11. 評価方法

精度、再現率
順位付き評価
学習と評価
評価者の一致性の評価

by 中川裕志（東京大学）

教師あり学習の評価

- 予測値の決め方

  if $p(1|\mathbf{x}) \geq \theta_{th}$ then $\hat{y} = +1$ (正解) otherwise $\hat{y} = -1$(不正解)

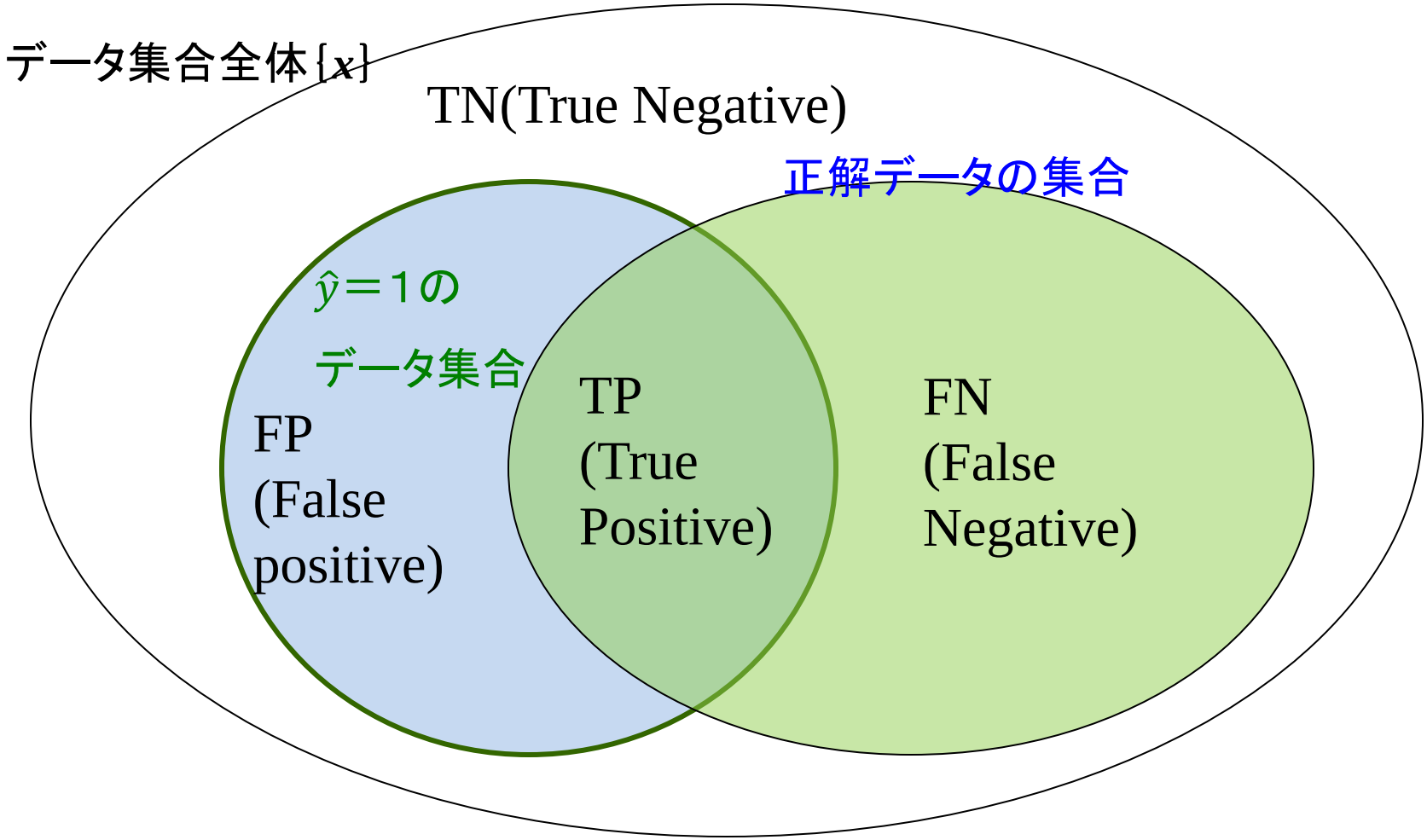
- 機械学習の結果の予測器によって

  $\boldsymbol{x}$が正解（１）である確率が閾値θ_{th}より大きければ予測値　$\hat{y}$＝＋１　小さければ　$\hat{y}$＝－１となる。

