Link Prediction & ROC / AUC

Seunghan Lee

2020-02-03

Contents

1. Link Prediction

node2vec

1. Link Prediction

4.7 Link prediction

In link prediction, we are given a network with a certain fraction of edges removed, and we would like to predict these missing edges. We generate the labeled dataset of edges as follows: To obtain positive examples, we remove 50% of edges chosen randomly from the network while ensuring that the residual network obtained after the edge removals is connected, and to generate negative examples, we randomly sample an equal number of node pairs from the network which have no edge connecting them.

(node2vec: Scalable Feature Learning for Networks)

Predict whether there is is a **connection(link) between two nodes**!

Binary Classification Problem!

Take one edge, and classify either into "**connected**" or "**un-connected**"

1. Link Prediction

4.7 Link prediction

In link prediction, we are given a network with a certain fraction of edges removed, and we would like to predict these missing edges. We generate the labeled dataset of edges as follows: To obtain positive examples, we remove 50% of edges chosen randomly from the network while ensuring that the residual network obtained after the edge removals is connected, and to generate negative examples, we randomly sample an equal number of node pairs from the network which have no edge connecting them.

(node2vec: Scalable Feature Learning for Networks)

Predict whether there is is a **connection(link) between two nodes**!

Binary Classification Problem!

Take one edge, and classify either into "**connected**" or "**un-connected**"

[Q] (in the case of weighted graph)
Can't it be a regression problem,
of predicting the "strength of the connection"(= weight) ?

Link Prediction Algorithm

Network with edges & nodes

Masking (Remove x% of edges)

Node2vec paper : mask 50%

[INPUT] Network with (100-x)% of edges

(removed) Network with x% of edges

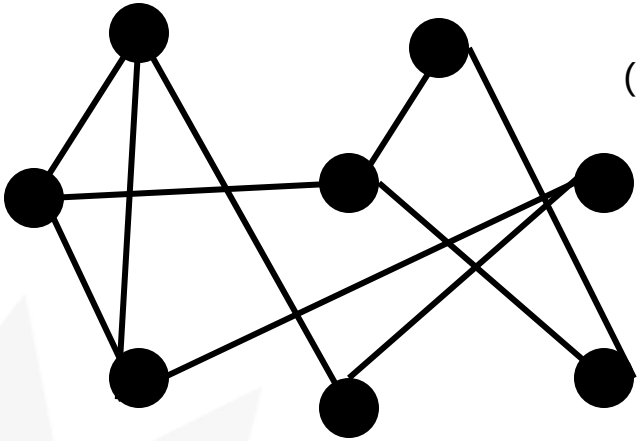
train

test

Classification Model

(whether the edge is connected(1) or not(0))

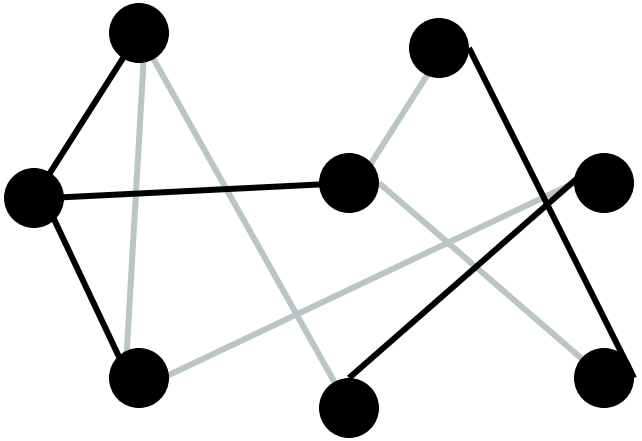
10 edges



(remove 5 edges (50%))

MASKING

5 edges left

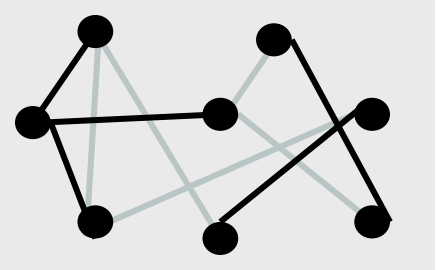


Remove x% of connected edges (grey edge)

