データマイニング

Data Mining

8: 分類③ Classification

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

ロジスティック回帰 Logistic Regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M$$

回帰をクラス分類に適用する

$$Sold\ within\ 1\ year\ or\ not = \beta_1 MedInc + \beta_2 HouseAge + \cdots \beta_8 Longitude$$

$Yes: 1, No: 0$

ダミー変数 Dummy Variable

$$House\ Value = \beta_1 MedInc + \beta_2 HouseAge + \cdots \beta_8 Longitude$$

量的変数 Quantitative Variable

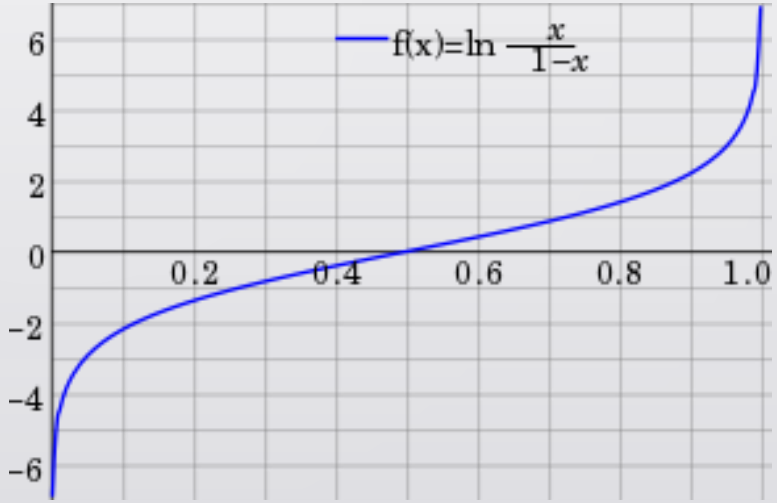
オッズとロジット Odds and Logit

$$Odds = \frac{P(Y=1)}{1-P(Y=1)}$$

ある事象（$Y=1$）が起こる確率と、
起こらない確率の比

Ratio between the probability that an event ($Y=1$) occurs and the probability that it does not occur

$$Logit(P) = log\frac{P(Y=1)}{1-P(Y=1)}$$



https://en.wikipedia.org/wiki/Logit

ロジスティック回帰 Logistic Regression

$$log\frac{P(Y=1)}{1-P(Y=1)} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M$$

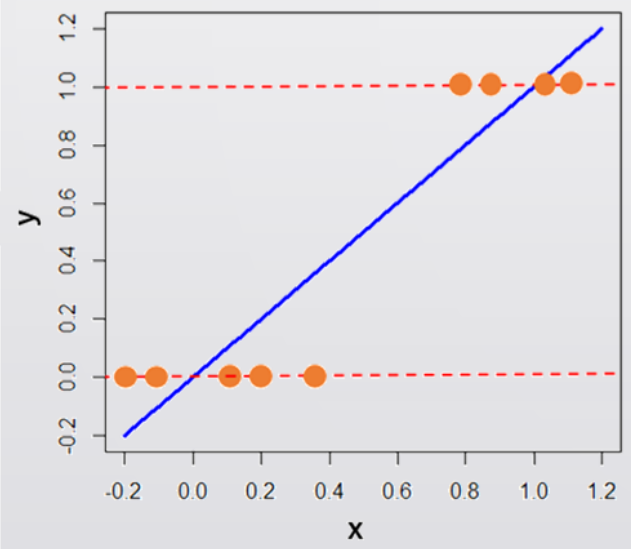
$$\frac{P}{1-P} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M}$$

