3-6 予測・判断

東京大学　数理・情報教育研究センター
2021年5月7日

概要

- データを用いた予測や予測の種類を学びます.
- 機械学習の予測モデルや，予測された結果の評価方法を学びます.

本教材の目次

予測とは

- 数理モデルを仮定して何かしらの推測することを予測と呼びます.

- 機械学習モデルによる予測
  - すでにあるデータから数理モデルを学習した後で,
    新たなデータに対する出力を数理モデルから予測します.
  - 線形回帰，ランダムフォレスト，サポートベクタマシン,
    ニューラルネットワークなどの，教師あり学習モデルを使用します.
- 時系列モデルによる予測
  - 現在までの時系列の変動から，未来の変動を予測します.
- 数値シミュレーションによる予測
  - 自然現象など観測の難しい物理現象を
    仮定した数理モデルからコンピュータ上での計算で予測します.

機械学習モデルによる予測(1)

- クラス分類…質的データの予測
  - 2クラス分類：はい／いいえ，陽性／陰性など
  - 多クラス分類：動物の種類，文字，人物など
- 回帰…量的データの予測
  - 価格，人数，重量，温度，体積など

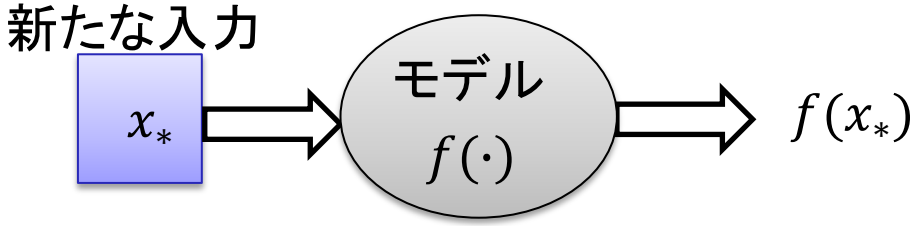
- クラス分類と回帰で，使用する機械学習モデルや評価基準が変化します．
- 一般化線型回帰モデルは，線形回帰モデルを回帰以外にも拡張したものです．この場合，クラス分類はロジスティック回帰，人数や回数などカウントデータの予測はポアソン回帰と呼ばれます．

機械学習モデルによる予測(2)

- モデル$f(\cdot)$に入力xを入力したときの$f(x)$が
  予測したい変数yに近づくようにモデル$f(\cdot)$を学習します.
- xは入力変数，説明変数，独立変数
  yは出力変数，被説明変数，従属変数　とそれぞれ呼ばれます.

