機械学習 第２回 機械学習の基本的な手順

立命館大学 情報理工学部

村上 陽平

Beyond Borders

講義スケジュール

- 担当教員１：村上、福森（第1回～第15回）

1	機械学習とは、機械学習の分類
2	**機械学習の基本的な手順**
3	識別（１）
4	識別（２）
5	識別（３）
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク
9	深層学習
10	アンサンブル学習
11	モデル推定
12	パターンマイニング
13	系列データの識別
14	強化学習
15	半教師あり学習

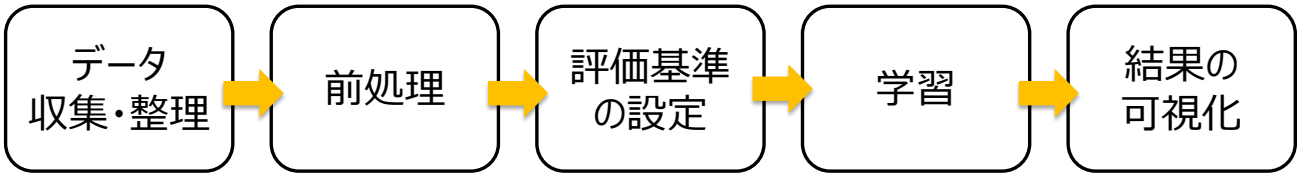
- 担当教員２：叶昕辰先生（第16回の講義を担当）

今回の講義内容

- 機械学習全体の手順
  - データ収集・整理
  - 前処理
    - 次元削減、標準化
  - 評価基準の設定
    - 分割学習法、交差確認法
  - 学習
    - 例としてk-NN法を紹介
  - 結果の可視化
    - 混同行列、F値、ROC曲線
- 演習問題

機械学習の流れ

- 機械学習の流れは、大きく分けて5つの処理で構成
  1. データ収集・整理
  2. 前処理
  3. 評価基準の設定
  4. 学習
  5. 結果の可視化



データ収集・整理

- データの収集例
  1. 予め存在するデータを使用
  2. 自分でタスクと問題を設定して、そのために必要なデータを収集
     - 教師あり学習の場合、さらに正解の付与作業が必要

- 第１回で述べた通り、**機械学習に用いる学習データは多次元ベクトル**
  - 多次元ベクトルの集合を機械可読な形式で表現する最も簡単な方法
    - CSV（Commna Speparated Values）形式
      - ベクトルの各要素をカンマで区切り、1行に1事例ずつ並べる