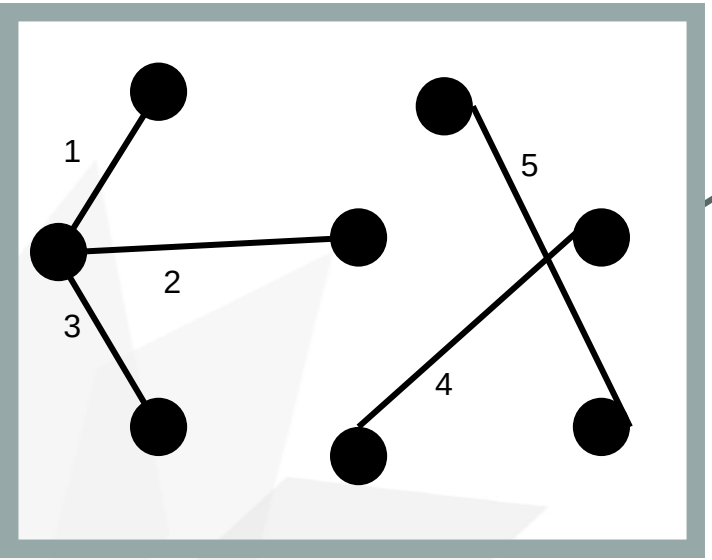
Train model only with the (100-x)% remaining edges (black edge)

HOW?

Train model only with the (100-x)% remaining edges (black edge)

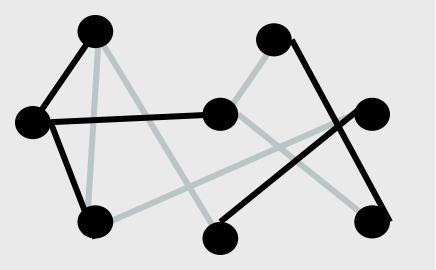


Black : remained
Grey : removed

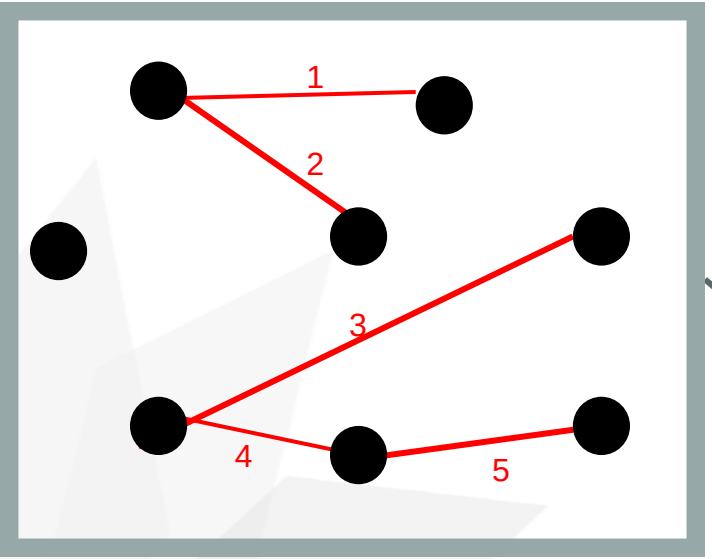


Positive Sample

(remaining 50% edges)

