機械学習における評価指標
~AUC&C-index~

長崎大学病院　初期研修医　中島誉也

参考文献

- イベント予測モデルの評価指標

- 岡田遥平先生のnote
  - 臨床における良い予測モデルとは？
  - 一番わかりやすいROC曲線とAUCとC統計量

Table of contents

Table of contents

評価指標の紹介

- **識別能（Discrimination）：**

「予測モデルが正しく予測対象のイベント発生のリスクが高い患者とリスクが低い患者を区別できる性能」

- **較正能（Calibration）：**

「予測がどれくらい実際に当たるか」

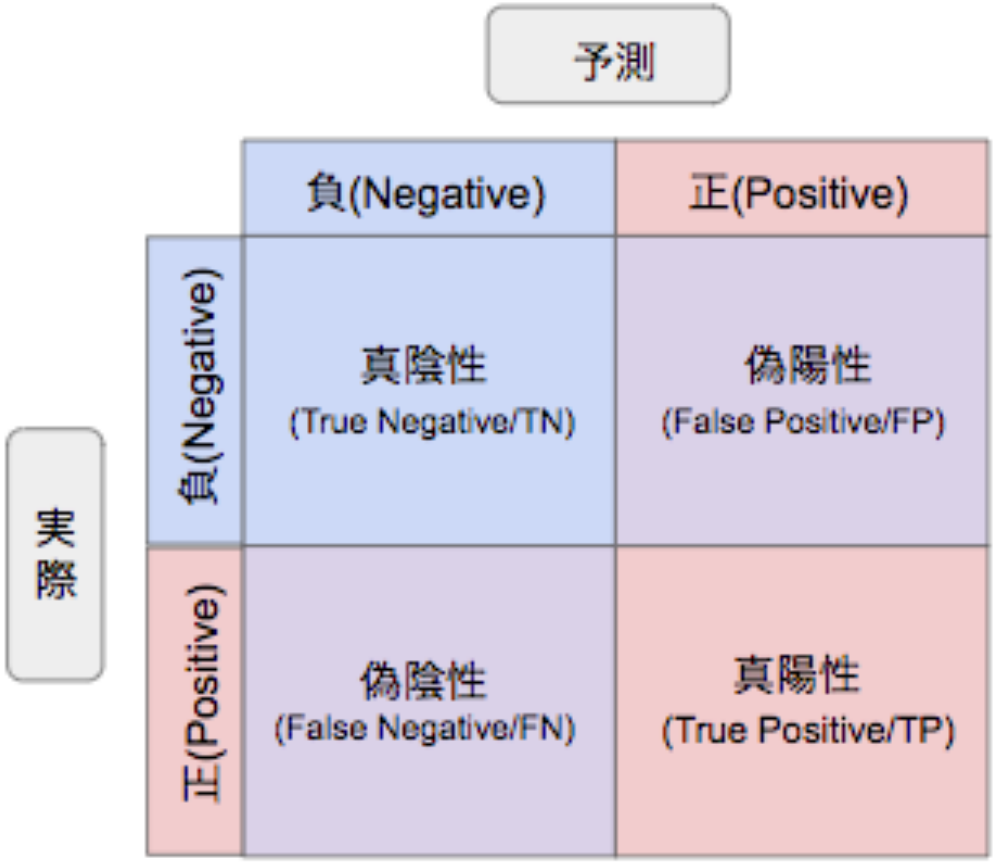
- **臨床的有用性（Clinical utility）**

予測モデルの究極的な目標は
「予測に基づき治療方針が変化し重要なアウトカムが改善する」

引用：臨床における良い予測モデルとは？

Table of contents

混同行列(Confusion Matrix)



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

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

$$Recall\ (True\ Positive\ Rate,\ TPR) = \frac{TP}{TP + FN}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Specificity\ (True\ Negative\ Rate,\ TNR) = \frac{TN}{TN + FP}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{TN + FP}$$