  - ここで、閾値への予測値の依存性に注意

一般的なデータ処理結果の状態

- 処理sで結果のデータ集合が得られた。しかし、結果の中には間違いもあるし、得られなかったデータの中にも正解がありうる。



性能評価尺度

- 再現率

$$recall = \frac{TP}{TP + FN}$$

- 適合率あるいは精度

$$precision = \frac{TP}{TP + FP}$$

- フォールアウト

$$fallout = \frac{TN}{TN + FP}$$

- 一般性

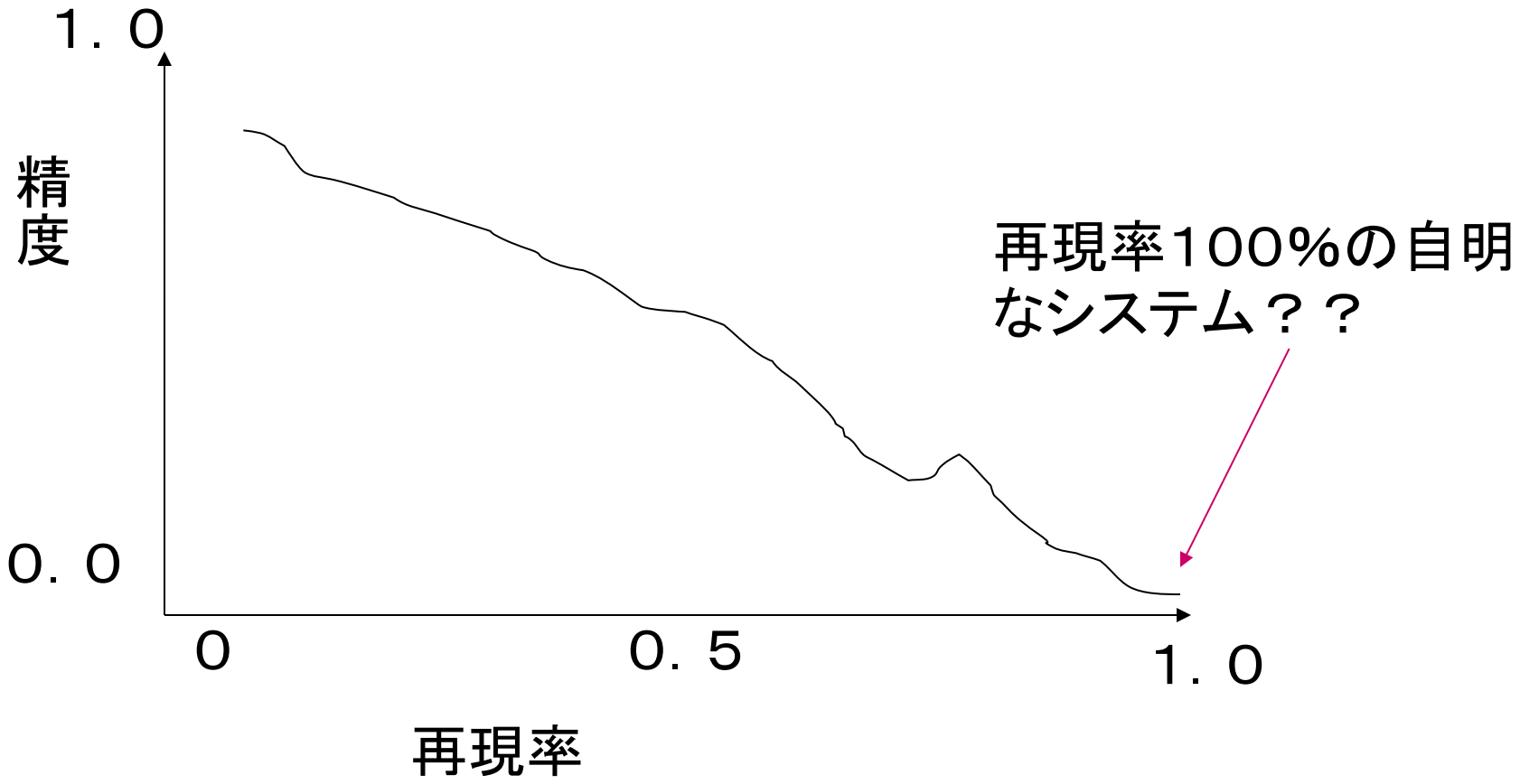
$$generarity = \frac{TP}{TP + TN + FP + FN}$$