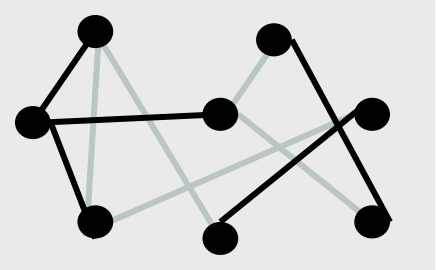


Black : remained
Grey : removed



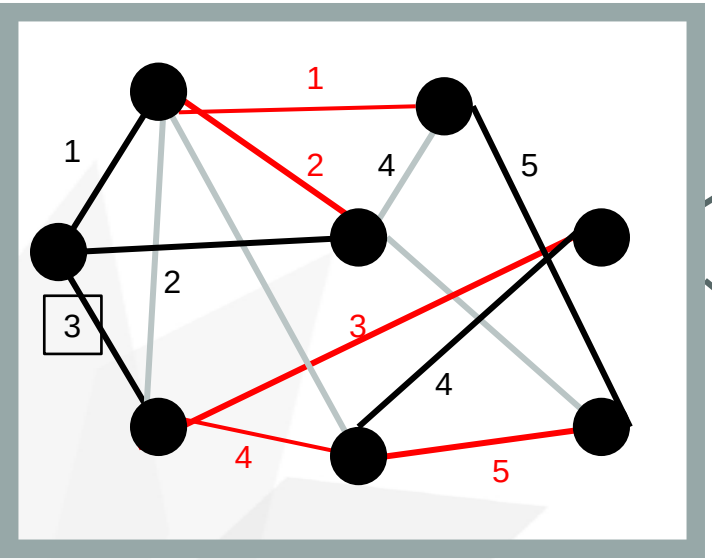
Negative Sample

(sampling UNCONNECTED nodes)



Black : remained
Grey : removed

Same number of pairs (pos = neg)

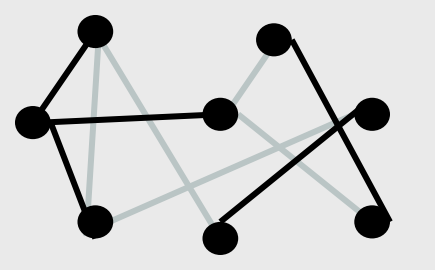


Positive Sample

(remaining 50% edges)

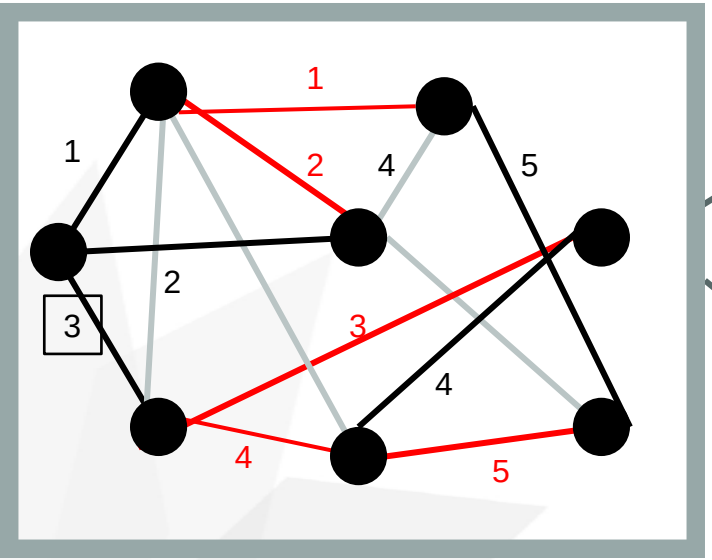
Negative Sample

(sampling UNCONNECTED nodes)



Black : remained
Grey : removed

Same number of pairs (pos = neg)



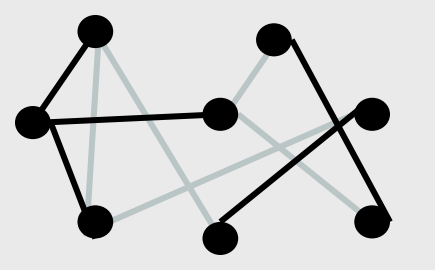
Positive Sample

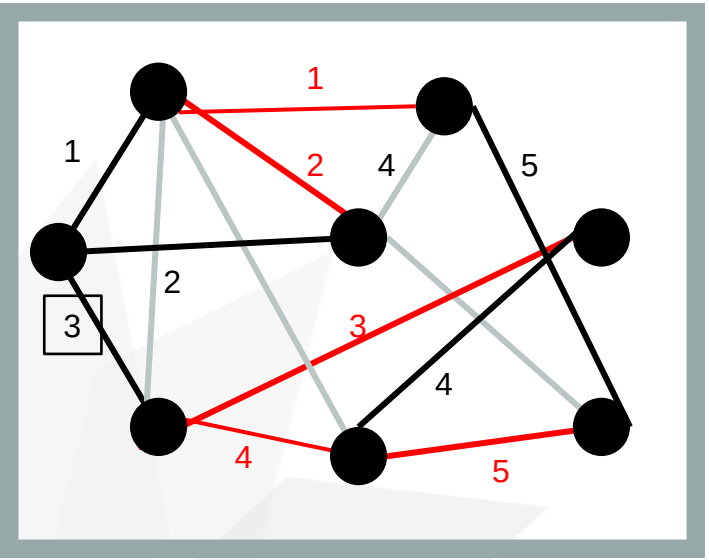
(remaining 50% edges)

Negative Sample

(sampling UNCONNECTED nodes)

The edges that are removed in the process of masking, (which are actually connected) can also be selected as negative samples!