$$P = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M)}}$$

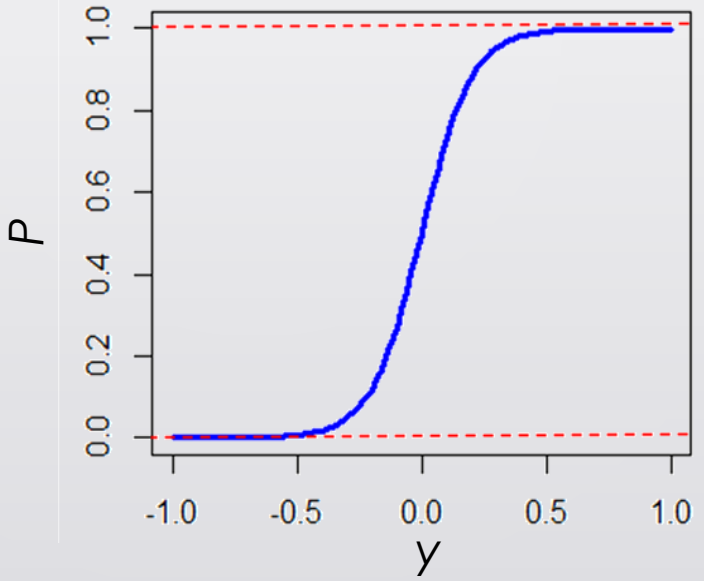
ロジスティック回帰 Logistic Regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M \qquad P = \frac{1}{1 + e^{-y}}$$

変換前 Before Conversion



変換後 Before Conversion



- ターゲット変数
  Target Variable

https://bellcurve.jp/statistics/course/26934.html

最尤推定法 Maximum Likelihood Estimation

与えられたデータが観測される確率が最大になるよう回帰係数 β_k を決定する

Determine regression coefficient β_k so as to maximize the probability that given data is observed

