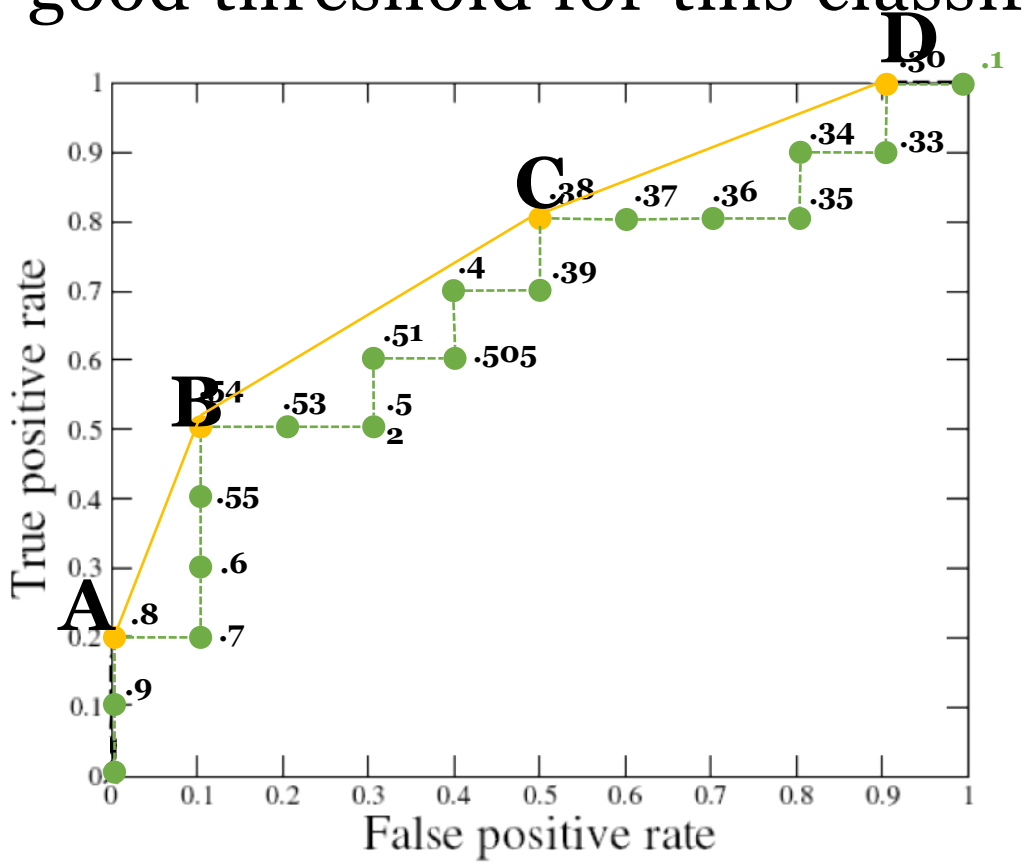
#p = 10, #n=10

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

ILLINOIS

Classification Metric: Roc Quiz

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



- ✓ A
- ✓ B
- ✓ C
- ✓ 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

ILLINOIS

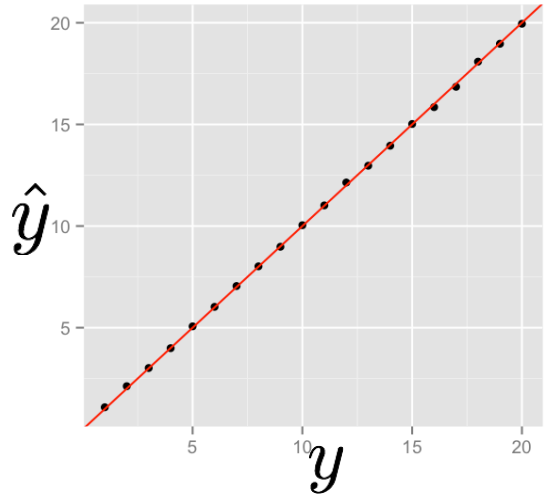
Regression Metrics: MAE, MSE

Mean Absolute Error (MAE)

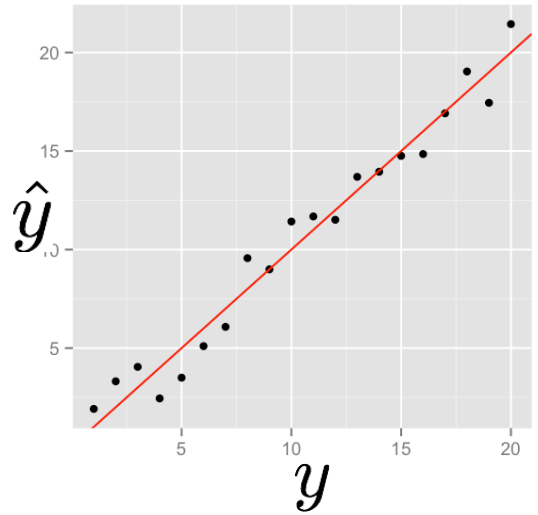
$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