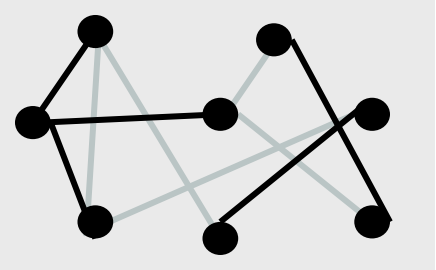




Make a classification model

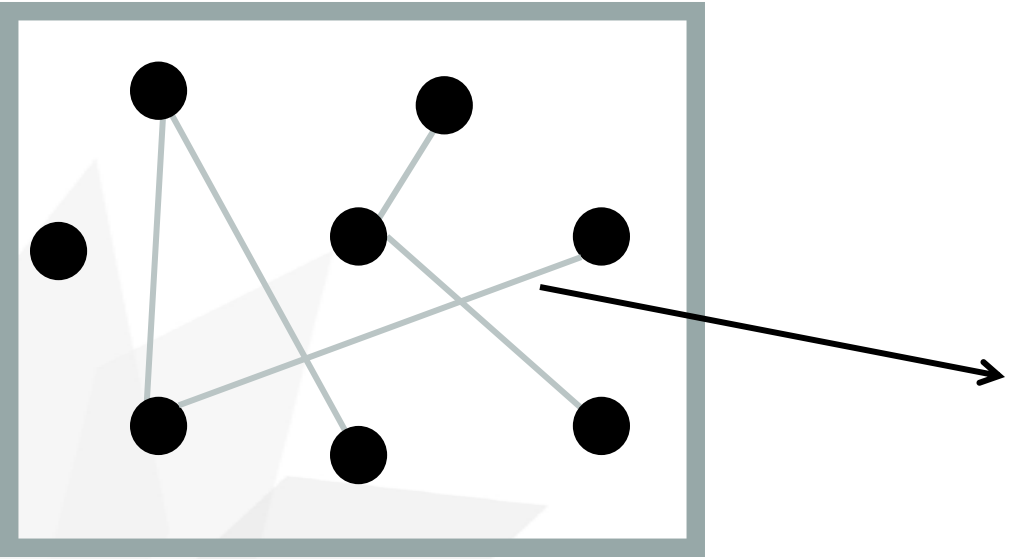
(with positive & negative sample)

*** equal number of pairs**



TEST with Masked Edges!

(Predict these missing edges!)



[ACTUAL] **1**

[PREDICTION] **?**

Metric for evaluating model performance in Link Prediction

1. ROC curve

2. AUC (Area Under Curve)

Before ROC & AUC…

		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

https://miro.medium.com/max/1194/0*wKaznIJzZF54b87B.jpg

[Confusion matrix]

Precision = TP / TP+FP

(Real Positive, among Predicted Positive)

Recall = TP / TP+FN

(Predicted Positive, among Actual Positive)

Specificity = TN / TN+FP

(Predicted Negative, among Actual Negative)

Before ROC & AUC…

		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

https://miro.medium.com/max/1194/0*wKaznIJzZF54b87B.jpg

[Confusion matrix]

Precision = TP / TP+FP

(Real Positive, among Predicted Positive)