$$Y_i \begin{cases} 1 \\ \\ 0 \end{cases}$$

$P_i = p(Y_i = 1)$ i番目のデータY_iが1である確率

Probability that i–th data Y_i is 1

$$L = \prod_{i=1}^{N} {P_i}^{Y_i}(1 - P_i)^{1-Y_i}$$

Lを最大化する

Maximize L

最尤推定法 Maximum Likelihood Estimation

$$L = \prod_{i=1}^{N} P_i{}^{Y_i}(1 - P_i)^{1-Y_i}$$

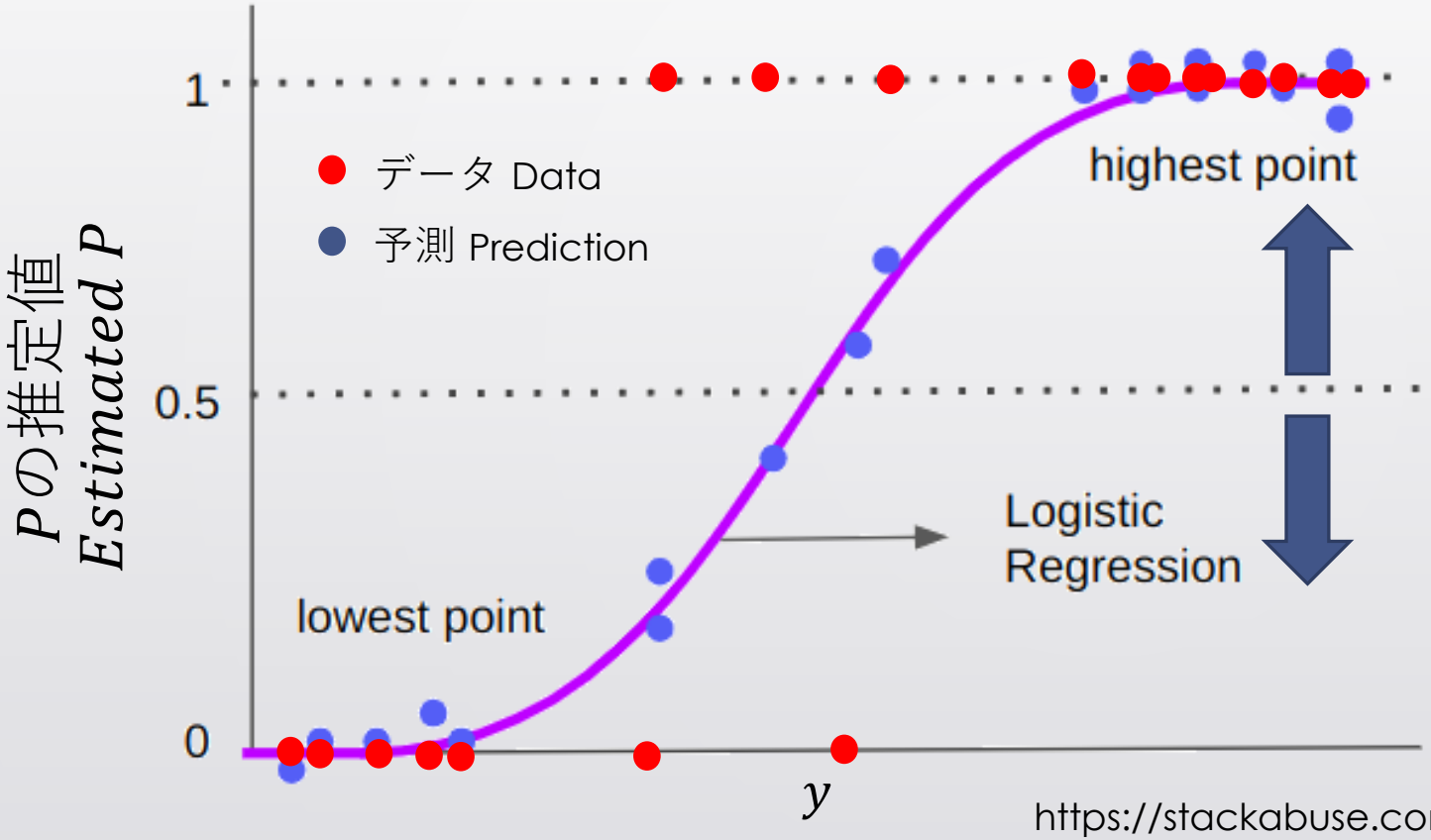
Lを最大化する
Maximize L

$$log(L) = \sum_{i=1}^{N} \{log(P_i)Y_i + log(1 - P_i)(1 - Y_i)\}$$

対数尤度関数
Log-likelihood function

ニュートン・ラフソン法で回帰係数をもとめる

閾値の設定 Setting Threshold



$P \geq$ 閾値ならばデータがクラス１に分類される

Data is classified into Class 1 if $P \geq Threshold$

どのように閾値を設定すればいいか？

How should we set the threshold?

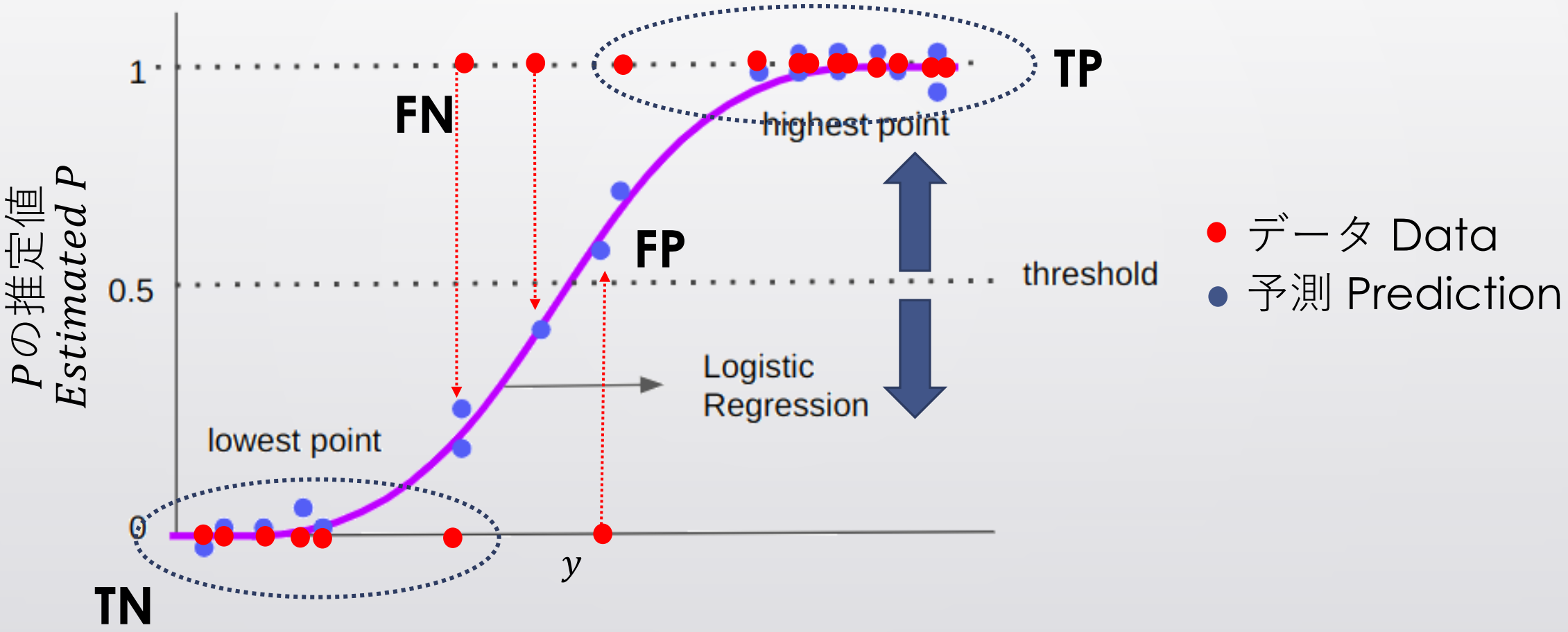
https://stackabuse.com/definitive-guide-to-logistic-regression-in-python/

