Evaluating predict

Monday, April 24, 2017 3:25 PM

Have test set $\{x_i, y_i\}$, want to evaluate predict method $\hat{y}_i = \begin{cases} 1, & s_i > \tau \\ 0, & s_i \leq \tau \end{cases}$

Confusion Matrix

	Predict 1	Predict 0
Actual 1	True Positive (TP)	False Negative (FN)
Actual 0	False Positive (FP)	True Negative (TN)

Predict

1. Calculate score s_i for each i.
2. Order scores : $s_{a_1} \geq s_{a_2} \geq s_{a_3} \geq \dots \geq s_{a_n}$ where $a_1, \dots, a_n$ is permutation of $\{1, \dots, n\}$.
3. Predict $\hat{y}_{a_1}, \dots, \hat{y}_{a_T} = 1$ and $\hat{y}_{a_{T+1}}, \dots, \hat{y}_{a_n} = 0$.

S :	3.1	3	2.8	2.6	2.3	2.1	1.1	1	.8	.6	.5	.2	
$\hat{y}$	1	1	1	1	1	1	0	0	0	0	0	0	$\tau = 6$
y	1	0	1	1	0	1	0	1	0	0	1	0	

Conf. TP = 4 | FN = 2

Conf. Matrix

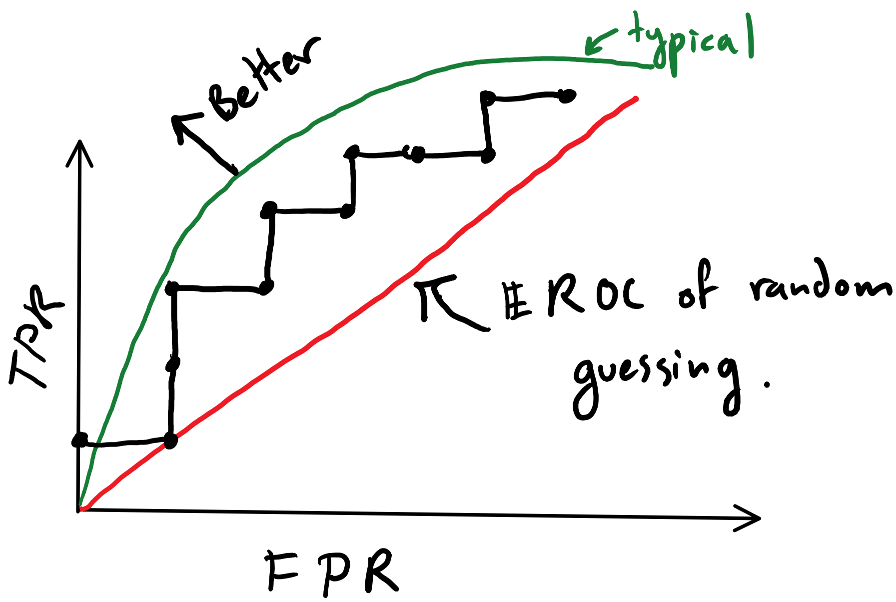
TP = 4	FN = 2
FP = 2	TN = 4

Reciever - Operating Characteristic (ROC)

True positive rate (TPR) (aka recall) $= \dfrac{TP}{\underbrace{TP + FN}_{\text{Actual Positive}}}$

False positive rate (FPR) $= \dfrac{FP}{\underbrace{FP + TN}_{\text{Actual Neg.}}}$

T	1	2	3	4	5	6	7	8
TPR	1/6	1/6	2/6	3/6	3/6	4/6	4/6	5/6
FPR	0/6	1/6	1/6	1/6	2/6	2/6	3/6	3/6



- Every element of confusion matrix is represented in ROC

Precision - Recall Curve

In recommendation systems, many 0's – few 1's.

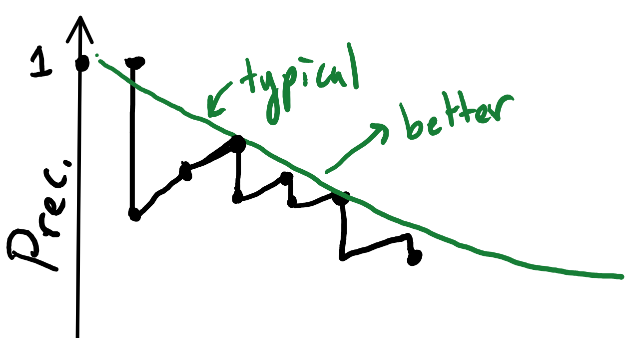
ex
- Link prediction in sparse graphs (friend recomm.)
- Search engine (most pages are not relevant)
- Document retrieval