- Accuracy
  or
  Rand Index

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

再現率　vs　精度

- よく使う評価の表現法



再現率　vs　精度に関連した尺度

- Break even point　　再現率と精度が一致する点
- 11点平均精度　再現率＝0.0 ，0.1，　0.2，….. 0.9，1.0 の11点における精度の平均値
- F値　　ただし、bは精度が再現率よりどれだけ重視されているかを示すパラメター　b＝1がよく使われる。

$$F = \frac{(1+\mathrm{b}^2)\times \mathrm{P}\times \mathrm{R}}{\mathrm{b}_2\times \mathrm{P}+\mathrm{R}}$$

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{P}+\frac{1}{R}}$$

表 1.1　分割表

	教師データで正解とラベル付け	教師データで不正解とラベル付け
正解と予測	TP	FP
不正解と予測	FN	TN

表中の TP，FP，FN，TN は評価データ集合における各欄に対応する個数である．
また，これらの略称は以下の表現の省略形である．

TP : True Positive,　　FP : False Positive

FN : False Negative,　　TN : Ture Negative

	大 ← $p(+1 \mid \boldsymbol{x})$ → 小							
	各データの正解ラベル (+ は +1，− は −1 を表す)							
理想的な $p(+1 \mid \boldsymbol{x})$ の場合	+	+	+	+	−	−	−	−
一般的な $p(+1 \mid \boldsymbol{x})$ の場合	+	+	−	−	+	−	+	−
	a			b			c	

図 1.3　$p(+1 \mid \boldsymbol{x})$ に並べたデータの正解ラベルと閾値

表 1.2　理想的な $p(+1 \mid \boldsymbol{x})$ の場合の分割表と精度，再現率，正解率

	$\theta_{th} = a$	
ラベル付け	$+1$	-1
$+1$ と予測	1	0
-1 と予測	3	4
精度	1/1	
再現率	1/4	
正解率	5/8	

	$\theta_{th} = b$	
ラベル付け	$+1$	-1
$+1$ と予測	4	0
-1 と予測	0	4
精度	4/4	
再現率	4/4	
正解率	8/8	

	$\theta_{th} = c$	
ラベル付け	$+1$	-1
$+1$ と予測	4	3
-1 と予測	0	1
精度	4/7	
再現率	4/4	
正解率	5/8	

表 1.3　一般的な $p(+1 \mid \boldsymbol{x})$ の場合の分割表と精度，再現率，正解率

	$\theta_{th} = a$	
ラベル付け	$+1$	-1
$+1$ と予測	1	0
-1 と予測	3	4
精度	1/1	
再現率	1/4	
正解率	5/8	

	$\theta_{th} = b$	
ラベル付け	$+1$	-1
$+1$ と予測	2	2
-1 と予測	2	2
精度	2/4	
再現率	2/4	
正解率	4/8	

	$\theta_{th} = c$	
ラベル付け	$+1$	-1
$+1$ と予測	4	3
-1 と予測	0	1
精度	4/7	
再現率	4/4	
正解率	5/8	