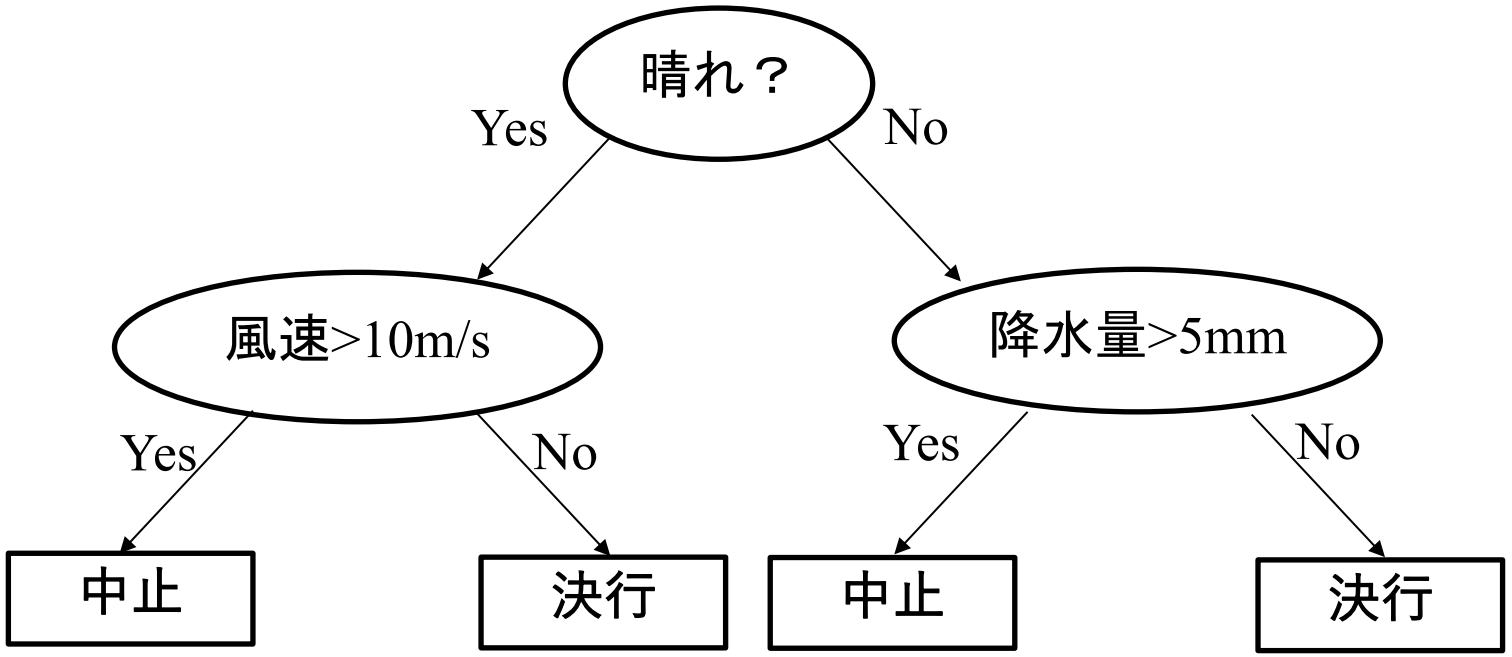
学習時

予測時



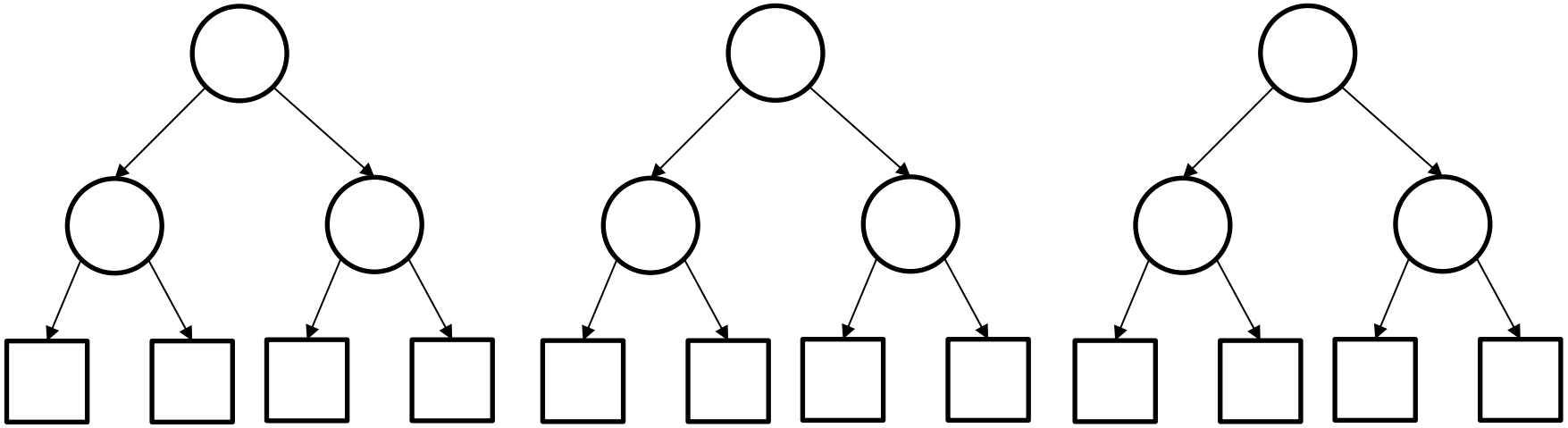
決定木

- 決定木は，Yes/Noを返す質問を繰り返す二分木の木構造によって予測を行う機械学習モデルです.
- どのような質問をすればよいかを学習します.