https://www.codexa.net/ml-evaluation-cls/

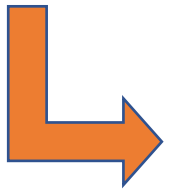
混同行列の問題点

1. データに偏りがある場合

元のデータでPositiveが99%，Negativeが1%のような場合，評価指標の数値は偏ったものになる．

2. カットオフ値の設定

カットオフ値をどこにするかで当然PositiveとNegativeの割合が変化し，
そのカットオフ値が適切なのかが曖昧になる．

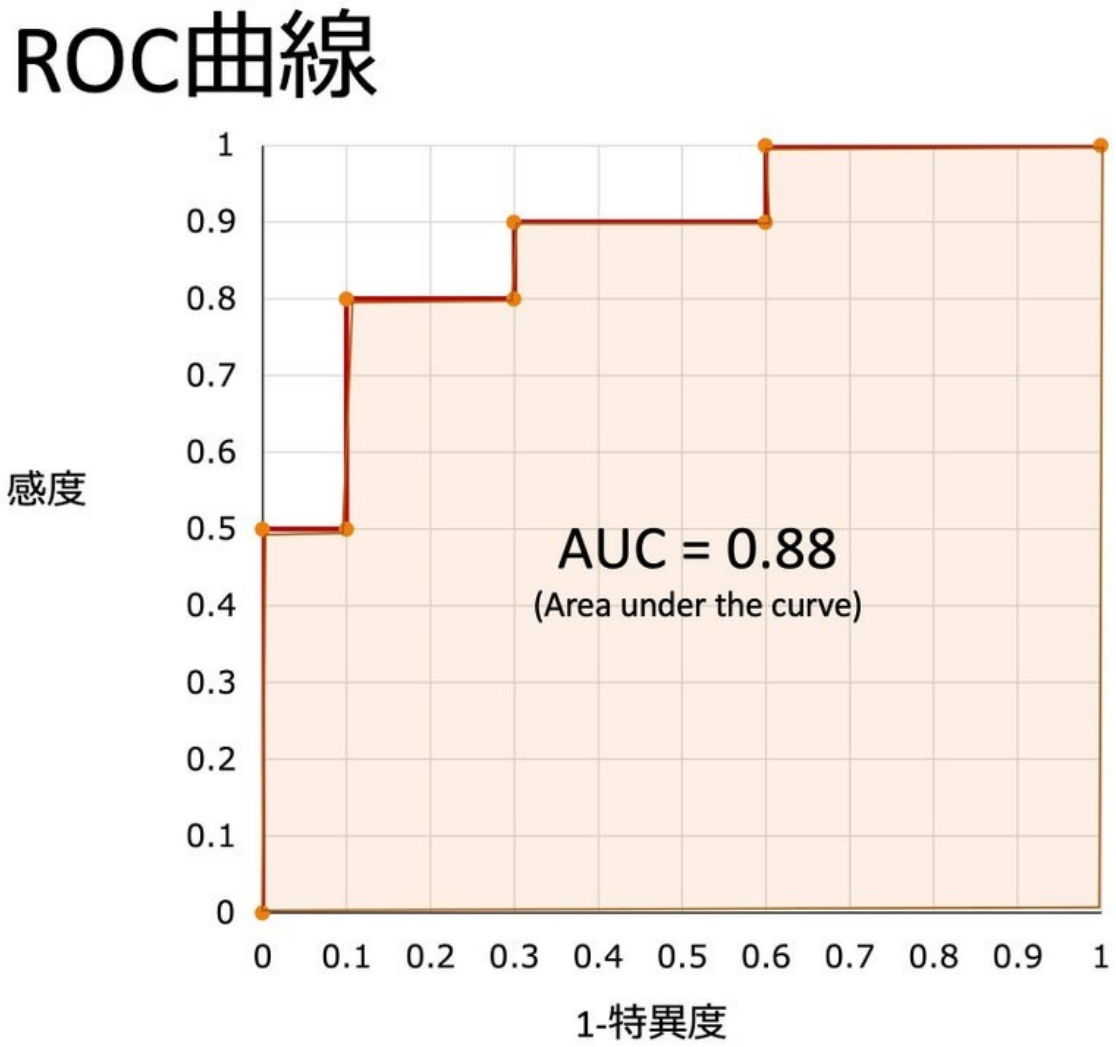


AUC(C統計量)による評価

AUCについて

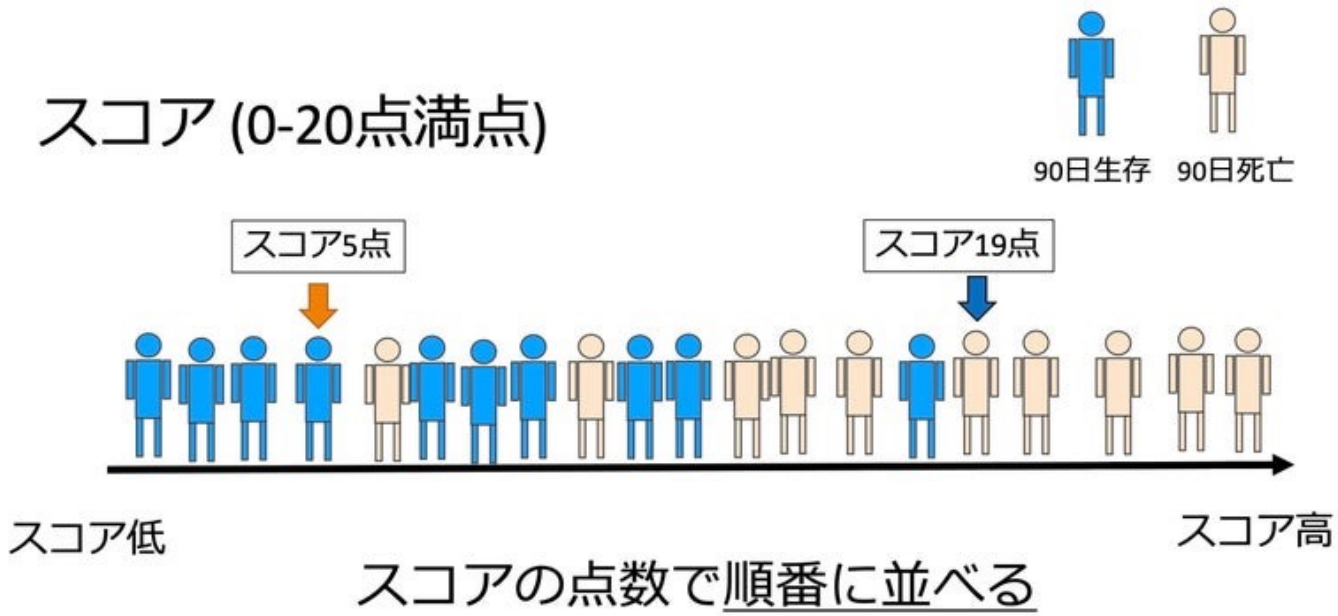
岡田遥平先生のnoteを引用しています

「対象としている連続変数（例：検査の値、スコア，または予測確率）が，二値のアウトカム（例えば生存/死亡）をどの程度正確に識別できるか」

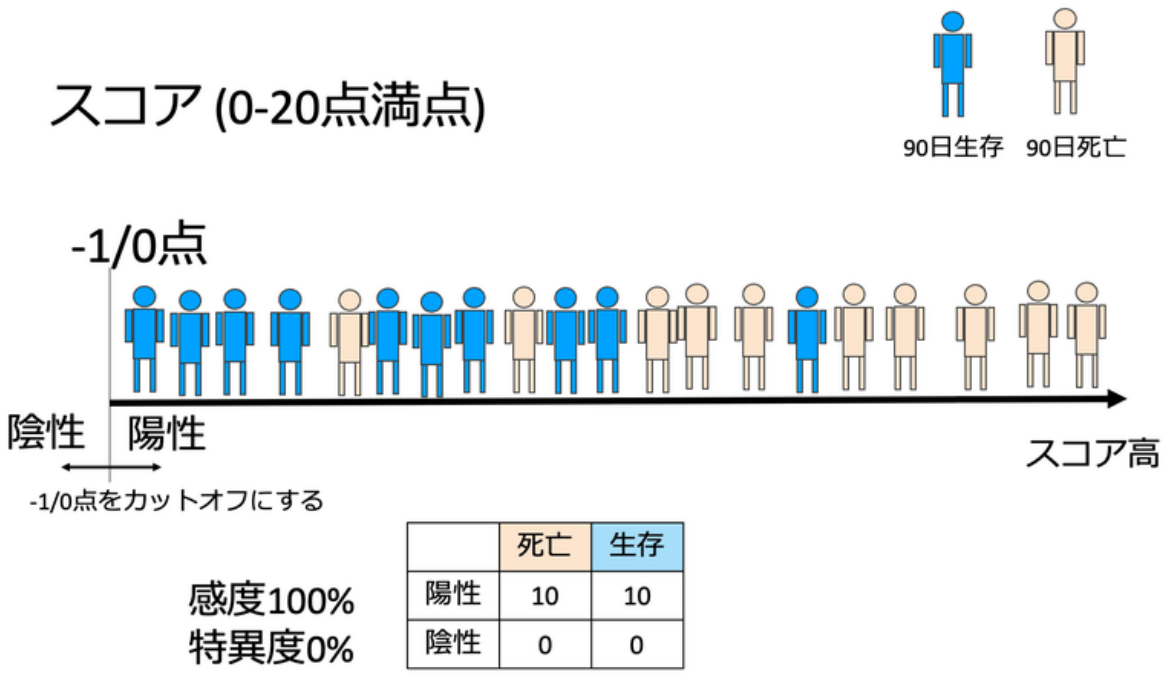


スコア	アウトカム
0	生存
1	生存
2	生存
5	生存
6	死亡
7	生存
8	生存
10	生存
11	死亡
12	生存
15	生存
16	死亡
16	死亡
17	死亡
18	生存
19	死亡
19	死亡
19	死亡
20	死亡
20	死亡