混同行列 Confusion Matrix

正解 Answer

分類 Classification		クラス1 Class 1	クラス0 Class 0
	クラス1 Class 1	真陽性 (**TP**) True Positive (**TP**)	偽陽性 (**FP**) False Positive (**FP**)
	クラス0 Class 0	偽陰性 (**FN**) False Negative (**FN**)	真陰性 (**TN**) True Negative (**TN**)

ロジスティック回帰と混同行列



分類性能の評価 Evaluation of Classification Performance

正答率

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

True Value = [1, 1, 1, 0, 1, 1, 1, 1, 0, 1]

Prediction = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

$$Accuracy = 0.8$$

分類性能の評価 Evaluation of Classification Performance

感度　　再現率

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

陽性データが正しく陽性と判定される確率

The probability that positive data is correctly classified as “positive”

適合率　　陽性的中率

$$Precision = Positive\ Predicitve\ Value = \frac{TP}{TP + FP}$$

陽性と判定されたデータが実際に陽性である確率

The probability that data classified as “positive” is truly positive

分類性能の評価 Evaluation of Classification Performance

特異度

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

陰性データを正しく陰性と判定する確率

The probability that negative data is correctly classified as “negative”

$$1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP} = \text{偽陽性率}$$

False Positive Rate

分類性能の評価 Evaluation of Classification Performance

$$Sensitivity = Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP}$$

再現率と適合率の間にはトレードオフがある

There is a trade-off between recall and precision

$$F1 = \text{再現率と適合度の調和平均} = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} = \frac{2Recall \times Precision}{Recall + Precision}$$