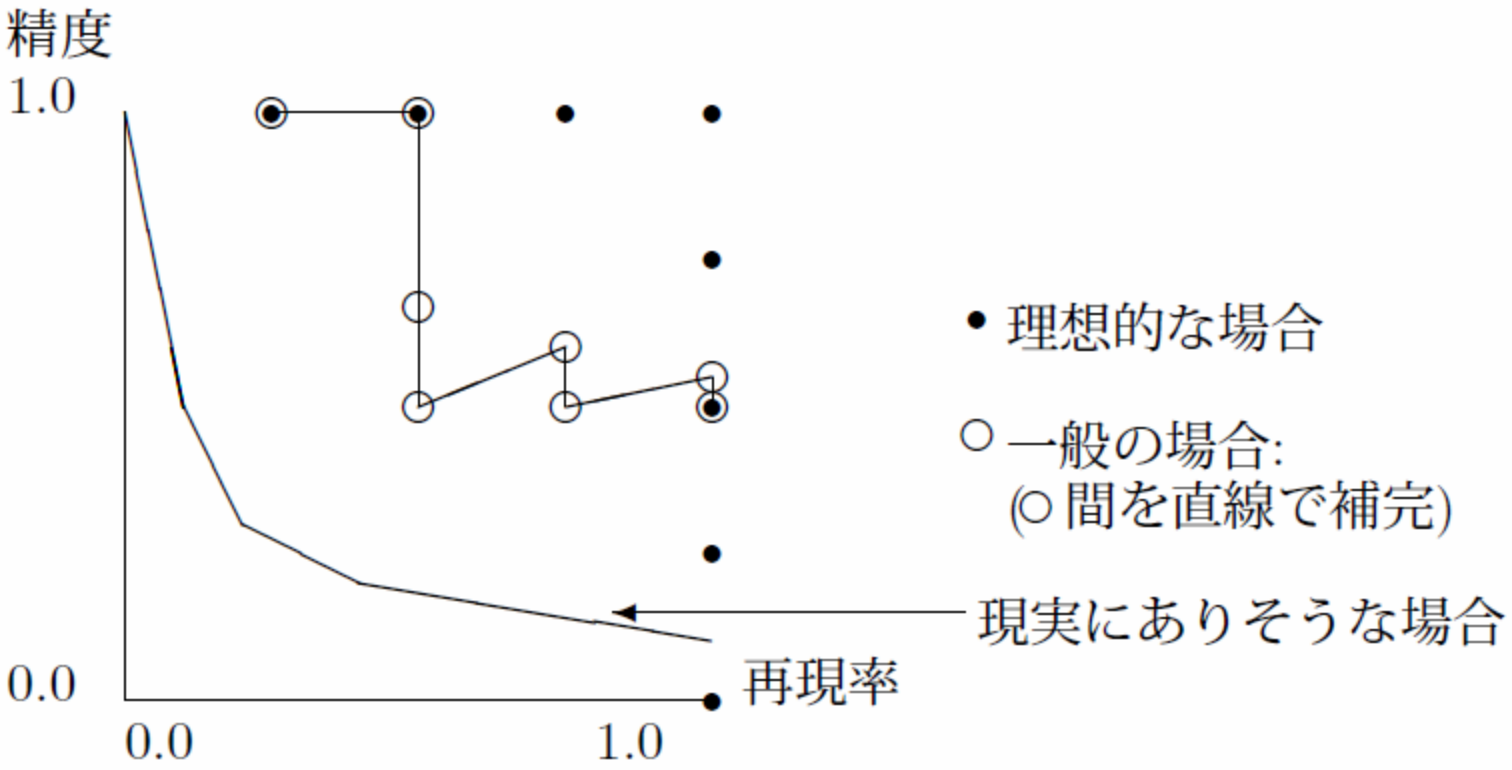