SOFAスコア：敗血症患者の重症度を表すスコア

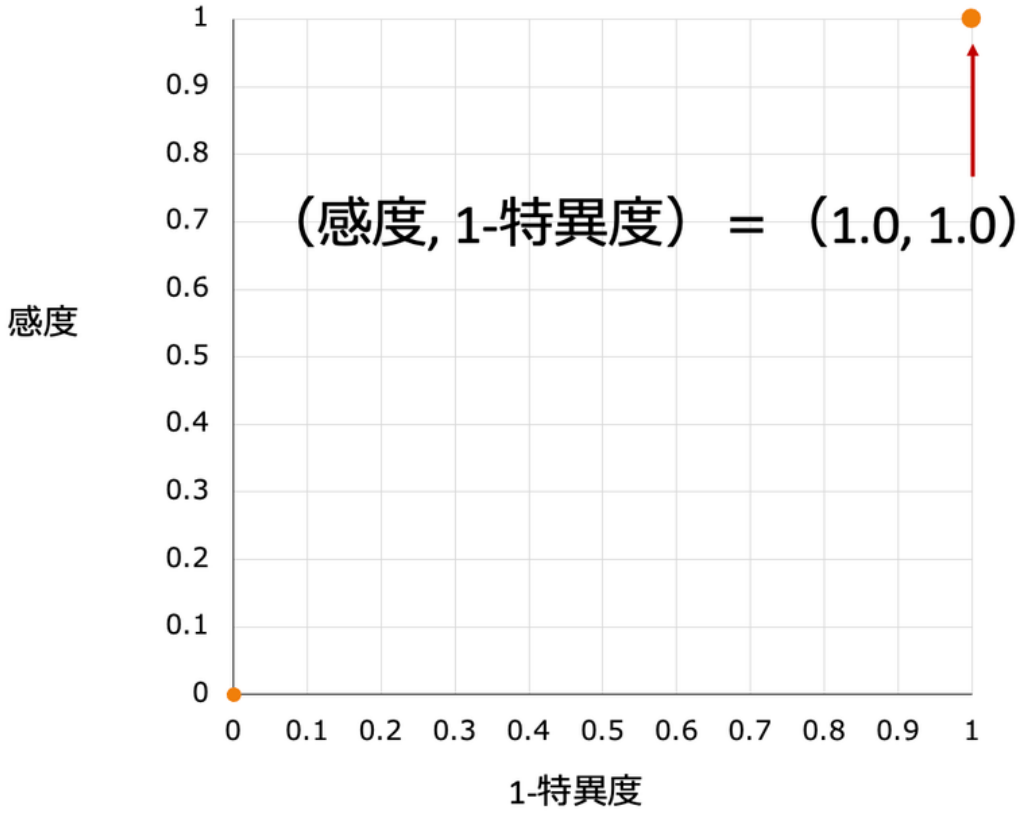


−１と０の間をカットオフにする

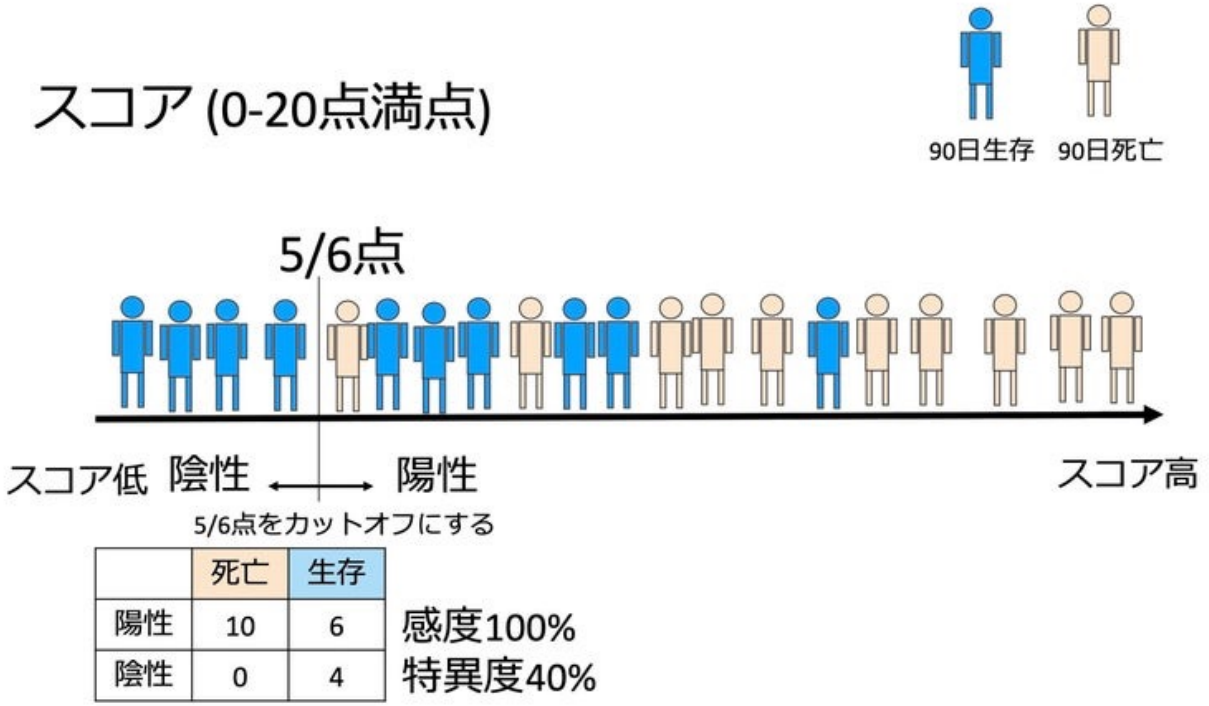


	死亡	生存
陽性	10	10
陰性	0	0

感度100%
特異度0%

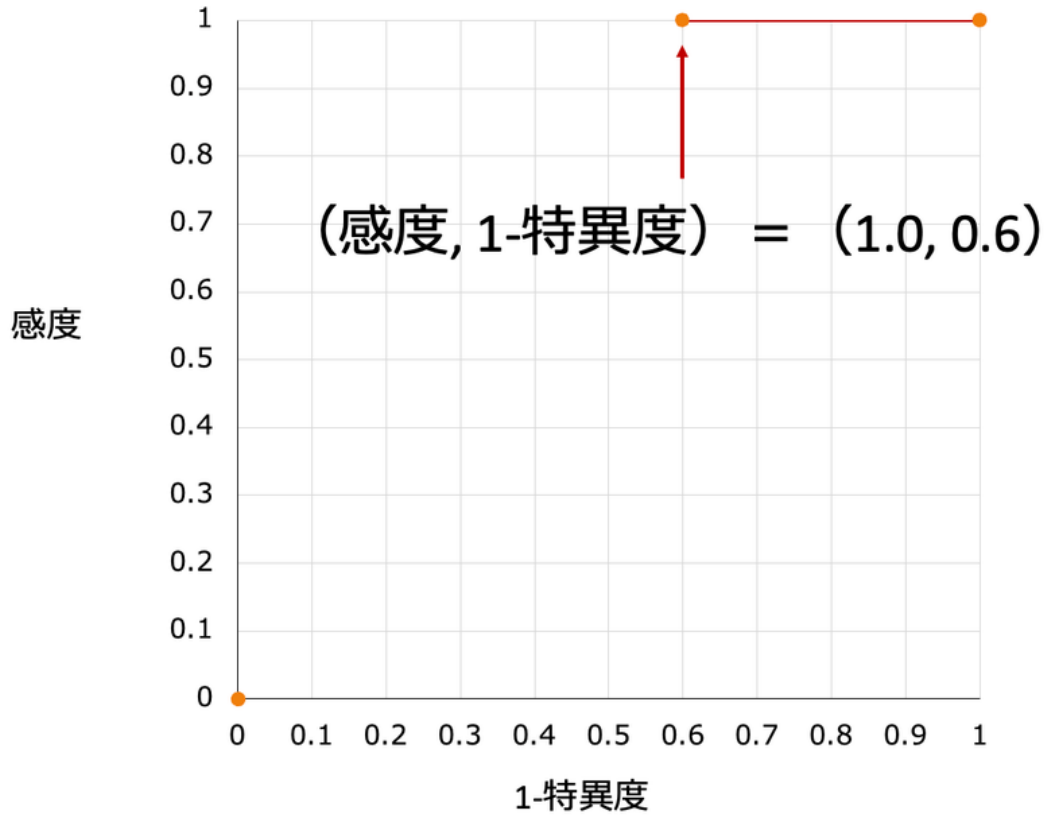