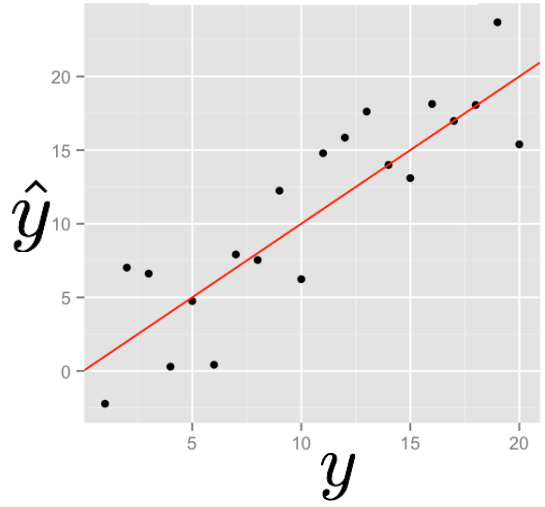
$$MSE = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$$



MAE = 0.0837
MSE = 0.0129



MAE = 0.7804
MSE = 1.1883



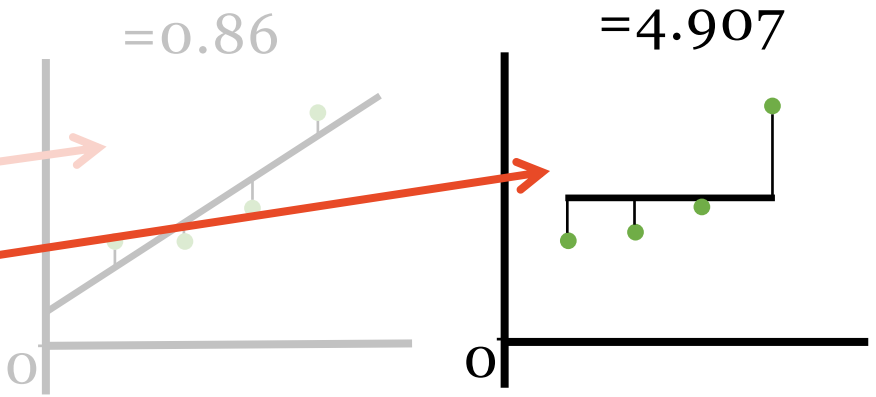
MAE = 3.4328
MSE = 18.6435

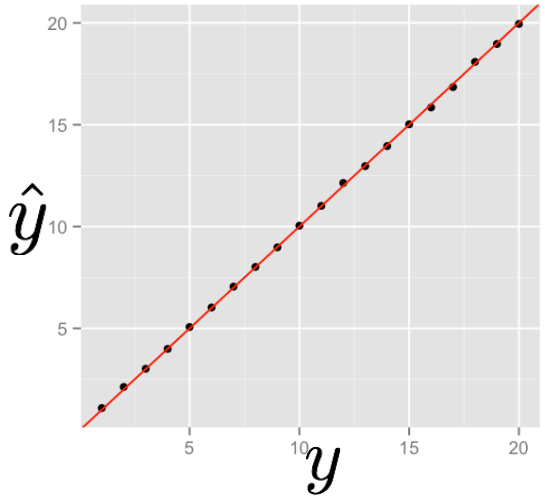
ILLINOIS

Regression Metrics: R²

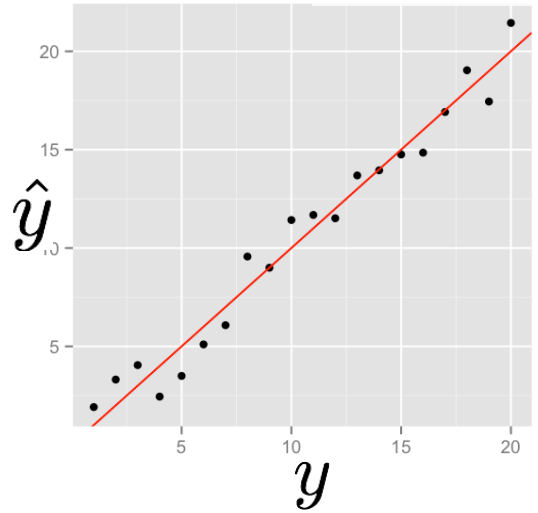
Coefficient of determination R²

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

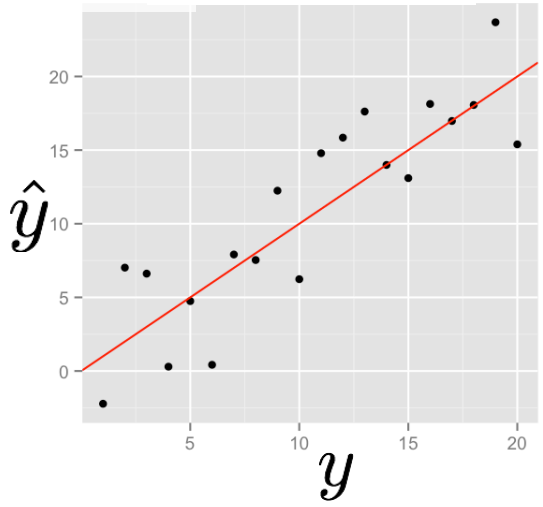




R² =0.9997



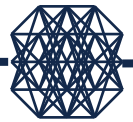
R² =0.7803

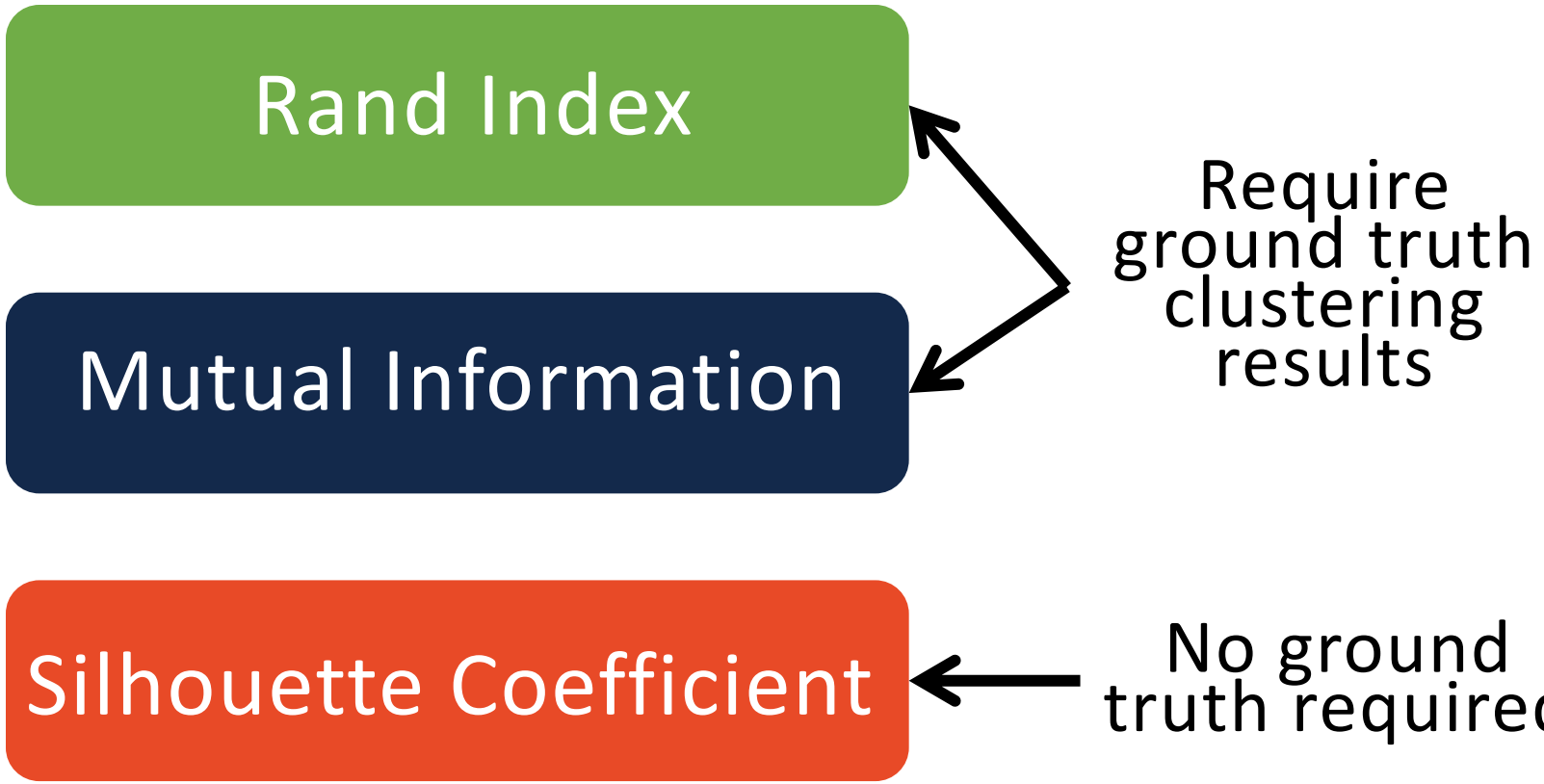


R² =0.7404

ILLINOIS

Clustering Evaluation Metrics





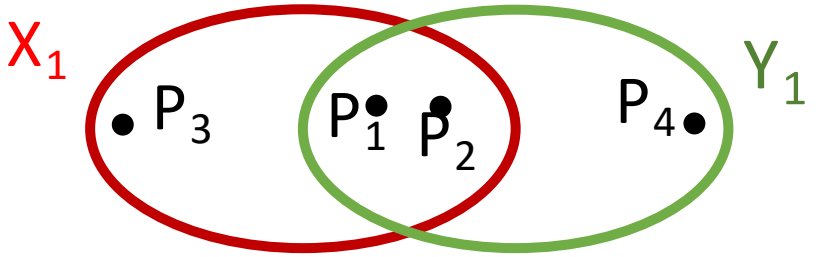
ILLINOIS

Rand Index (RI)

n data points

X: clustering assignments

Y: ground truth



a: # of pairs that belong to same cluster in X and Y