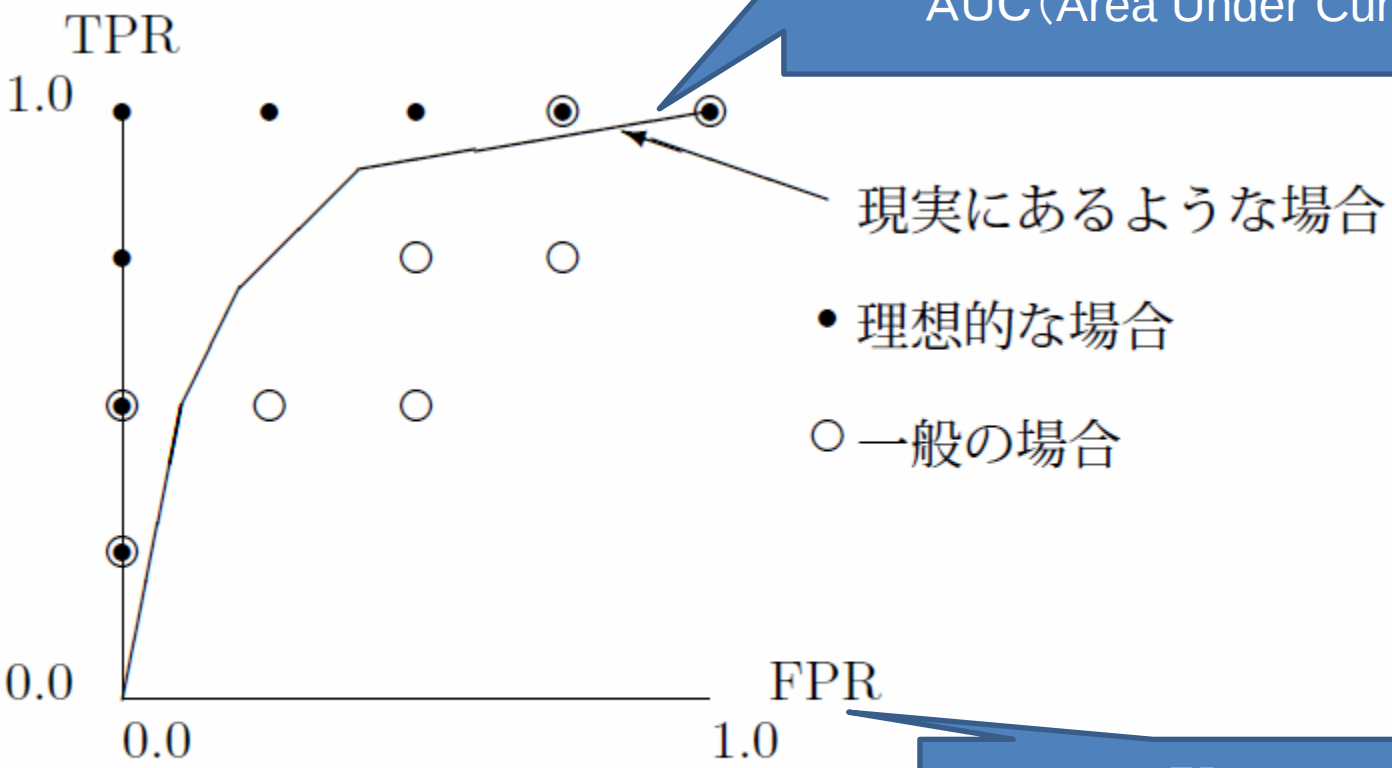