5と6の間をカットオフにする

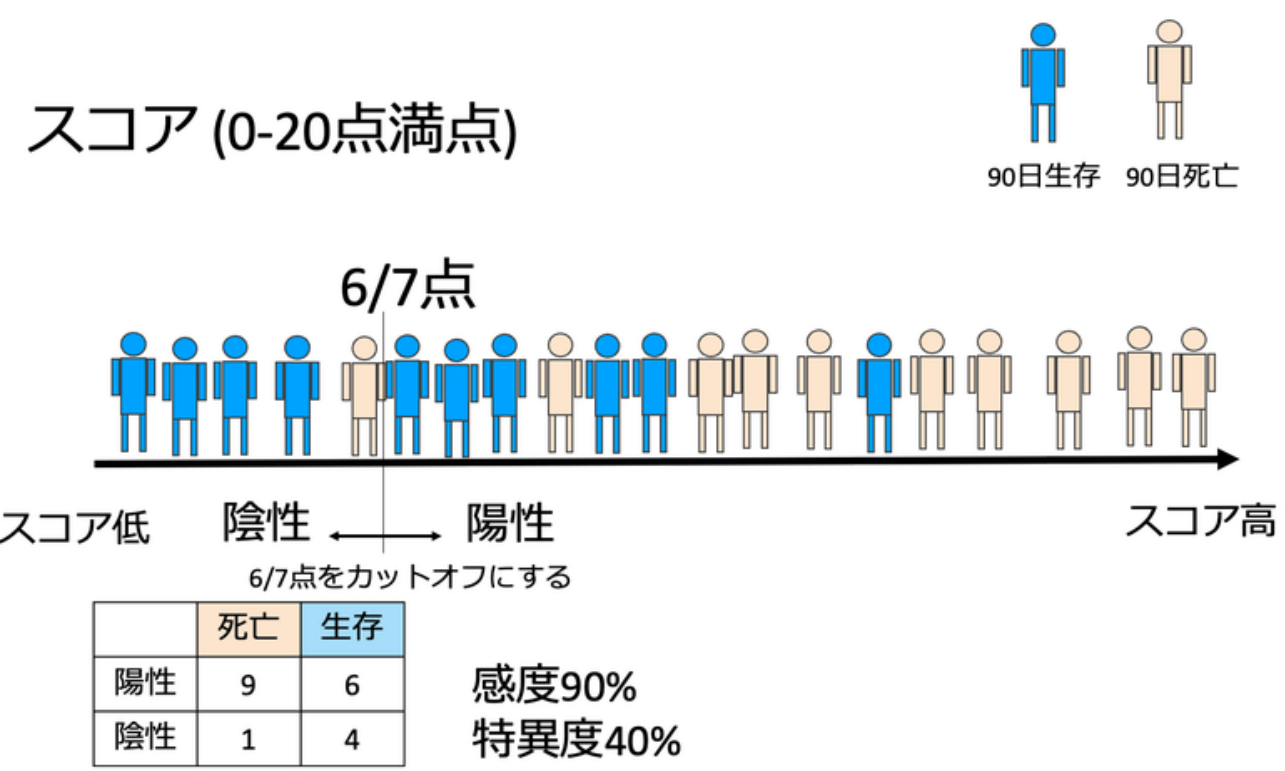


	死亡	生存
陽性	10	6
陰性	0	4

感度100%
特異度40%

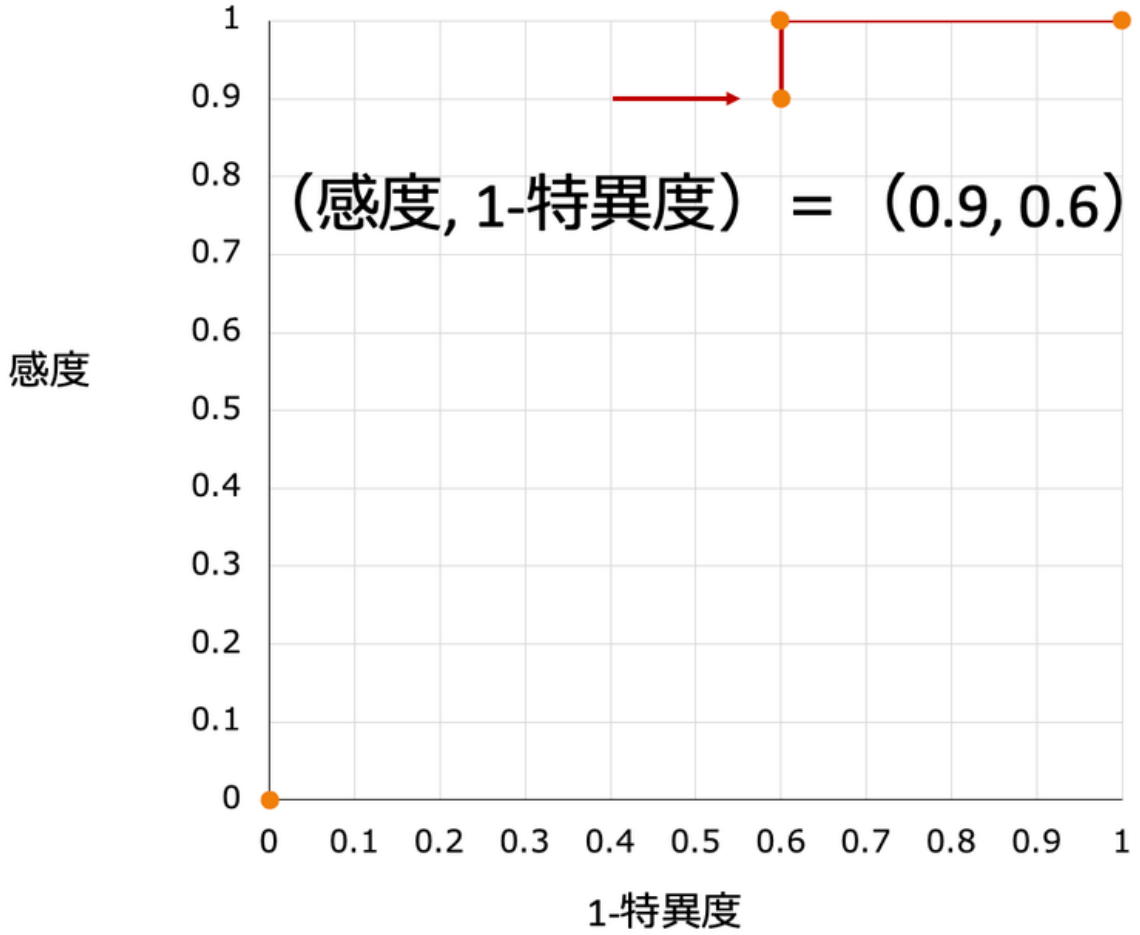


6と7の間をカットオフにする

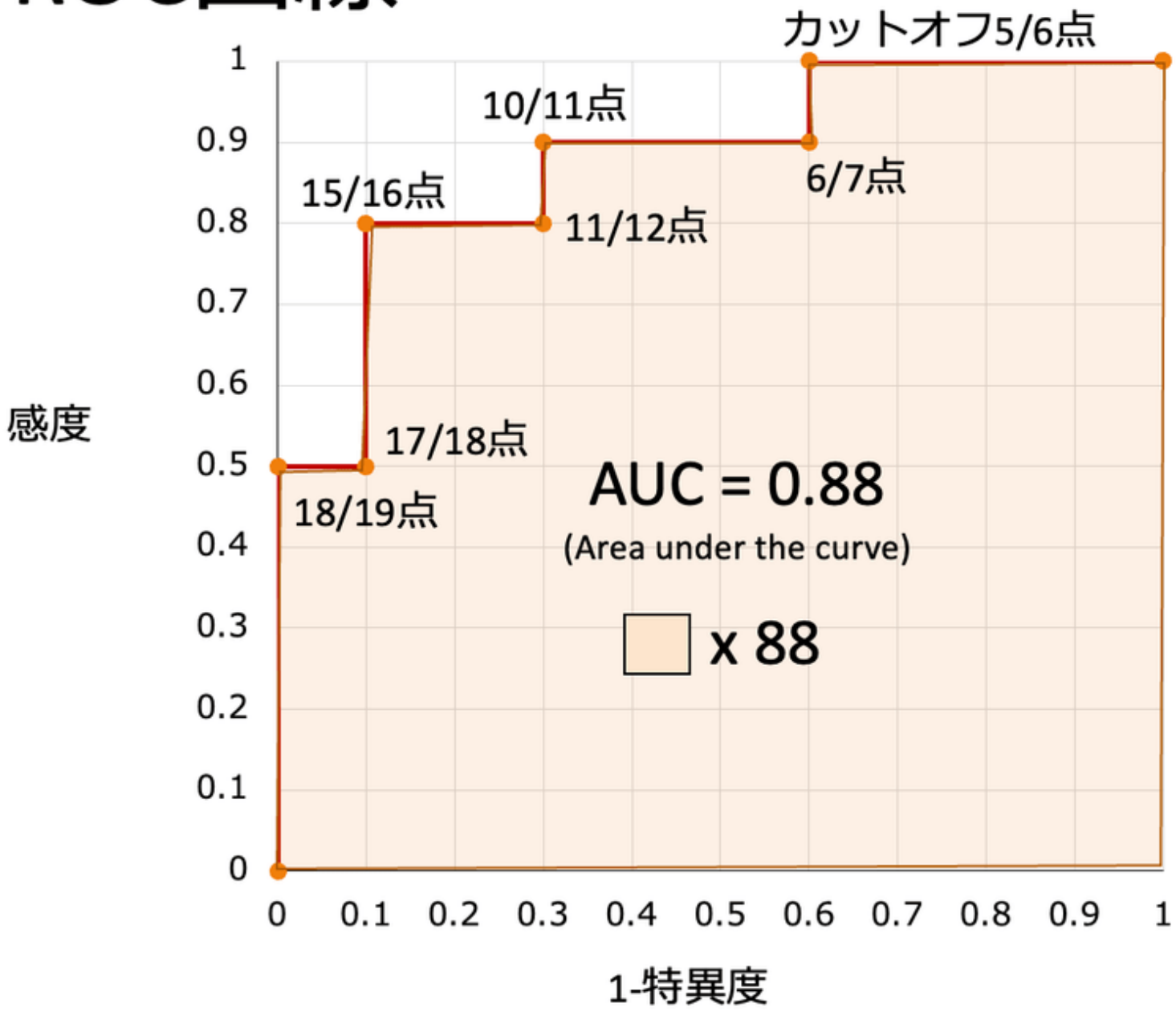


	死亡	生存
陽性	9	6
陰性	1	4

感度90%
特異度40%



ROC曲線


カットオフ5/6点
10/11点
6/7点
15/16点
11/12点
17/18点
18/19点
AUC = 0.88
(Area under the curve)
x 88
感度
1-特異度
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