前処理

- 機械学習アルゴリズムに収集したデータを用いる前に、そのデータに何かしら手を加えること
  - 機械学習の性能を向上させるために重要な処理

- 本講義で紹介する前処理の手法
  - 次元削減
  - 標準化

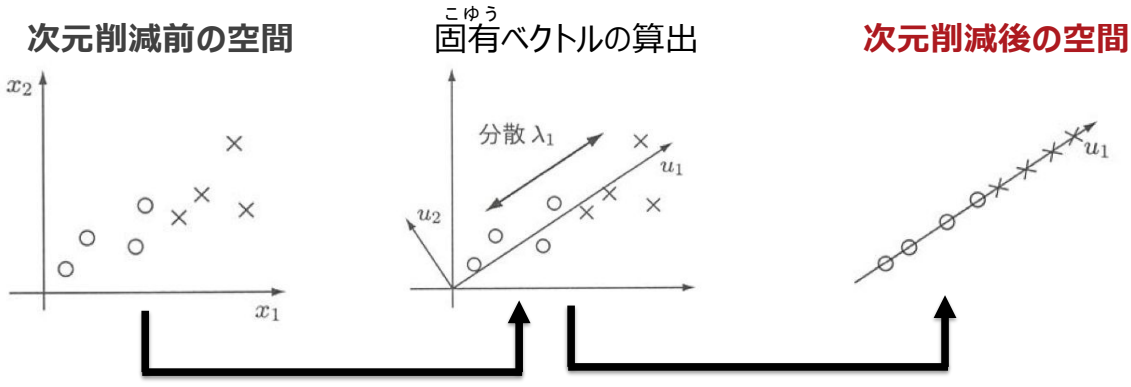
前処理：次元削減

- 次元数が増えると…
  - 高次元空間上に、学習データが疎らに存在することになり、そのようなデータから得られたモデルの汎化能力が低い
    - これを「次元の呪い」と呼ぶ。
    - **汎化能力**（または汎化性能）
      学習データにない入力に対して、いかに良い結果を出力できるか？
      学習データから、いかに一般化されたモデルが獲得されているか？

- **次元削減**（または次元圧縮）
  - 特徴ベクトルの次元数を減らすこと
  - 汎化能力の高いモデルを学習する上で重要な前処理

前処理：次元削減

- 主成分分析（Principal Component Analysis; PCA）
  - 相関が高い特徴を含むような冗長な高次元空間を、冗長性の少ない低次元空間に写像する行列を求める操作



STEP1
共分散行列Σの計算

STEP2
固有値と固有ベクトルの計算

STEP3
固有ベクトルを用いて次元削減

前処理：次元削減

- STEP1：共分散行列の計算
  - 特徴空間上におけるデータの散らばり具合を
    データの統計的性質を表す共分散行列を用いて表現

学習データ $\{\boldsymbol{x}|\boldsymbol{x} \in D\}$の共分散行列$\boldsymbol{\Sigma}$の計算式

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{\boldsymbol{x}\in D}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T$$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{\boldsymbol{x}\in D}\boldsymbol{x}$$

D：学習データの集合
N：Dの要素数
$\boldsymbol{x}$：学習データ
$\boldsymbol{\mu}$：Dの平均ベクトル

前処理：次元削減

- STEP1：共分散行列の計算（つづき）
  - ２次元データの場合、平均ベクトル $\boldsymbol{\mu} = (\overline{x_1}, \overline{x_2})^T$とすると
    共分散行列$\boldsymbol{\Sigma}$は、以下の通りとなる。

$$\boldsymbol{\Sigma} = \frac{1}{N}\begin{pmatrix} \sum_{\boldsymbol{x}\in D}(x_1-\overline{x_1})^2 & \sum_{\boldsymbol{x}\in D}(x_1-\overline{x_1})(x_2-\overline{x_2}) \\ \sum_{\boldsymbol{x}\in D}(x_1-\overline{x_1})(x_2-\overline{x_2}) & \sum_{\boldsymbol{x}\in D}(x_2-\overline{x_2})^2 \end{pmatrix}$$

※ 対角成分：次元ごとの分散（データの散らばり具合）
※ 非対角成分：次元間の相関（共分散）

前処理：次元削減

- STEP2：共分散行列の固有値と固有ベクトルの計算
  - 共分散行列$\boldsymbol{\Sigma}$は
    - 半正定値：固有値がすべて0以上の実数
    - 対称行列：固有ベクトルが実数かつ直交

    であるため、以下のように分解できる。

$$\boldsymbol{\Sigma}' = \boldsymbol{U}^T \boldsymbol{\Sigma} \boldsymbol{U} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

λ_1, λ_2：固有値（$\lambda_1 > \lambda_2$）

$\boldsymbol{U}$：それぞれの固有値に対応する固有ベクトル$\boldsymbol{U}_1, \boldsymbol{U}_2$を並べた行列 $[\boldsymbol{U}_1, \boldsymbol{U}_2]$

  - **固有値の大きい順に、それに対応する固有ベクトルの方向**
    - データの散らばりが大きい方向
    - 言い換えると、識別するにあたって情報が多い方向