図 1.4　再現率 vs 精度

ROCとAUC

$$\frac{TP}{TP + FN}$$

ROC曲線
ROC曲線の下の部分の面積が
AUC（Area Under Curve)

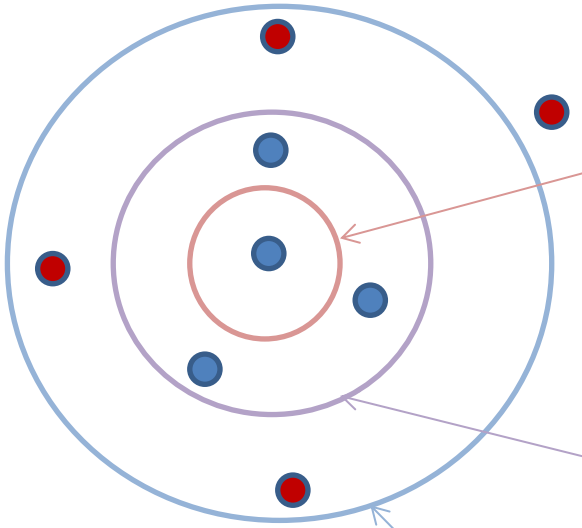


$$\frac{FP}{FP + TN}$$

図 1.5　ROC 曲線

理想的な場合（表1．2）

● ： 正解
● ： 不正解



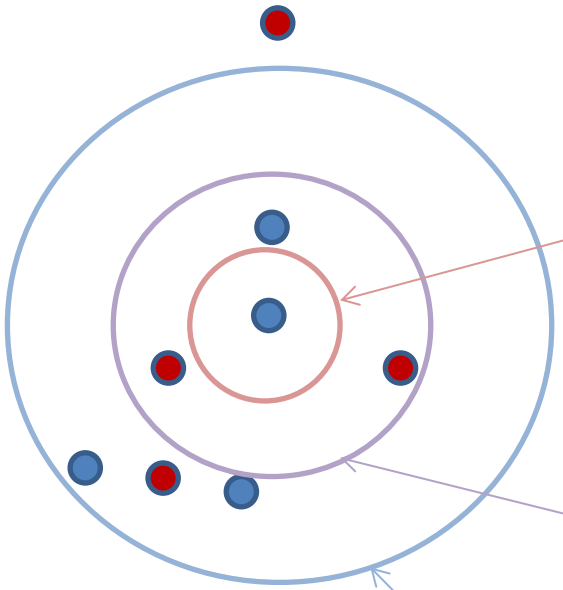
θth=a: TPR=1/4
FPR=0

θth=b: TPR=4/4
FPR=0

θth=c: TPR=4/4
FPR=3/4

現実的な場合(表1．3）

● ： 正解
● ： 不正解



θth=a: TPR=1/4
FPR=0

θth=b: TPR=2/4
FPR=2/4

θth=c: TPR=4/4
FPR=3/4

順位つき結果の評価

- 単純な識別では結果は全て同等
- 生成モデルの場合は、結果が適合性のよい順番に並ぶ。（表示も適合順）
- この場合の評価法について

Recall , Precision

- 処理qに適合する結果（以下、正解、という）の数: $|D_q|$
- 処理システムの順位つけられた結果: $(d_1 \ldots\ldots d_n)$
- d_i が処理qへの正解なら $r_i=1$、 そうでなければ $r_i=0$ とする。すると、
- 第k順位まで拾ったときの

$$\mathrm{Recall}(k) = \frac{1}{|D_q|} \sum_{1 \le i \le k} r_i$$

$$\mathrm{Precision}(k) = \frac{1}{k} \sum_{1 \le i \le k} r_i$$

平均適合率：average precision

$$\mathrm{AveragePrecsion} = \frac{1}{|D_q|}\sum_{1\le k\le N} r_k \times \mathrm{precision}(k)$$

ただし、Nは正解が最後に現れた順位

- 例：

$$AvPrec = \frac{1}{2}\left(\frac{1}{1}+\frac{2}{4}\right) = 0.75$$

順位	正解か
1	○
2	
3	
4	○
5	
6	

平均逆順位：Mean Reciprocal Rank(MRR)

$RR = \frac{1}{n}$　　：nは初めて正解がでた順位

もし、正解がひとつも現れなければ　$MRR = 0$

MRR = 全質問に対するRRの平均値

- 例

$$MRR = \left(\frac{1}{1} + \frac{1}{4}\right) / 2 = 0.625$$

順位	正解か	
	第1問	第2問
1	○	
2		
3	○	
4		○

nDCG

- DCG(Discounted Cumulative Gain)
  - 結果には関連度(relevancy)：Rが与えられている。Rは適当な範囲の数値
  - 順位i番目の結果の関連度をRiとする
  - *p*位までの結果に対するCG(Cumulative Gain):

$$CGp = \sum_{i=1}^{p} R_i$$

  - *CGp*に順位が低いものに関連度Rの高いものが現れた場合のペナルティを考慮したのが***DCGp***

$$DCGp = R_1 + \sum_{i=2}^{p} \frac{R_i}{\log_2 i}$$

or more generally $\sum_{i=1}^{p} f_i(R_i)$: f_iはiの減少関数

- DCGはRiの決め方や関数f_iの定義に強く依存
- そこで理想的な場合のDCG（＝$IDCG$）と実際の結果に対するDCGの比を使う $nDCG$

$$nDCGp = \frac{DCGp}{IDCGp}$$

DCG,nDCGの例

- 結果：R1=4, R2=1, R3=4, R4=2, R5=1
- $\log_2 3=1.58$, $\log_2 4=2$, $\log_2 5=2.32$
- $DCG_5=4+1+4/1.58+2/2+1/2.32=8.96$
- $IDCG_5=4+4+2/1.58+1/2+1/2.32=10.70$
- $nDCG_5=8.96/10.70=0.83$
- もし、結果が関連度Rの大きい順に並んでいれば、
  DCG=IDCGだから nDCG=1
- もし、結果が逆順なら（1，1，2，4，4）
  $DCG_5=1+1+2/1.58+4/2+4/2.32=6.98$
  → $IDCG_5=6.98/10.70=0.65$

学習と評価（教師ありの場合）

- 正解データがある場合。
- 正解データ全部を教師データとして機械学習。学習結果のシステムをs
- ｓ　を教師データで評価
- ｓ　を未知のデータで評価
  - 本当は、未知データでの評価をしたいが、なにしろ未知
- 正解データを教師データとテストデータに分割
  - 教師データで学習し、テストデータを未知データとみなして評価
  - 正解データが少ない場合：N-fold cross validation(N-交差検定）
    - 正解データをN等分。N-1個を教師データとして学習し、残りの1個で評価。これをN種類繰り返す。
    - 特殊なケースとして、1個だけを除いて学習し、その1個で評価。これをデータ数繰り返す。Leave-one-out法