Harmonic mean of recall and precision

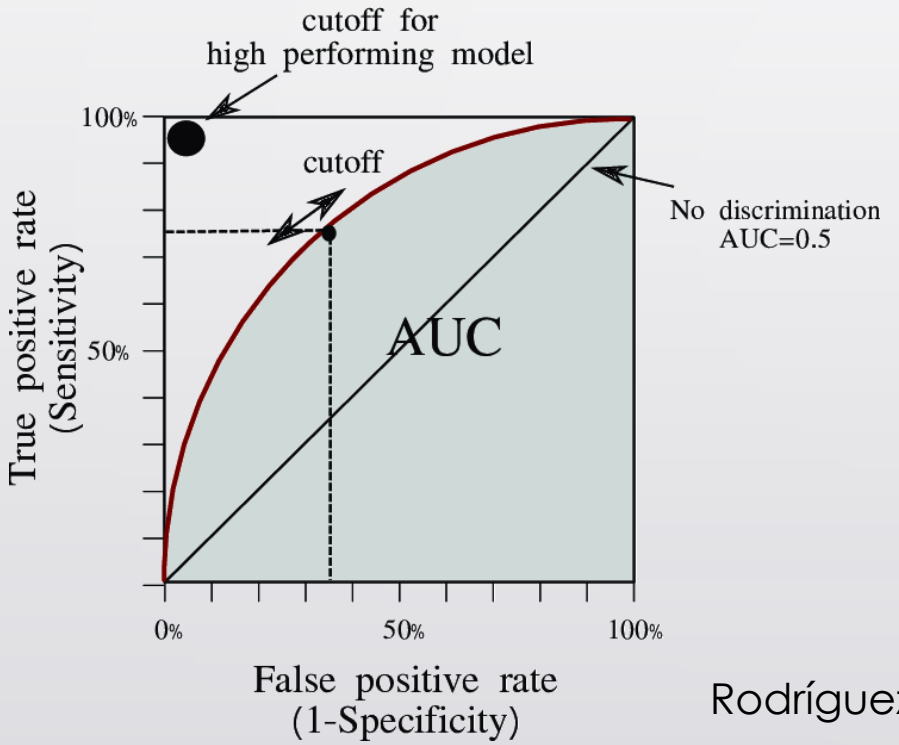
F1値は再現率と適合度のバランスを反映する

F1-value represents balance between recall and precision

ROC曲線 ROC(Receiver-Operator Characteristics) Curve

良い分類器は、感度が高く偽陽性率が低い

A good classifier has high sensitivity and low false positive rate

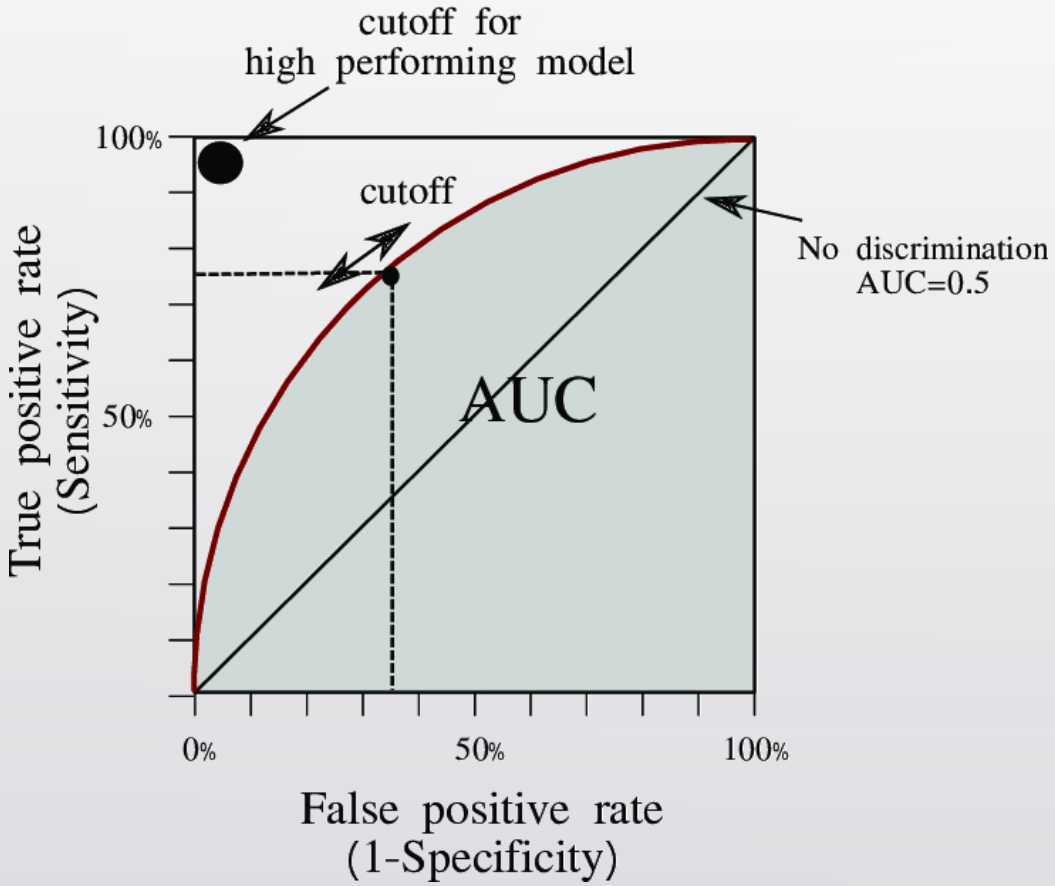


ROC曲線は、異なる閾値における感度・偽陽性率を表す

ROC represents relationship between sensitivity and false positive rate under varying threshold