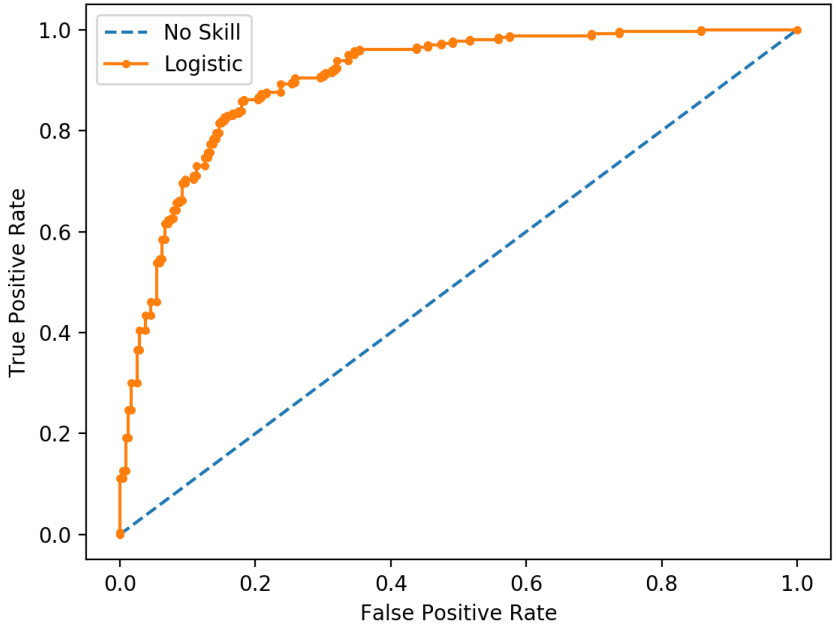
Recall = TP / TP+FN

(Predicted Positive, among Actual Positive)

Specificity = TN / TN+FP

(Predicted Negative, among Actual Negative)

1. ROC Curve



(= FP / FP+TN)

X axis : False Positive Rate

(= 1 – Specificity)

The bigger, the better

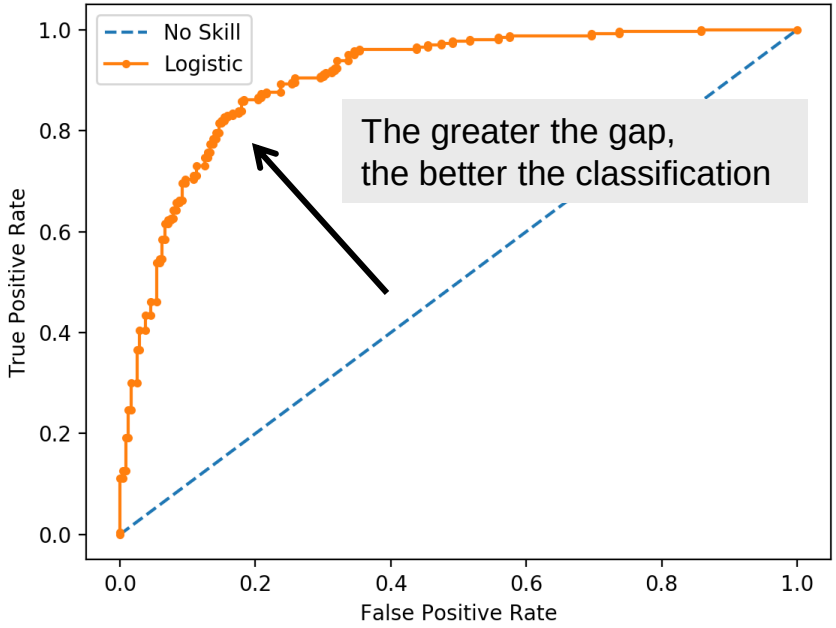
(= TP / TP+FN)

Y axis : True Positive Rate

(= Recall)

The bigger, the better

1. ROC Curve



(= FP / FP+TN)

X axis : False Positive Rate

(= 1 – Specificity)

The bigger, the better

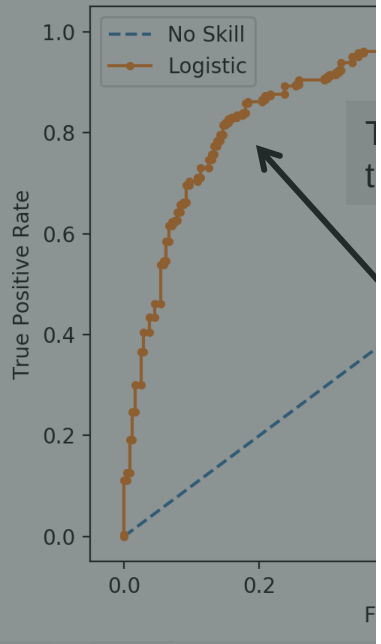
(= TP / TP+FN)