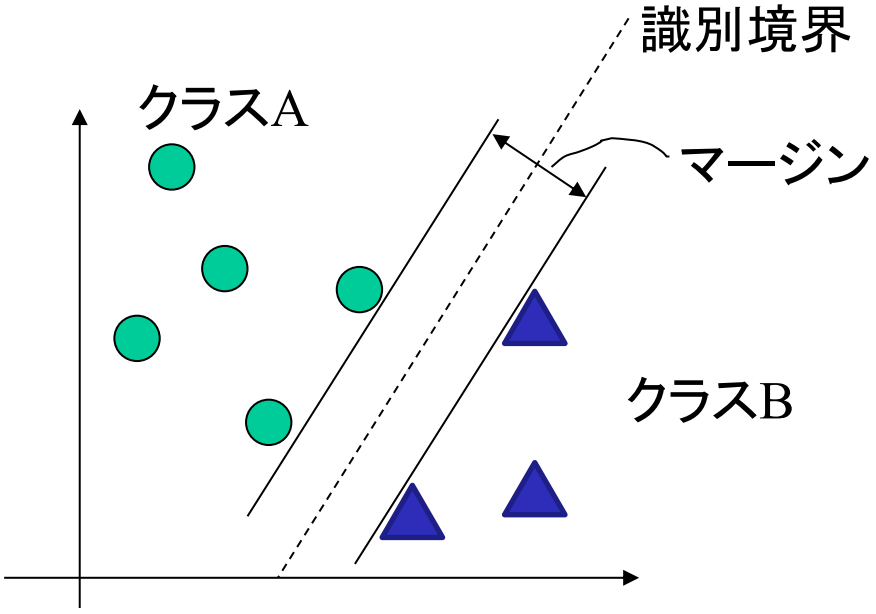
ランダムフォレスト

- 決定木は解釈性が高いものの，性能はあまり高くありません.
- ランダムフォレストは複数の決定木（いわゆる“森”）を組合せることで推定精度を向上させる手法です.
- 複数の弱い機械学習モデル（弱学習器）を組み合わせる手法は**アンサンブル学習**と呼ばれ決定木以外でも広く使用されています.



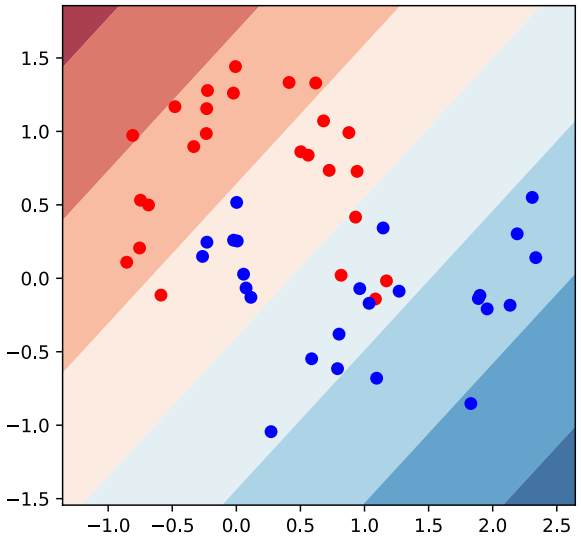
サポートベクタマシン

- サポートベクタマシン（Support Vector Machine; SVM）は
  データの代表点であるサポートベクタによって予測の識別境界を
  決定する手法です.
- 異なるクラス間の距離（マージン）が最大になるように
  学習を行います.
- 識別境界が直線で表されるSVMは線形SVMと呼ばれます.

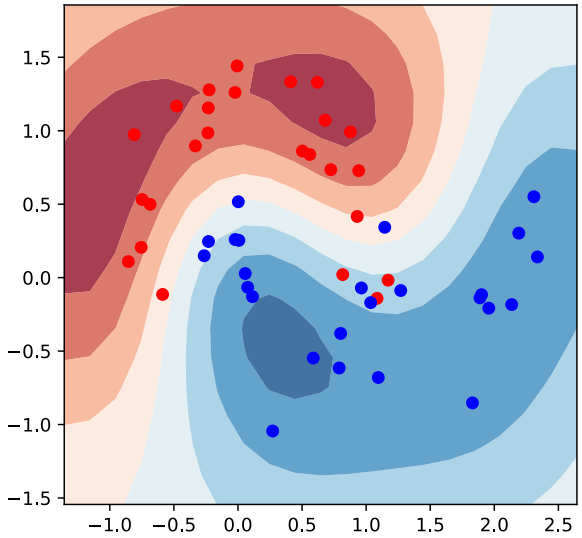


カーネルSVM

- 線形SVMは分離境界が直線になるため，データによっては必ずしもうまく予測ができません.
- カーネルSVMは，データ点同士の類似度を表すカーネル関数を用いる手法で，複雑な識別境界を実現可能です.