C統計量について

岡田遥平先生のnoteを引用

「アウトカムのイベントなし（生存）の患者とアウトカムのイベントあり（死亡）の患者を、ランダムに1人ずつ抽出した時に、それぞれの患者のスコアの大小関係が,（抽出した生存の患者のスコア）＜（抽出した死亡の患者のスコア）となる確率」

構築したスコアやモデルが適切に識別できている確率

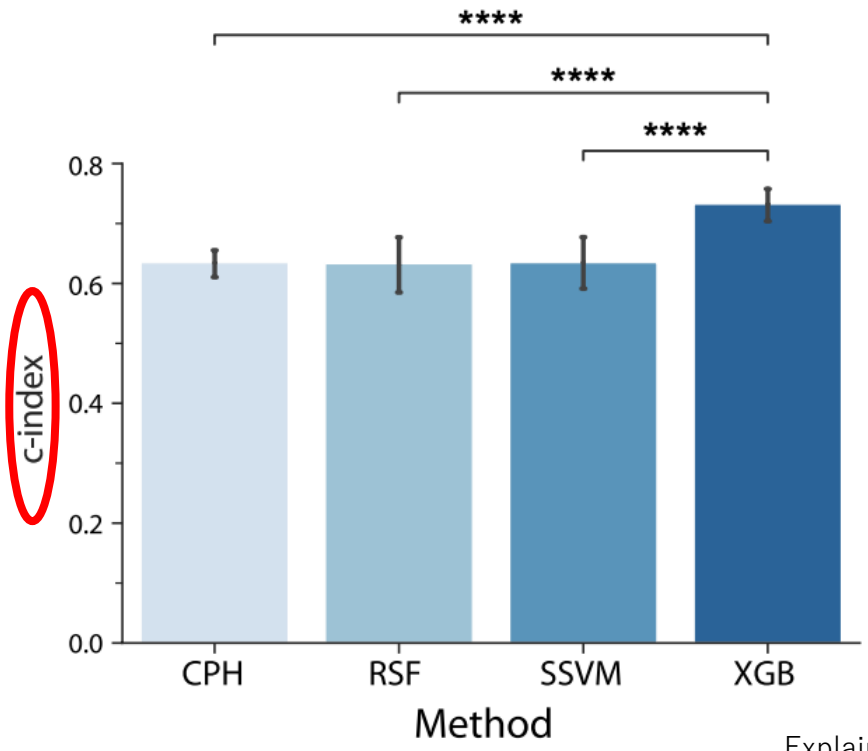
計算してみよう！！

岡田遥平先生のnoteを参照

Table of contents

C-indexについて

- 生存時間解析の分野においても，機械学習を用いた予測モデルの開発の研究が盛んに行われている.
- 打ち切りの概念を取り入れなければならないため，AUCで評価することは困難.
- 打ち切りの概念を取り入れつつ，AUCのように評価したい！！ → C –indexの登場



Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival

様々な生存時間解析×機械学習モデル

Task	Model	sksurv	lifelines	statsmodels	pycox
Survival Function Estimation	Non-parametric	✓	✓	✓	✓
	Parametric	✗	✓	✗	✗
CHF Estimation	Non-parametric	✓	✓	✗	✗
	Parametric	✗	✓	✗	✗
Linear Regression	Cox	✓	✓	✓	✓
	Cox + Elastic-Net	✓	✓	✓	✗
	Log-normal AFT	✓	✓	✗	✗
	Log-logistic AFT	✗	✓	✗	✗
	Weibull AFT	✗	✓	✗	✗
	Piecewise exponential	✗	✓	✗	✓
Non-linear Regression	Aalen	✗	✓	✗	✗
	Gradient Boosted AFT	✓	✗	✗	✗
	Gradient Boosted Cox	✓	✗	✗	✗
	Heterogenous Ensemble	✓	✗	✗	✗
	NN (Grouped survival times)	✗	✗	✗	✓
	NN (Proportional hazards)	✗	✗	✗	✓
	NN (Piecewise exponential)	✗	✗	✗	✓
	Random Survival Forest	✓	✗	✗	✗
	Survival SVM	✓	✗	✗	✗
	Survival Tree	✓	✗	✗	✗
Evaluation	Brier Score	✓	✗	✗	✓
	Concordance Index	✓	✓	✗	✓
	Time-dependent ROC	✓	✗	✗	✗

Table 1: Availability of methods. AFT: Accelerated Failure Time. CHF: Cumulative Hazard Function. NN: Neural Network. SVM: Support Vector Machine.

Pythonでのパッケージ(Scikit-survival)
SurvCART: Constructing Survival Tree in R | R-bloggers
https://www.r-bloggers.com › 2022/01

C-indexについて

異なるアウトカムじゃないと×

C統計量(Harrell's C-Statistic.)：

イベント発症者(Yi=1)の予測確率 Pi の分布と非発症者(Yj=0)の予測確率 Pj の分布それぞれからランダムサンプリングした時，発症者での予測確率の方が高くなる確率.

???

同じアウトカムでも良い！！

C-index(C-statistic by Uno et al.)：

標本からランダムに2つの異なるデータi，jを取り出してペアにした時，観測打ち切り時点τまでの生存時間Tの短長と予測確率P(t)の大小が一致する確率.

標本からランダムに2つの異なるデータi, jを取り出してペアにした時,
観測打ち切り時点τまでの生存時間Tの短長と予測確率P(t)の大小が一致する確率.