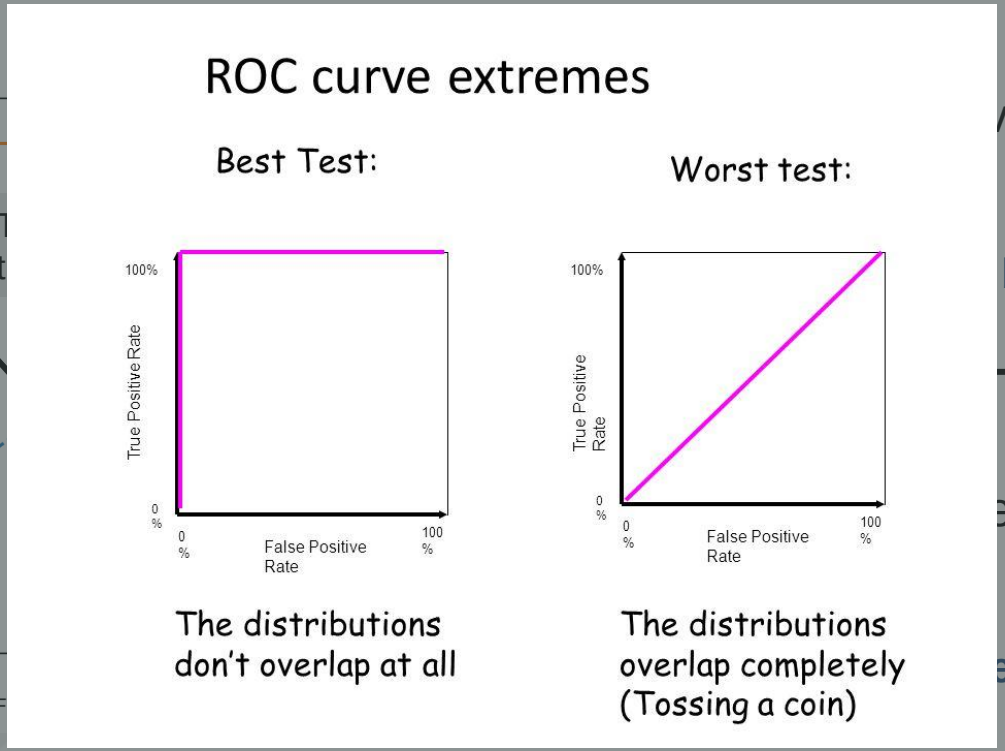
Y axis : True Positive Rate

(= Recall)

The bigger, the better

1. ROC Curve





(= FP / FP+TN)

ve Rate

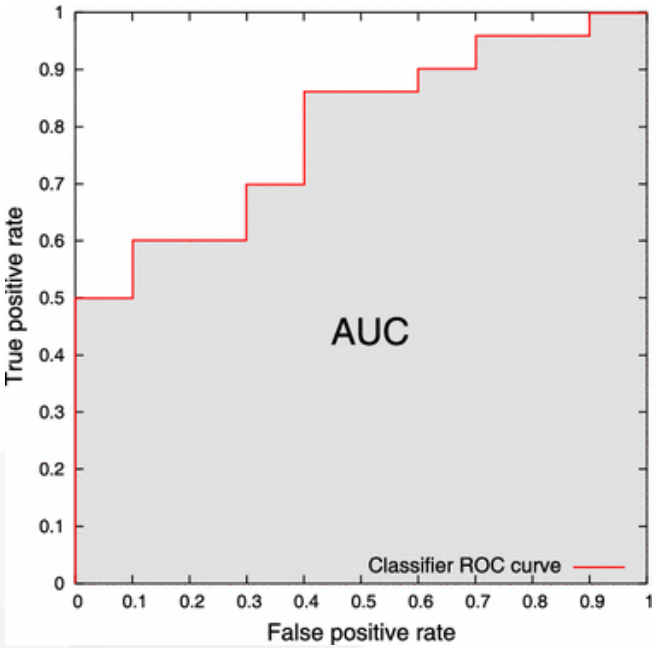
better

(= TP / TP+FN)

e Rate

er

2. AUC (Area Under Curve)



Area below the ROC curve!