線形SVM

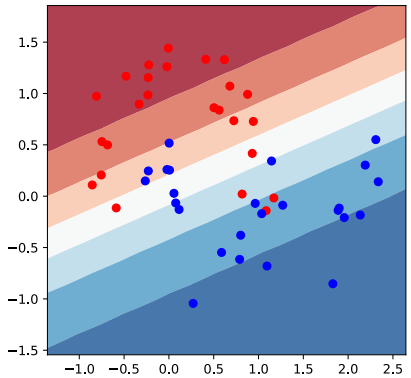


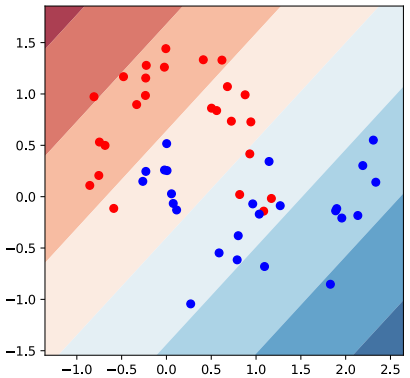
RBFカーネルSVM

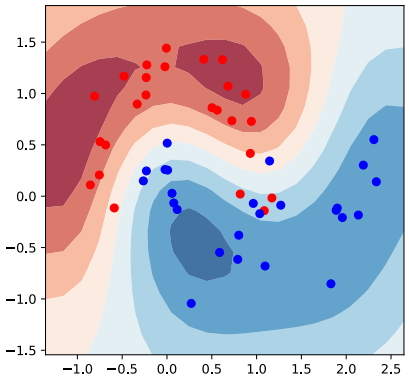
2次元入力変数を2クラス（赤or青）で2クラス分類した例．赤い領域は赤のクラスに，青の領域は青のクラスに分類されます.

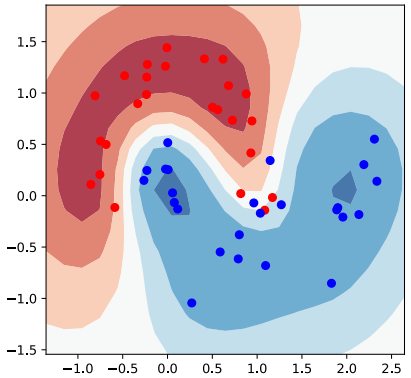
2クラス分類の手法の比較

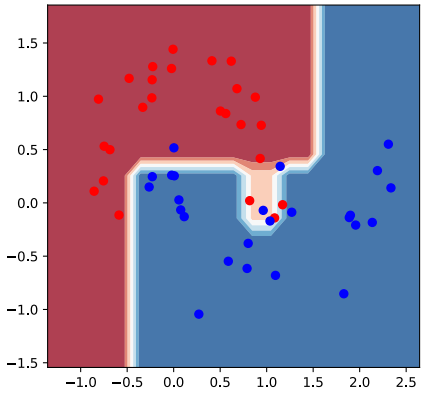
- 予測手法ごとに，得られる識別境界は大きく異なります．
  そのため，データに合った予測手法の選択が重要です．(参考)

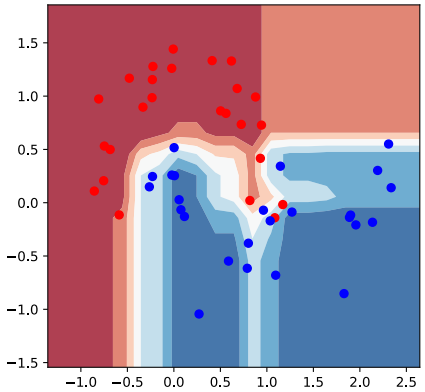

線形ロジスティック回帰(1-4参照)