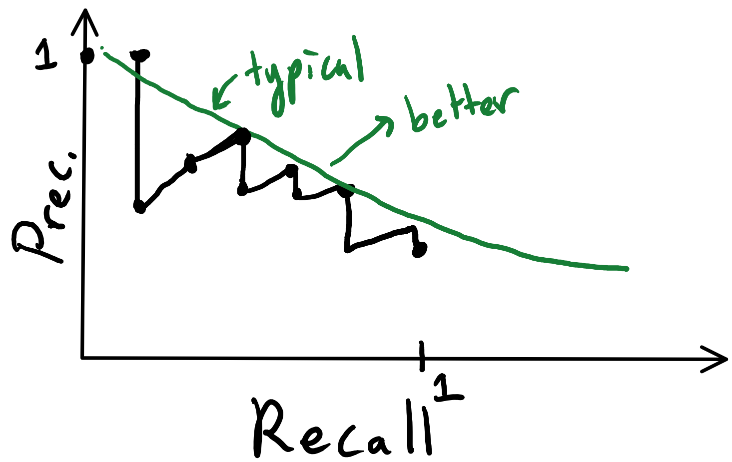
$$\text{Recall} = \text{TPR} = \frac{TP}{TP + FN}$$

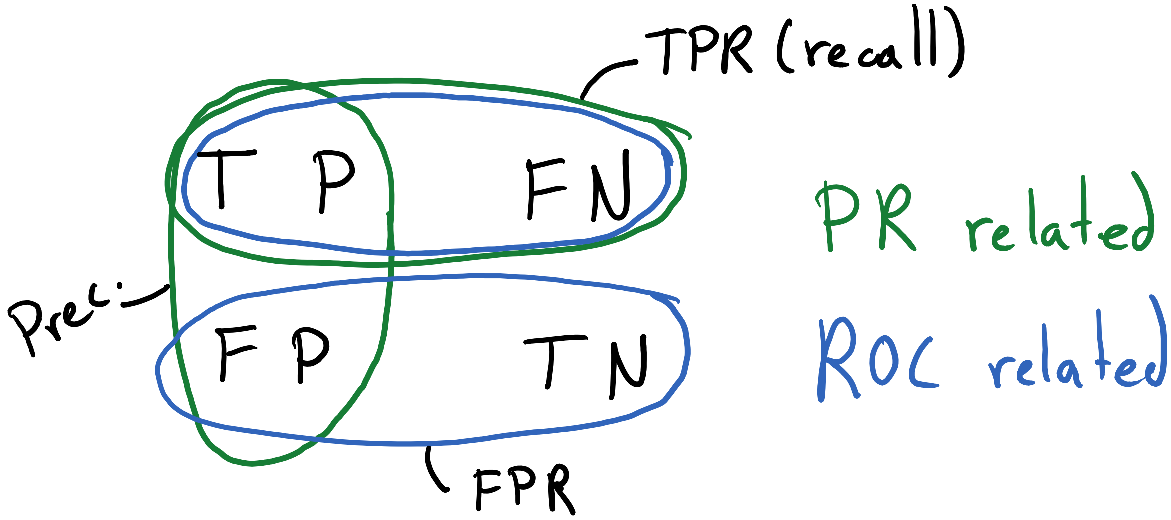
$$\text{Precision} = \frac{TP}{TP + FP}$$

Neither is sensative to increasing TN

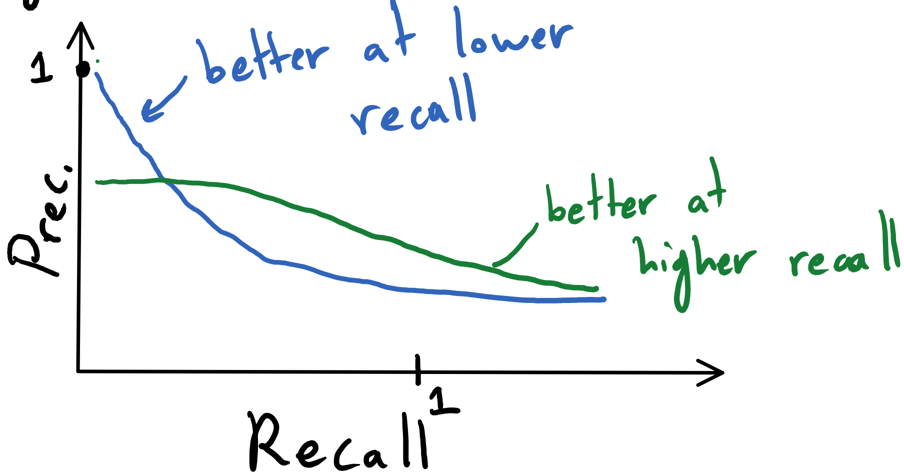
T	1	2	3	4	5	6	7	8	
Recall	1/6	1/6	2/6	3/6	3/6	4/6	4/6	5/6	…
Prec.	1	1/2	2/3	3/4	3/5	4/6	4/7	5/8	…