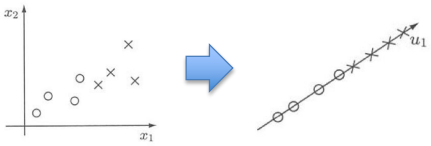
前処理：次元削減

- STEP3：固有ベクトルを用いて次元削減
  - 2 次元空間から 1 次元空間に次元削減

$$u_1 = \boldsymbol{U}_1^T \boldsymbol{x}$$

次元削減後の特徴量（1次元）

次元削減前の特徴ベクトル（2次元）



次元削減前　　次元削減後

  - 固有ベクトル同士は直交するので、固有値の大きい順に軸とすると、特徴空間を構成できる。
  - 上位n位までを用いるとn次元空間が構成される。
    - これらは、元々の高次元特徴空間のデータの散らばりを最もよく保存したn次元空間となる。

前処理：次元削減

- STEP3：固有ベクトルを用いて次元削減（つづき）
  - 特徴空間の次元数を削減すると
    - 学習において推定するべきパラメータ数が少なくなるので、学習結果の信頼性が向上する
    - **過度に次元を削減すると、もとのデータ情報が大きく損なわれる**
  - 削減後の次元数nは、累積寄与率をもとに調整すると良い
  - **累積寄与率**
    - 「すべての固有値の和」に対する「採用した軸の固有値の和」の比
    - 主成分分析によって構成した軸では、対応する固有値が分散になるので、累積寄与率によって「次元削減後の空間が、もとのデータの情報をどの程度保存しているのか」を表現できる。
      - 2次元→1次元の場合：寄与率 $= \lambda_1/(\lambda_1 + \lambda_2)$

前処理：標準化

- 標準化（standardization）
  - 特徴の値の範囲を揃える操作
  - 特徴は各次元で独立の基準で計測・算出するので、その絶対値や分散が大きく異なることがある。
  - これを**ベクトルとして組み合わせて、そのまま学習を行うと、絶対値の大きい特徴量の寄与が大きくなりすぎる**ので、事前に各次元での値のスケールを合わせる必要がある。
  - 下式に従って、各次元の平均値を0に、標準偏差を1に揃えるのが一般的

$$\text{標準化後の値} = \frac{\text{もとの値} - \text{その次元の平均値}}{\text{その次元の標準偏差}}$$

演習問題2-1（10分間）

- 以下5名分の身長と体重のデータについて、身長と体重の平均値を0、標準偏差を1となるように標準化せよ。

標準化前

番号	身長 [cm]	体重 [kg]
1	160	54
2	166	58
3	168	60
4	172	62
5	184	66
平均		
標準偏差		

標準化後

番号	身長	体重
1		
2		
3		
4		
5		
平均	0	0
標準偏差	1	1

【ヒント】標準化の方法

$$\text{標準化後の値} = \frac{\text{もとの値} - \text{その次元の平均値}}{\text{その次元の標準偏差}}$$

評価基準の設定：分割学習法

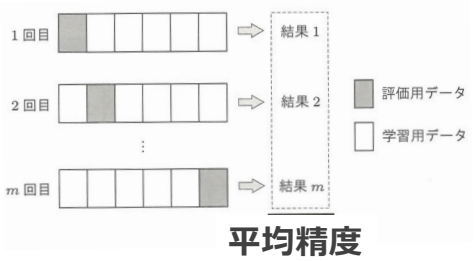
- 学習結果の評価基準
  - **未知データに対してどれだけの正解率が期待できるか？**がポイント
    - 学習データに対して100%では意味がない
- 未知データの評価方法
  - **分割学習法**
    - 半分を学習用、残り半分を評価用として分割する方法
    - モデルのパラメータを調整するときは
      データを学習用・調整用・評価用として分割する場合もある
    - 学習データが大量にある場合に有効
      - 評価用データが少ないと、未知データの分布と全く異なる可能性が高くなり、評価そのものが信頼できなくなる