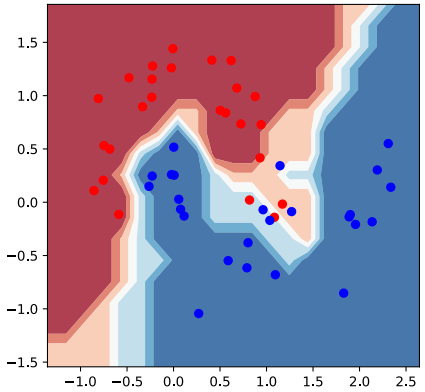

線形SVM

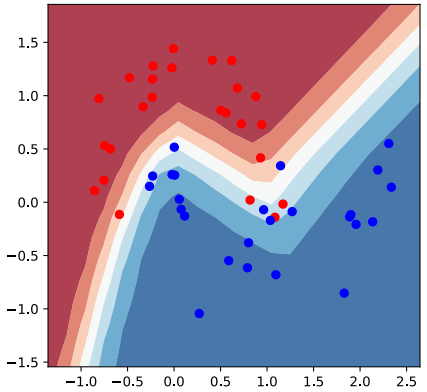

カーネルSVM


ガウス過程による分類


決定木

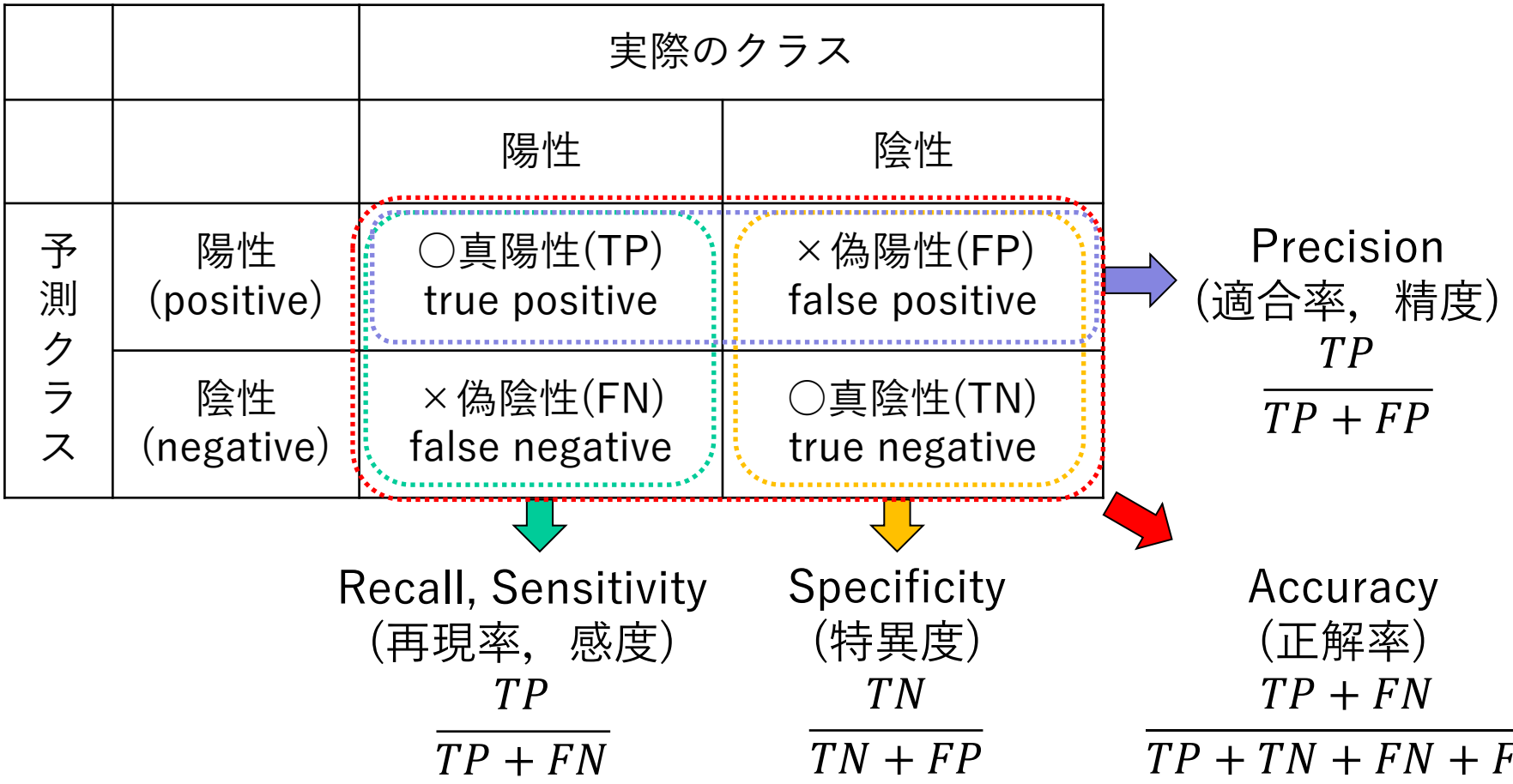

ランダムフォレスト


k-近傍法


ニューラルネット(3-4 参照)

2クラス分類の評価

- 2クラス（陽性／陰性）の分類の評価はRecallやPrecisionなど複数の指標によって評価されます.



ACCURACY/RECALL/PRECISION

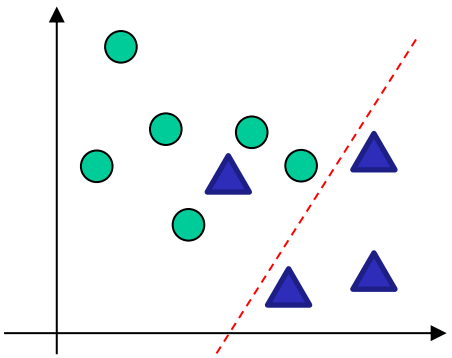
- Accuracy
  - 単純に予測モデルがどの程度正解できるかを表す指標です.
    陰性が珍しい場合などバランスが異なるときには適していません.
- Recall(Sensitivity, 再現率, 感度)
  - 実際の陽性を，どれだけ検出できるかを表す指標です.
- Specificity(特異度)
  - 実際の陰性を，どれだけ検出できるかを表す指標です.
- Precision(適合率，精度)
  - 陽性と判断されたうち，どれだけが正解かを表す指標です.
- F値とはトレードオフの関係にあるRecallとPrecisionを
  まとめて評価する指標で次の式で定義されます.

$$F = 2\frac{\mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}}$$

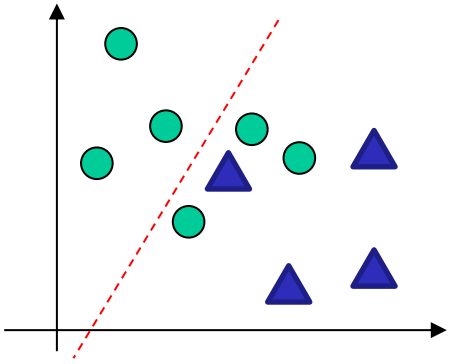
ROC曲線／AUC (1)

- RecallとPrecisionのどちらが有効かはシステムに依存します.
  - 病気を検出する場合，病人を取りこぼさないことを重要視したいときには，Recallを重視します.
- 予測モデルによっては，陽性／陰性の判断基準の調整が可能で，Recall・Precisionのどちらを重要視するか選べます.