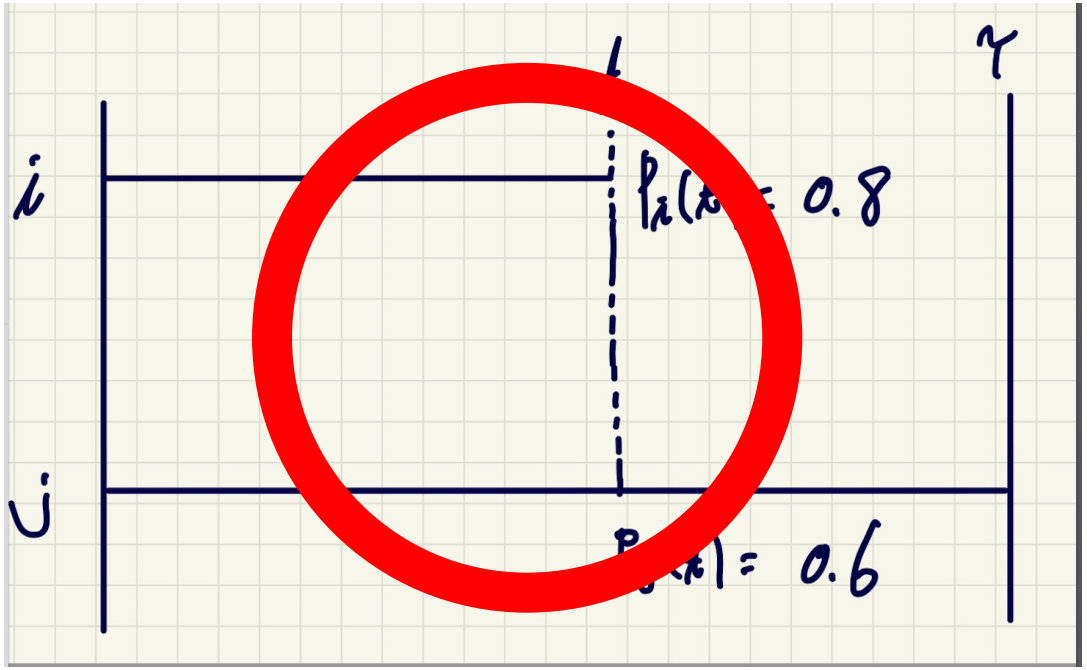
↓

ランダムに2人(iとj)を抽出した時に，観測期間τまでの生存期間の長さの短長と
そのモデルによって予測される時点tの死亡確率の大小が一致する確率.

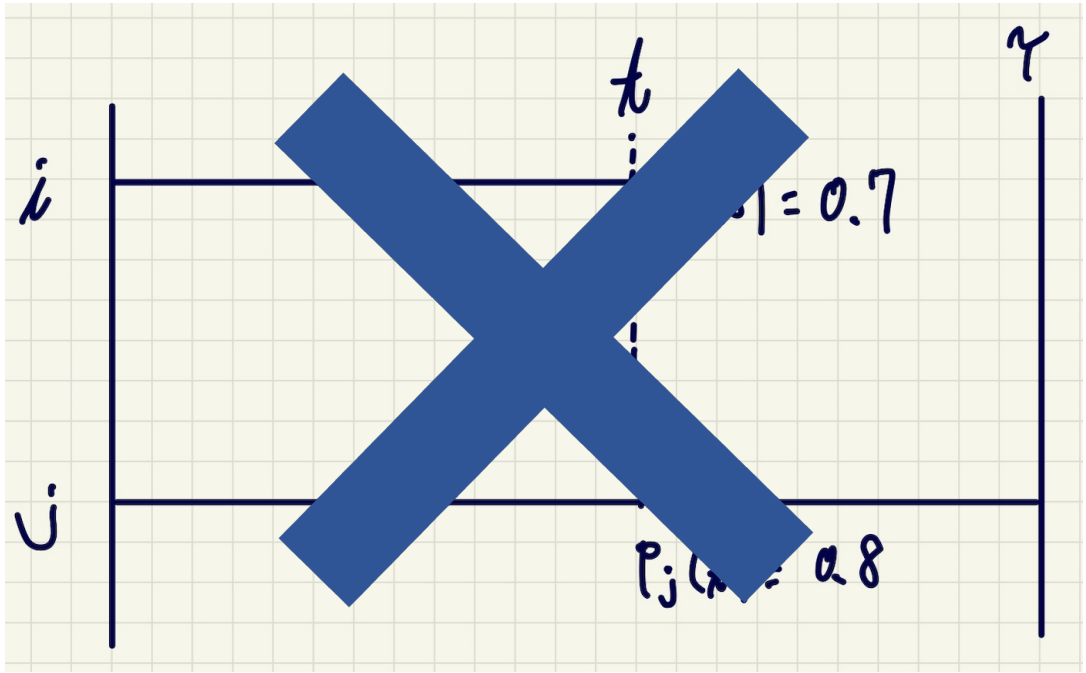
↓

適切に識別できてる！！

iとjを比較した時に，iの方が，実際のデータでもjより早く死亡(＝時点t)しており，
かつ
モデルによって予測される時点tにおける死亡確率がjよりも大きければcountできる！



iの方が，早く死亡(＝時点t)しており，
予測される時点tでの死亡確率がjよりも大きい
（＝正しい予測）



iの方が，早く死亡(＝時点t)しているが，
予測される時点tでの死亡確率がjよりも小さい
（＝誤った予測）

C-indexの式(Uno, et al., 2011)

$$\widehat{C}_\tau = \frac{\sum_{i,j} D_i \, I(T_i^* < T_j^*) \, I\{P_i(t) > P_j(t)\}}{\sum_{i,j} D_i \, I(T_i^* < T_j^*)}$$

- 分母が比較可能な（comparable）ペアの総数
- 分子はそのうち T と P(t) の短長と大小が揃っている（concordant）ペアの総数を表す.
- ペアが比較可能である必要十分条件は，ペア内で観察時間 T * の短い方のイベントが観測されていること（＝iにアウトカムが発生していること）である．(Di = 1)

計算してみよう！！

No.	Outcome	Time(year)
1	生存	10
2	死亡	3
3	死亡	7
4	生存	10
5	生存	10
6	死亡	2
7	死亡	8
8	生存	10
9	死亡	1
10	生存	10

心筋梗塞による死亡をアウトカムとした生存時間解析
観測期間τは10年とする.