Rodríguez-Hernández et al, 2021

ROC曲線 ROC(Receiver-Operator Characteristics) Curve



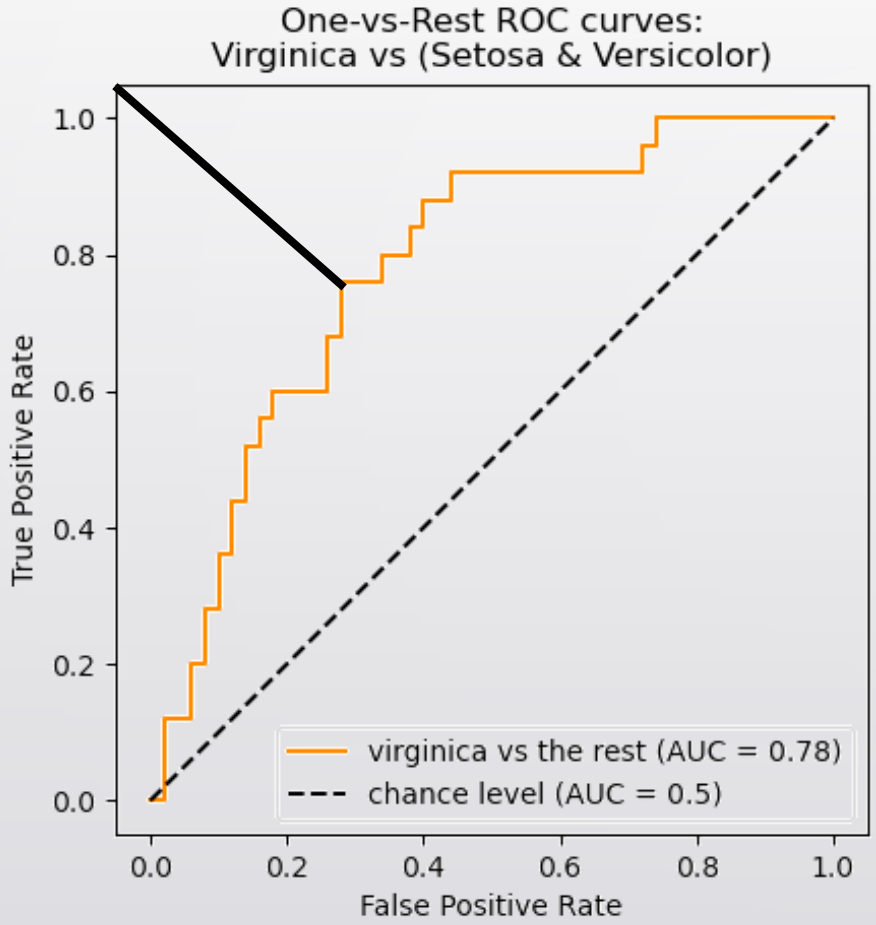
AUC: Area Under Curve

AUCが大きいほど、分類器の性能が良い

Larger AUC indicates better performance of classifier

AUC	
0.9 - 1.0	**High accuracy**
0.9 - 0.7	**Moderate accuracy**
0.5 - 0.7	**Low accuracy**

カットオフの決定方法 How to determine “Cut-Off”