(range : 0.5 ~ 1)

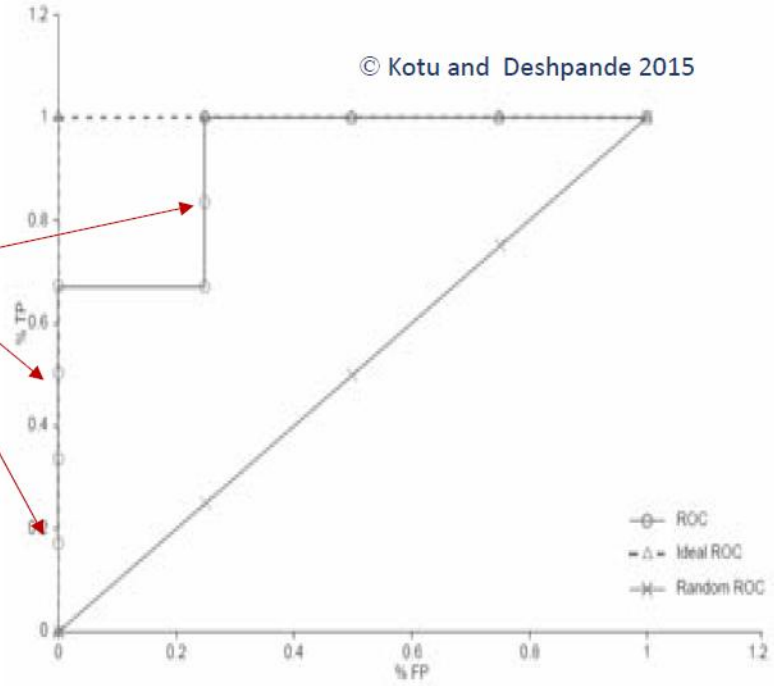
* 0.5 : tossing a coin

https://media.springernature.com/original/springer-static/image/art%3A10.1007%2Fs10115-017-1022-8/MediaObjects/10115_2017_1022_Fig1_HTML.gif

(optional) How to draw an ROC curve?

[표 8.3] ROC 곡선을 생성하기 위한 분류기의 성능 데이터

Actual Class	Predicted Class	Confidence of "response"	Type?	Number of TP	Number of FP	Fraction of FP	Fraction of TP
response	response	0.902	TP	1	0	0	0.167
response	response	0.896	TP	2	0	0	0.333
response	response	0.834	TP	3	0	0	0.500
response	response	0.741	TP	4	0	0	0.667
no response	response	0.686	FP	4	1	0.25	0.667
response	response	0.616	TP	5	1	0.25	0.833
response	response	0.609	TP	6	1	0.25	1
no response	response	0.576	FP	6	2	0.5	1
no response	response	0.542	FP	6	3	0.75	1
no response	response	0.530	FP	6	4	1	1
no response	no response	0.440	TN	6	4	1	1
no response	no response	0.428	TN	6	4	1	1
no response	no response	0.393	TN	6	4	1	1
no response	no response	0.313	TN	6	4	1	1
no response	no response	0.298	TN	6	4	1	1
no response	no response	0.260	TN	6	4	1	1
no response	no response	0.248	TN	6	4	1	1
no response	no response	0.247	TN	6	4	1	1
no response	no response	0.241	TN	6	4	1	1
no response	no response	0.116	TN	6	4	1	1



[그림 8.1] 표 8.3의 예제 데이터에 대한 무작위 분류기와 이상적인 분류기의 ROC 곡선 비교

SUMMARY

1. Link Prediction

- Remove some edges with masking, for testing in the future!
- Make binary classification model with “remained positive edges” & “sampled negative edges”
- Predict the removed edges (for testing)

2. ROC & AUC

- Metric for evaluating binary classification model performance
- ROC : checking how well a classification is done visually (by comparing FPR & TPR)
- AUC : checking how well a classification is done numerically, (by finding the area under the ROC Curve)
- (both ROC & AUC) the BIGGER, the BETTER

Thank you