教師なしの場合

- クラスタリングの場合
  - 人手で正解データを作っておき、教師あり学習と同じような評価。
  - 一応、再現率も計測できる。
- 正解データが存在しない場合
  - 学習結果をサンプリングして、人手で評価するしかない。
  - 再現率は評価できない。

クラスタリングの評価：Purity

- 生成されたクラスタがどれだけ多数派で占められているかを表す尺度

N：データ数, C：真のクラス集合 $= (C_1, \ldots C_K)$,

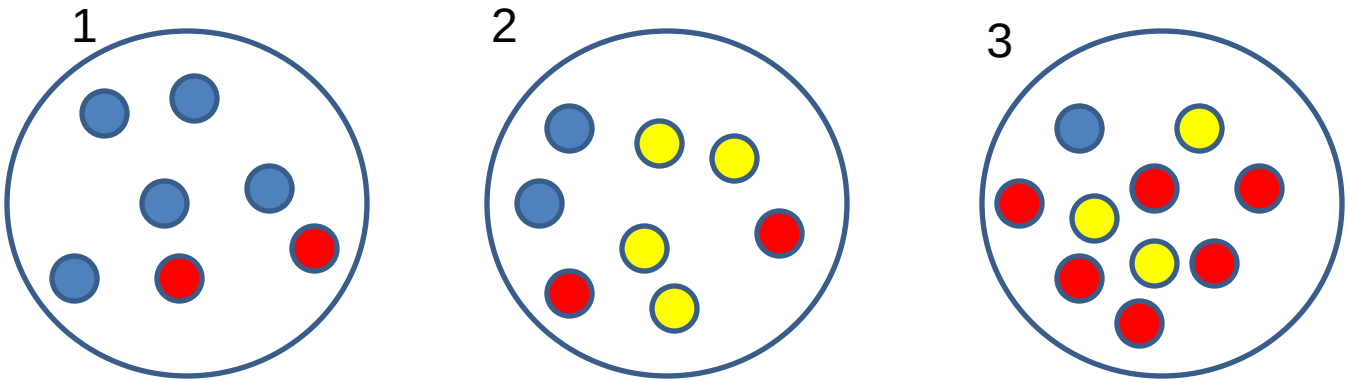
生成されたクラスタ数 $= L$

$n_{i,j}$：生成されたi番目のクラスタにおいて

j番目の真のクラスに属するデータ数

$$local \quad purity = \frac{1}{\sum_{j=1}^{L} n_{i,j}} \max\left(n_{i,j}\right)$$

$$global \quad purity = \frac{1}{\sum_{i=1}^{L}\sum_{j=1}^{K} n_{i,j}} \sum_{i=1}^{L} \max_{j}\left(n_{i,j}\right) = \frac{1}{N} \sum_{i=1}^{L} \max_{j}\left(n_{i,j}\right)$$



- local purity $purity(1)=\frac{5}{7},\ purity(1)=\frac{4}{8},\ purity(1)=\frac{6}{10}$
- global purity $purity=\frac{5+4+6}{7+8+10}=\frac{15}{25}=0.6$

- 問題点　何もしない場合
  - 全データが同一クラスタ　$purity=\frac{1}{N}\max_{i,j}\left(n_{i,j}\right)$
  - 1クラスタが1データ　$purity=\frac{1}{N}\sum_{i=1}^{L}\max_{j}\left(n_{i,j}\right)=\frac{1}{N}\sum_{i=1}^{L}1=\frac{N}{N}=1$