Recallを重視した
識別境界



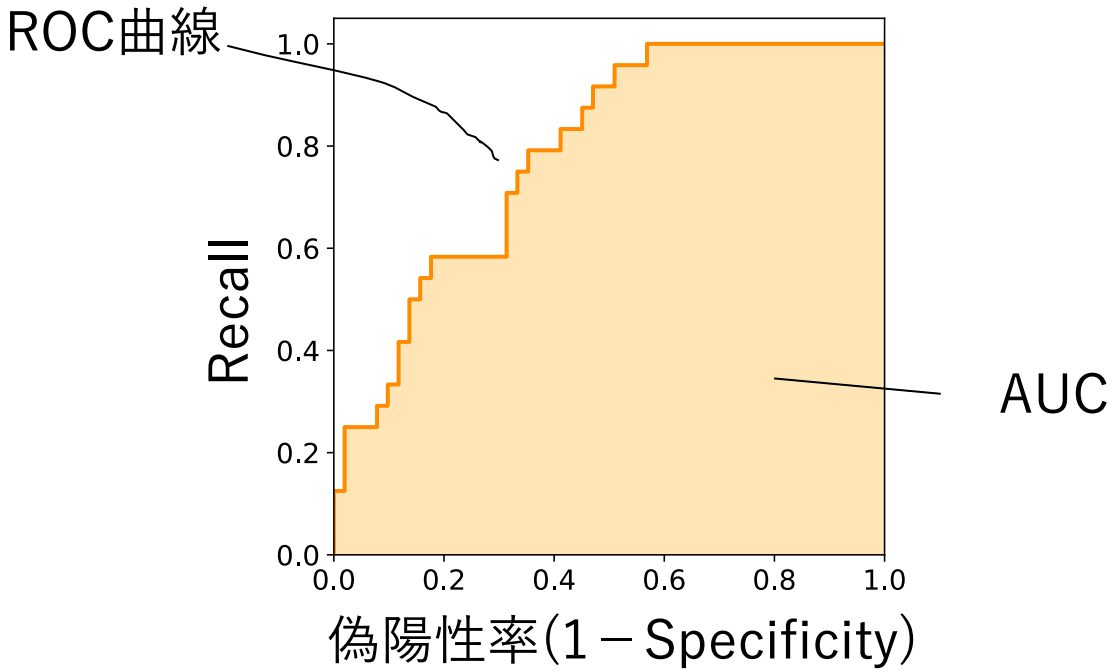
Precisionを重視した
識別境界



陽性
陰性

ROC曲線／AUC (2)

- ROC曲線(receiver operating characteristics curve)は，偽陽性率(=1－Specificity)を横軸，Recallを縦軸にとって陽性/陰性の判断基準を変えたときの変化を表す曲線です．
- AUC(area under curve)は，ROC曲線とx軸の間の面積で，値が1に近いほど，予測モデルの性能が高いと判断できます．



多クラス分類の評価

- 多クラス分類ではAccuracyの他に，混同行列(confusion matrix)を用いた評価も有効です．
- 混同行列によって予測の誤りの傾向の調査が可能です．

実際のクラス

予測されたクラス	A	B	C	D
A	25	10	5	2
B	3	13	3	9
C	0	3	20	4
D	2	4	2	15

回帰問題の評価

- 実際の数値と予測した数値の誤差を調べることによって，
回帰モデルの性能評価が可能です．
- 実際のデータy_iと予測した値$\overline{y_i}$の差の二乗に対して，
データ全体で平均をとったものを，
平均二乗誤差(mean squared error; MSE)と呼びます．

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2$$

- 平均絶対誤差(mean absolute error; MAE)を使用することもあります．

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \overline{y_i}|$$

- MSEはMAEと比較して，大きい予測誤差の影響を受けやすい
特徴があります．

数値シミュレーションによる予測

- 数値シミュレーションは，現実に起こりうる問題に対して，あるモデルを仮定し，数値計算によってモデルから予測を行う手法です.
- 離散型シミュレーションは、ATMなどの混雑状況をコンピュータの数値計算によって予測する手法です。
  - 人の到着確率および処理の終了時間の分布によって定義される待ち行列モデルに対して，乱数を実際に与えることで，待ち行列の長さが予測可能です.
- 連続型シミュレーションは，複雑な物理現象など微分方程式で与えられるモデルに対して，数値計算によって現象を予測する手法です.

気象予測

- 気象予測は気象学から得られる物理モデルや，観測データに基づく機械学習モデルによって行われます.
- 大気の遷移などを表した数値予報モデルから
  未来の大気の状態を数値的に予測することができます.
- さらに，大気の状態と日照量や降水量などを関連付けた
  過去のデータから機械学習モデルを学習することで,
  未来の気象を予測することができます.
- 物理モデルで予測した空間的に疎な背景値と，実際の観測値との差分を用いて，モデルの修正を行う手法をデータ同化と呼びます.