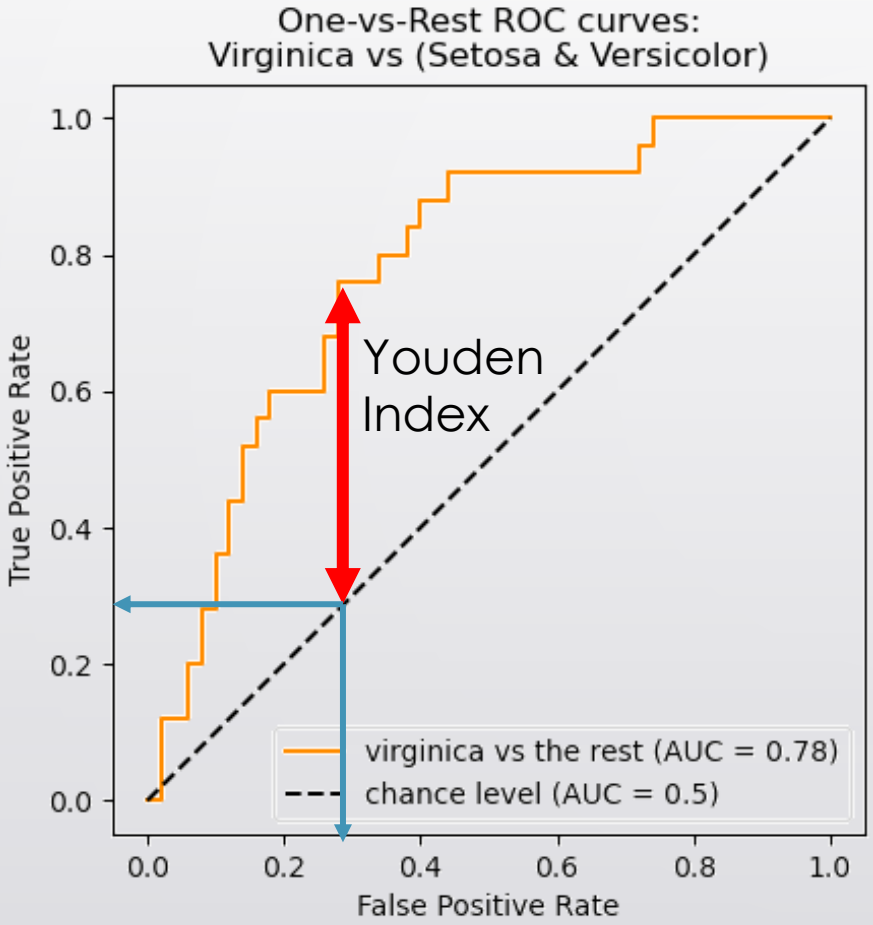


最適な性能との距離が最小になる閾値

The threshold at which distance from the optimal performance is minimized

https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

カットオフの決定方法 How to determine “Cut-Off”