Inverse Purity

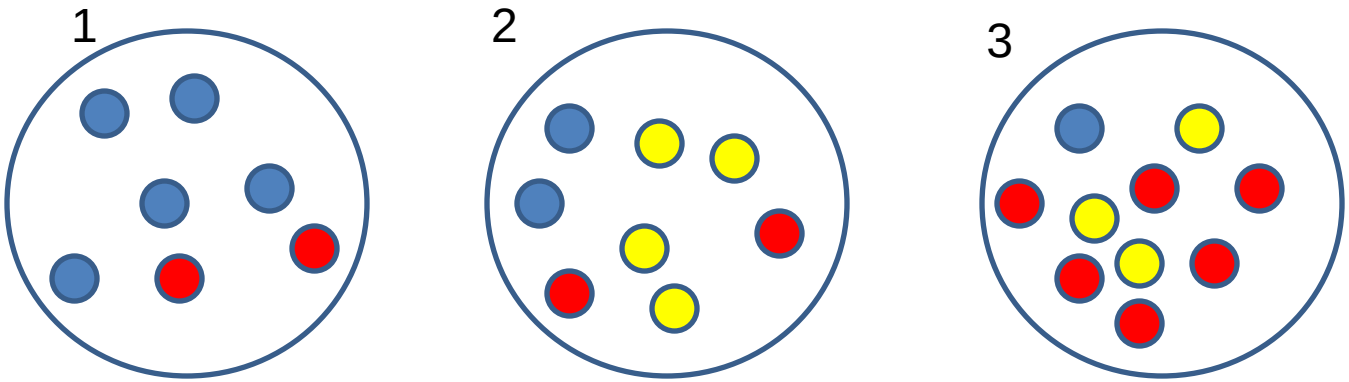
$$InversePurity = \frac{1}{N}\sum_{i=1}^{K}\left(\frac{\sum_{j=1}^{L} n_{i,j}}{\sum_{i=1}^{K} n_{i,j}}\max_{j}\left(n_{i,j}\right)\right)$$

クラスタiのデータ数

真のクラスjのデータ総数

1クラスタに1個のデータしかない場合も
Inverse Purityは1より小さい。
そこでPurityとの調和平均であるF値で評価

$$F値 = \frac{2}{\frac{1}{Purity} + \frac{1}{InversePurity}}$$



- $Purity = \frac{5+4+6}{7+8+10} = \frac{15}{25}$ ●8個、●7個、●10個

- $InversePurity = \frac{1}{25}\left(\frac{7}{8}\times 5 + \frac{8}{7}\times 4 + \frac{10}{10}\times 6\right) = 0.598$

- $F値 = \frac{2}{\frac{1}{0.6}+\frac{1}{0.598}} = 0.599$

評価者の一致性の評価

- κ計数
  - ある事象集合に対して評価がC1からCNのN種類。評価者1，2が各々評価点i,jをつける確率がp_{ij}

	評価者2				
評価者1		C1	…	CN	Σ
	C1	p_{11}		p_{1N}	$P_{1\cdot}$
	⋮			⋮	
	CN	p_{N1}	…	p_{NN}	$p_{N\cdot}$
	Σ	$p_{\cdot 1}$		$p_{\cdot N}$	1

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{N} p_{ii} - \sum_{i=1}^{N} p_{i\bullet} p_{\bullet i}}{1 - \sum_{i=1}^{N} p_{i\bullet} p_{\bullet i}}$$

- κ計数が1に近いほど評価者1，2の評価が一致している（評価行列が対角の場合）

例

	評価者2			
評価者1		0	1	Σ
	0	0．6	0	0.6
	1	0	0．4	0.4
	Σ	0.6	0.4	1

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{2} p_{ii} - \sum_{i=1}^{2} p_{i\bullet} p_{\bullet i}}{1 - \sum_{i=1}^{2} p_{i\bullet} p_{\bullet i}} = \frac{(0.6 + 0.4) - (0.6 \times 0.6 + 0.4 \times 0.4)}{1 - (0.6 \times 0.6 + 0.4 \times 0.4)} = 1$$

	評価者2			
評価者1		0	1	Σ
	0	0．25	0.25	0.5
	1	0.25	0.25	0.5
	Σ	0.5	0.5	1

$$\kappa = \frac{\sum_{i=1}^{2} p_{ii} - \sum_{i=1}^{2} p_{i\bullet} p_{\bullet i}}{1 - \sum_{i=1}^{2} p_{i\bullet} p_{\bullet i}} = \frac{(0.25 + 0.25) - (0.5 \times 0.5 + 0.5 \times 0.5)}{1 - (0.5 \times 0.5 + 0.5 \times 0.5)} = 0$$

テストコレクション

- (a) 入力データ集合、(b) 解くべき問題（識別など）、(c)問題において＜入力データ、推測結果＞対の集合、を組にしたデータベースをテストコレクションと呼び、機械学習システムの性能評価において必須の資源である

- ある入力データに対応する推定結果の個数が多いような問題（例えば、情報検索）では、＜入力データ、推測結果＞の大規模な集合を作ることは大規模テストコレクションでは困難
- Pooling method:、 同一の入力データ集合に対して、多数のシステムで同じ問題に対して出した上位N 個の結果を全て集める。N の値として、100 程度が多い。この結果に対してのみその適合性を人手で判断し、それを正解の集合とする