Comparing PR Curves

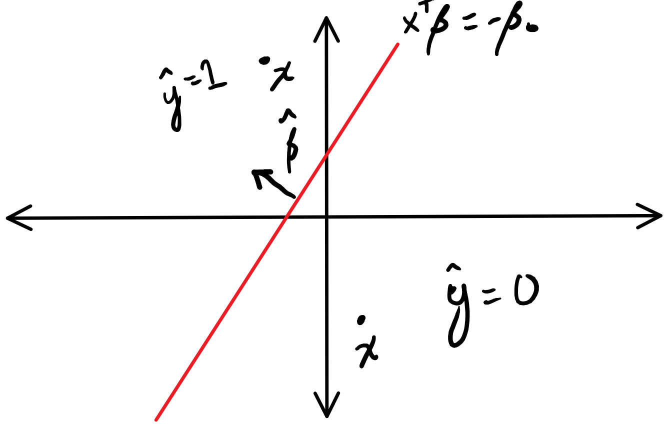


F1-score : $\dfrac{2 \text{ prec} \cdot \text{recall}}{\text{prec} + \text{recall}}$

Margin Based Methods

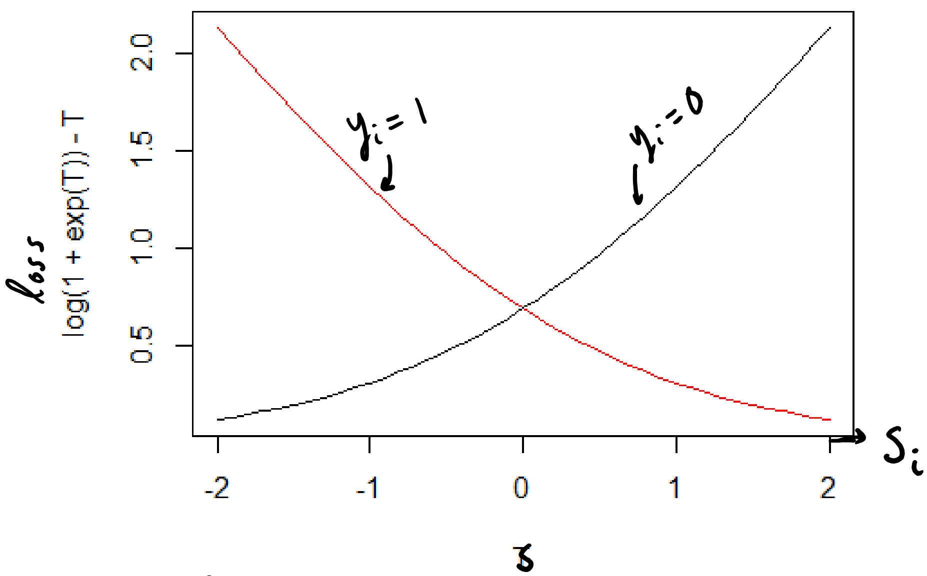
Tuesday, April 25, 2017 4:22 PM

$$\hat{y} = \begin{cases} 1, & \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \geq 0 \\ 0, & \text{otherwise} \end{cases}$$



Empirical Risk Minimization

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} loss\,(y_i, x_i, \beta)$$



$$S_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

0-1 loss $(y_i = 1)$

$$loss_{0\text{-}1} = \begin{cases} 1, & S_i < 0 \\ 0, & \text{otherwise} \end{cases}$$

logistic regression $(y_i = 1)$

$$\min \frac{1}{n} \sum_i loss_{0\text{-}1}(y_i, x_i, \beta)$$

is very hard to optimize!

Support vector machines

$$\text{loss} = \log\left(1+e^{-s_i}\right)$$

$$\left(\begin{array}{c}\text{for } y_i = 0 \text{ switch} \\ s_i \leftarrow -s_i\end{array}\right)$$

$$\text{loss} = \begin{cases} 1 - s_i & , \; s_i < 1 \\ 0 & , \; s_i > 1 \end{cases}$$