評価基準の設定：交差確認法

未知データの評価方法（つづき）

■ 交差確認法（Cross Validation method; CV法）

- 学習データをm個の集合に分割し、そのうちの$m-1$個で学習し、除外した残りの一つで評価する。
  - mは交差数。例えば、$m=10$の場合は「10-fold CV」と表記する。
  - mがデータ個数の場合は、一つ抜き法（leave-one-out method）と呼ぶ。
- 除外するデータを順に交換することで、合計m回の学習と評価を行う。
  - 全データが一通り評価に使われ、かつその評価時に用いられる識別器は評価用データを除いて構築されたものとなっている。



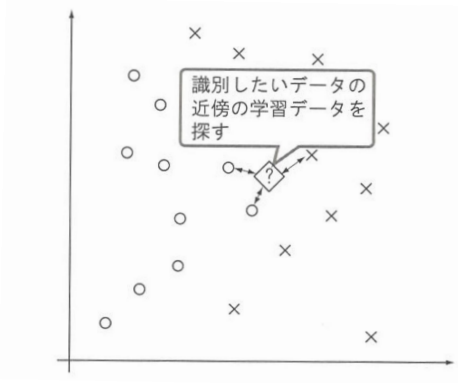
← 交差確認法

学習

k-NN法（k-Nearest Neighbor method）

■ 入力されたデータに近い学習データを近い順にk個選び、多数決などで所属クラスを決定する手法

- 音声対話アプリで実現されている発話理解手法の一部にk-NN法の考え方に近いものが採用されている



K-NN法による識別

K-NN法で調整するべきパラメータ

1. 近傍として探索するデータ数 k
2. 距離尺度
   - 通常は、以下のユークリッド距離を用いる。

$$\mathrm{Dist}(\boldsymbol{x},\boldsymbol{x}') = \sqrt{\sum_{i=1}^{d}\left(x_i - x_i'\right)^2}$$

3. 探索方法
   - 通常は、**入力と全データとの距離を計算**して並べ替える。データが多いときは、事前にデータを木構造化し、効率よく探索する場合もある。

結果の可視化

- 結果の可視化とは、識別結果から
  いくつかの評価指標を計算し、表やグラフとして表示すること
  - 今回は、2クラス識別問題の評価法
    （入力データが、ある概念に当てはまるか判定する）を考える
    - 例えば、受信したメールが迷惑メールなのか判定
  - 正例（positive）と負例（negative）
    - 正例：設定した概念に当てはまる学習データ
    - 負例：設定した概念に当てはまらない学習データ

結果の可視化：混同行列

- 混同行列（confusion matrix）
  - 以下の要素で構成された表
  - 分割表（contingency table）とも言う。
  - 対角成分が正解数、非対角成分が間違いの数を示す

混同行列

	予測＋ （識別器が正と判定したもの）	**予測－** （識別器が負と判定したもの）
正解＋ （正例）	**true positive (TP)** 正例に対して識別器が正（positive）であると正しく(true)判定した数	**false negative (FN)** 正例に対して識別器が負（negative）であると間違って（false）判定した数
正解－ （負例）	**false positive (FP)** 負例に対して識別器が正（positive）であると間違って（false）判定した数	**true negative (TN)** 負例に対して識別器が負（negative）であると正しく（true）判定した数

※ 前の語が判定の正否（true or false）
※ 後の語が判定結果（positive or negative）

結果の可視化：評価指標

	予測＋	予測－
正解＋	true positive (TP)	false negative (FN)
正解－	false positive (FP)	true negative (TN)

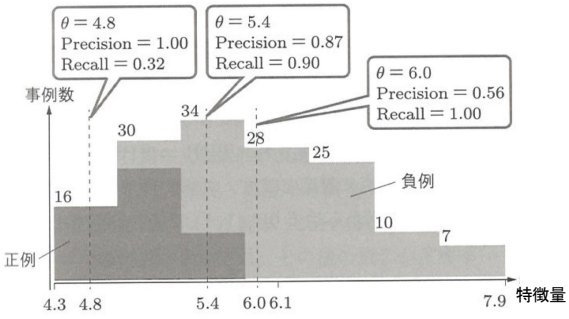
- 評価指標
  - 正解率（Accuracy）
    - $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}}$ ⇒ 識別器が正しい答えを出した割合
  - 精度（precision） ※ 適合率とも呼ぶ
    - $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$ ⇒ 識別器が正と判断したときに、どれだけ信頼できるか？
  - 再現率（recall）
    - $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$ ⇒ 正例がどれだけ正しく判定されているか？
  - F値（F-measure）
    - $\text{F-measure} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$ ⇒ 精度と再現率の調和平均