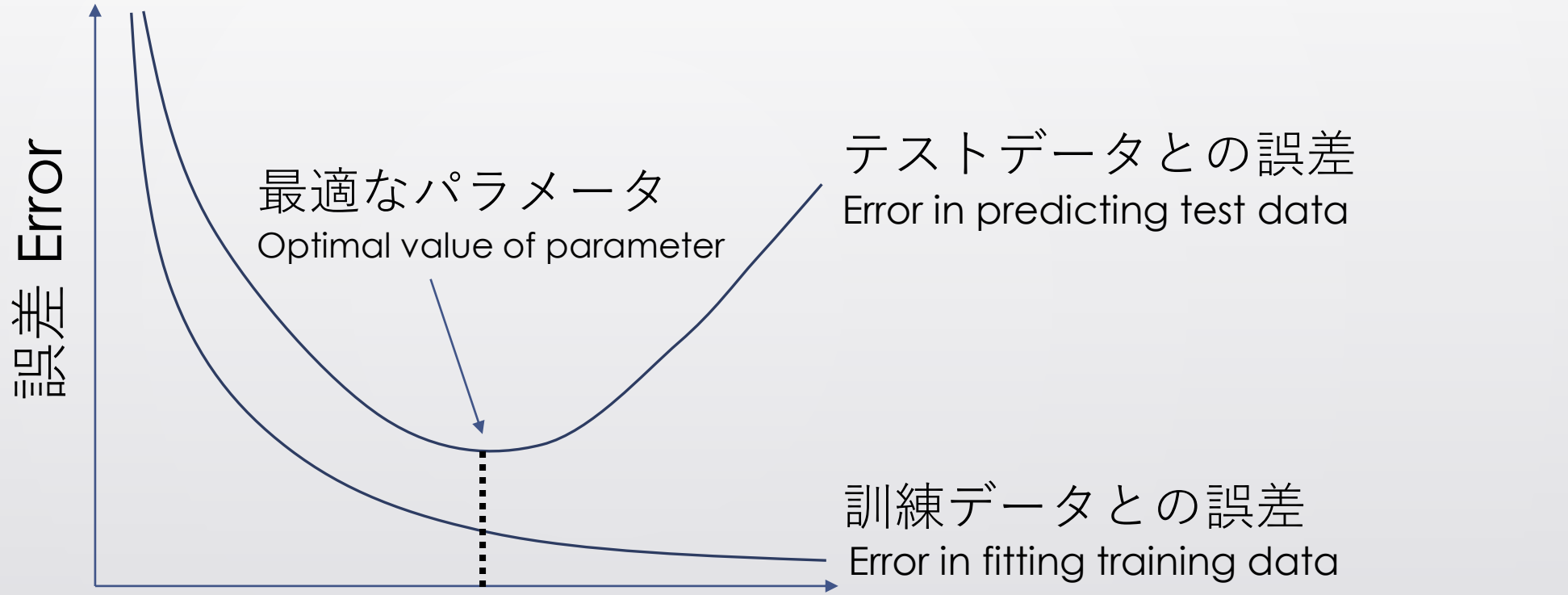
Youden Indexが最大になる閾値

The threshold at which Youden Index is maximized

$$Youden\ Index = Sensitivity - False\ Positive\ Rate$$
$$= Sensitivity - (1 - Specificity)$$
$$= Sensitivity + Specificity - 1$$

https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

過学習のサイン Signs of Overfitting



交差検証 Cross validation

1. データを学習(訓練)データとテストデータに分割する
   Splitting data into training and test data

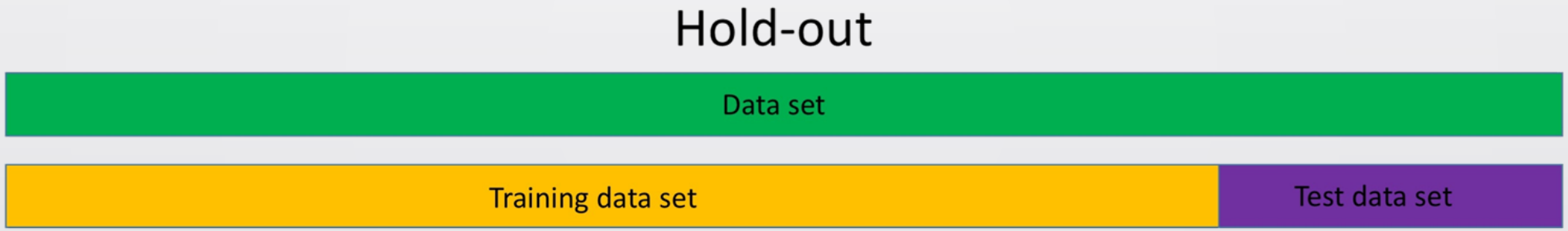
2. 学習(訓練)データを使って分類モデルを作る
   Create classification model based on training data

3. 分類モデルの予測性能をテストデータで検証する
   Evaluate prediction performance of classification model using test data

ホールドアウト法 Hold-out Method

データを一定の比率で学習データとテストデータに分割し性能検証を行う

Evaluate classification performance by dividing the dataset to training/test data with certain proportion



https://qiita.com/ZESSU/items/8aaad3cdfeae35fa0820

k-分割交差検証 *k*-fold cross validation

すべてのデータを学習/テストデータとして使用する

Use all the data as training/test data

$k = sample\ size$の時がLeave-one-out交差検証

Equals to Leave-one-out cross validation when $k = sample\ size$

http://ethen8181.github.io/machine-learning/model_selection/model_selection.html

ロジスティック回帰 Logistic Regression

$$log\frac{P(Y=1)}{1-P(Y=1)} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M$$

$$\frac{P}{1-P} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M}$$

$$P = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_M x_M)}}$$

多クラス分類への拡張 Extension to Multiclass Classification

データが複数に属する可能性がある場合

When a data can belong to multiple classes simultaneously

各クラスごとに回帰モデルを作る

Estimate regression model for each class

$$log\frac{P_{C_k}}{1-P_{C_k}} = \beta_{1,C_k}\,x_1 + \beta_{2,C_k}x_2 + \beta_{3,C_k}x_3 \ldots \beta_{M,C_k}x_M$$

P_{C_k}: データがクラスC_kに属する確率

The probability that data belongs to class C_k

多クラス分類への拡張 Extension to Multiclass Classification

データが一つのクラスのみに属する場合

When a data can be classified into only one class

ソフトマックス関数で各データが観測される確率を計算する

Compute the probability that each data is observed by softmax function

$$f_{C_k} = \beta_{1,C_k} x_1 + \beta_{2,C_k} x_2 + \beta_{3,C_k} x_3 \ldots \beta_{M,C_k} x_M$$

$$P_{C_k} = \frac{exp(f_{C_k})}{\sum_1^K exp(f_{C_i})}$$

K: クラスの総数

Total number of classes