b: # of pairs that belong to different clusters in X and Y

$$\text{RI} = \frac{a+b}{\#\ of\ pairs}$$

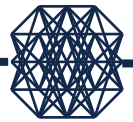
$\#\ of\ pairs = \frac{n(n-1)}{2}$

RI = 0 bad clustering

RI = 1 perfect clustering

ILLINOIS

Mutual Information



X = $\{x_1, x_2, \ldots, x_k\}$ clustering assignments

Y = $\{y_1, y_2, \ldots, y_r\}$ ground truth

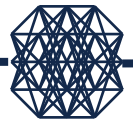
Entropy $\mathrm{H}(X) = \sum_{x \in X} p(x) \log p\,(x)$

deterministic random

$[0, 1]$

ILLINOIS

Mutual Information



X = $\{x_1, x_2, \ldots, x_k\}$ clustering assignments

Y = $\{y_1, y_2, \ldots, y_r\}$ ground truth

deterministic random

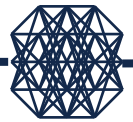
Entropy $\mathrm{H}(X) = \sum_{x \in X} p(x) \log p(x)$ $[0, 1]$

$$\mathrm{MI}(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

$$\text{Normalized\_MI}(X, Y) = \frac{MI(X, Y)}{\sqrt{H(X)H(Y)}}$$

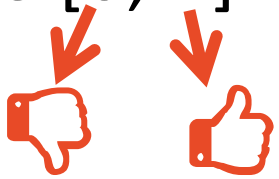
ILLINOIS

Summary Of Rand Index and Mutual Information



PROS

- Bounded range [0, 1]



- No assumption on cluster shapes

CONS

- Require ground truth

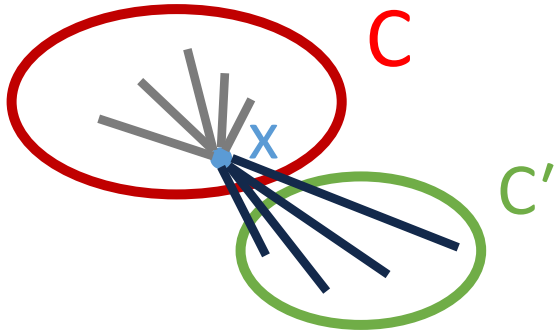
ILLINOIS

Silhouette Coefficient

x: data point

C: cluster containing x

C′: next nearest cluster to x



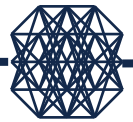
$$a = \frac{1}{|C|}\sum_{y \in C} \| x - y \|$$

$$b = \frac{1}{|C'|}\sum_{y \in C'} \| x - z \|$$

$$s(x) = \frac{b - a}{\max(a, b)}$$

ILLINOIS

Silhouette Coefficient



PROS

- Bounded between -1 and 1

Bad clustering assignment

Good clustering assignment

CONS

- Assume spherical clusters

ILLINOIS