Random Survival Forestを用いて行った死亡確率の予測結果

観測時点 (時点t)	No.1	No.2 (死亡)	No.3 (死亡)	No.4	No.5	No.6 (死亡)	No.7 (死亡)	No.8	No.9 (死亡)	No.10
1	0.03	0.12	0.1	0.03	0	0.2	0.11	0.06	0.31	0.08
2	0.06	0.19	0.12	0.04	0.02	0.3	0.17	0.06	0.34	0.11
3	0.1	0.22	0.12	0.05	0.06	0.31	0.25	0.07	0.41	0.14
7	0.1	0.3	0.13	0.09	0.07	0.4	0.23	0.08	0.42	0.17
8	0.12	0.32	0.2	0.13	0.17	0.44	0.27	0.09	0.43	0.17
10	0.16	0.36	0.26	0.18	0.18	0.47	0.3	0.13	0.46	0.19

観測時点 (時点t)	No.1	No.2 (死亡)	No.3 (死亡)	No.4	No.5	No.6 (死亡)	No.7 (死亡)	No.8	No.9 (死亡)	No.10
1	0.03	0.12	0.1	0.03	0	0.2	0.11	0.06	0.31	0.08
2	0.06	0.19	0.12	0.04	0.02	0.3	0.17	0.06	0.34	0.11
3	0.1	0.22	0.12	0.05	0.06	0.31	0.25	0.07	0.41	0.14
7	0.1	0.3	0.13	0.09	0.07	0.4	0.23	0.08	0.42	0.17
8	0.12	0.32	0.2	0.13	0.17	0.44	0.27	0.09	0.43	0.17
10	0.16	0.36	0.26	0.18	0.18	0.47	0.3	0.13	0.46	0.19

$$\widehat{C}_\tau = \frac{\sum_{i,j} D_i \, I(T_i^* < T_j^*) \, I\{P_i(t) > P_j(t)\}}{\sum_{i,j} D_i \, I(T_i^* < T_j^*)}$$

e.g.)
i をNo.9とすると，時点t = 1, τ = 10であるから，
比較可能となる候補 j は，No.1, 2, 3, 4, 5, 6, 7, 8, 10（9組）

例えば，i = No.9, j = No.1とすると，

$$P_i(t) = P_{No.9}(t = 1) = 0.31$$
$$P_j(t) = P_{No.1}(t = 1) = 0.03$$
$$P_i(t) > P_j(t)$$であるから，

i の方が，早く死亡(＝時点t)しており，

予測される時点tでの死亡確率がjよりも大きい　→ countされる

観測時点 (時点t)	No.1	No.2 (死亡)	No.3 (死亡)	No.4	No.5	No.6 (死亡)	No.7 (死亡)	No.8	No.9 (死亡)	No.10
1	0.03	0.12	0.1	0.03	0	0.2	0.11	0.06	0.31	0.08
2	0.06	0.19	0.12	0.04	0.02	0.3	0.17	0.06	0.34	0.11
3	0.1	0.22	0.12	0.05	0.06	0.31	0.25	0.07	0.41	0.14
7	0.1	0.3	0.13	0.09	0.07	0.4	0.23	0.08	0.42	0.17
8	0.12	0.32	0.2	0.13	0.17	0.44	0.27	0.09	0.43	0.17
10	0.16	0.36	0.26	0.18	0.18	0.47	0.3	0.13	0.46	0.19

$$\widehat{C}_\tau = \frac{\sum_{i,j} D_i \, I(T_i^* < T_j^*) \, I\{P_i(t) > P_j(t)\}}{\sum_{i,j} D_i \, I(T_i^* < T_j^*)}$$

J = No.2の場合もcountされる．(∵0.31>0.12)
J = No.3の場合もcountされる．(∵0.31>0.1)
J = No.4の場合もcountされる．(∵0.31>0.03)
J = No.5の場合もcountされる．(∵0.31>0)
J = No.6の場合もcountされる．(∵0.31>0.2)
J = No.7の場合もcountされる．(∵0.31>0.11)
J = No.8の場合もcountされる．(∵0.31>0.06)
J = No.10の場合もcountされる．(∵0.31>0.08)

$$C_{i=No.9} = \frac{9}{9} = 1$$

観測時点 (時点t)	No.1	No.2 (死亡)	No.3 (死亡)	No.4	No.5	No.6 (死亡)	No.7 (死亡)	No.8	No.9 (死亡)	No.10
1	0.03	0.12	0.1	0.03	0	0.2	0.11	0.06	0.31	0.08
2	0.06	0.19	0.12	0.04	0.02	0.3	0.17	0.06	0.34	0.11
3	0.1	0.22	0.12	0.05	0.06	0.31	0.25	0.07	0.41	0.14
7	0.1	0.3	0.13	0.09	0.07	0.4	0.23	0.08	0.42	0.17
8	0.12	0.32	0.2	0.13	0.17	0.44	0.27	0.09	0.43	0.17
10	0.16	0.36	0.26	0.18	0.18	0.47	0.3	0.13	0.46	0.19

$$\widehat{C}_\tau = \frac{\sum_{i,j} D_i \, I(T_i^* < T_j^*) \, I\{P_i(t) > P_j(t)\}}{\sum_{i,j} D_i \, I(T_i^* < T_j^*)}$$

同様に,

i = No.6の時, 比較可能群$(= \boldsymbol{j})$は$\boldsymbol{No.1,2,3,4,5,7,8,10}$（8組）， $\boldsymbol{C_{i=No.6} = \frac{8}{8} = 1}$

i = No.2の時, 比較可能群$(= \boldsymbol{j})$は$\boldsymbol{No.1,3,4,5,7\,8,10}$（7組）， $\boldsymbol{C_{i=No.2} = \frac{6}{7} \fallingdotseq 0.86}$

i = No.3の時, 比較可能群$(= \boldsymbol{j})$は$\boldsymbol{No.1,4,5,7,8,10}$（6組）， $\boldsymbol{C_{i=No.3} = \frac{4}{6} \fallingdotseq 0.67}$

i = No.7の時, 比較可能群$(= \boldsymbol{j})$は$\boldsymbol{No.1,4,5,8,10}$（5組）， $\boldsymbol{C_{i=No.7} = \frac{5}{5} = 1}$

平均 $\boldsymbol{\widehat{C} \fallingdotseq 0.91}$

AUC(C統計量)，C-indexの弱点

予測確率がランダムに選ばれたもう1つのサンプルが大きいかどうかしか見ていない

→　どれくらい正確な予測ができているかどうかまでは分からない

→　正しい予測ができているのかをCalibrationしてあげる必要がある

おまけ

1. 予測確率が同じ P のデータを集めると確率 P でイベントを生じるように
「同じ予測値を与える状況ではその予測値は較正(calibration)されるべき」

2. 不確実な個々の観測には確率 1 か 0 を個々に割り当てる予測より，0<Pi<1 という確率 Pi で予測する方が良い

3. このような確率的な予測 Pi においては「なるべく Pi の分布が極端になるように(0か1に近くなるように)予測確率を割り当てるべき」

Brier Score $$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$