結果の可視化：評価指標

- 以下４つの評価指標を評価する意味とは？
  - 正解率（Accuracy）
  - 精度（precision）
  - 再現率（recall）
  - F値（F-measure）
- 機械学習の評価は、正解率を算出するだけでは不十分
  - 例えば、**正例に比べて、負例が大量にあるデータを考える。**
    - 「ウィルス感染者（正例）or 否（負例）」のデータを取り扱う場合など
    - 何も考えずに全て感染していないと判定すると、正解率は相当高い。
      - これではウィルス感染者を正しく検出できないのに、良い機械学習アルゴリズムと判断される。
  - 正解率以外の指標も使って、アルゴリズムを評価することが大事

結果の可視化：精度と再現率の関係

- 精度と再現率はトレードオフの関係
  - ある特徴量を用いて閾値θを設定し、入力がθより小さければ正例と判定する識別器を考える。
    - 精度と再現率の両方が1となる閾値θはない。
    - θが小さい（例：$\theta = 4.8$）：精度は1だが、再現率が低い。
    - θを高い（例：$\theta = 6.0$）：再現率は1だが、精度が低い。



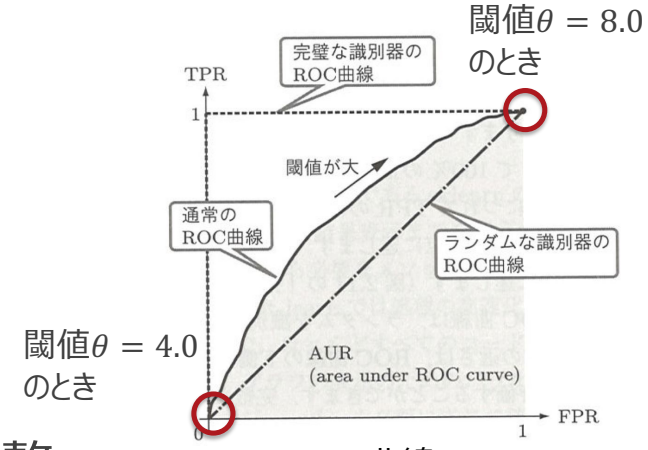
ある特徴量による識別

一般的には、タスクによって、精度と再現率の一方を重視してパラメータを設定する。

重視するものがなければ、F値で性能を測定するのが妥当。

結果の可視化：ROC曲線

- ROC曲線（Receiver Operating Characteristic curve）
  - 閾値θを変えたときの、FPRとTPRをプロットした図
  - FPR（false positive rate）
    - $\mathrm{FPR} = \dfrac{\mathrm{FP}}{\text{負例数}}$
  - TPR（true positive rate）
    - $\mathrm{TPR} = \dfrac{\mathrm{TP}}{\text{正例数}}$



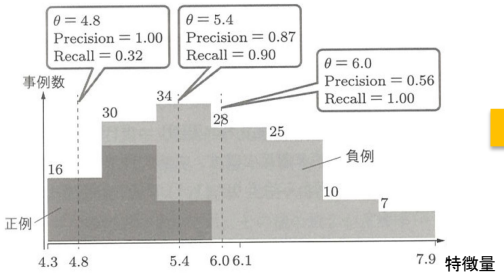
ROC曲線

- ROC曲線を用いた結果の調整
  - 閾値θを変えたときの精度と再現率の関係を見れば、タスクで要求される適切な閾値θを設定できる。

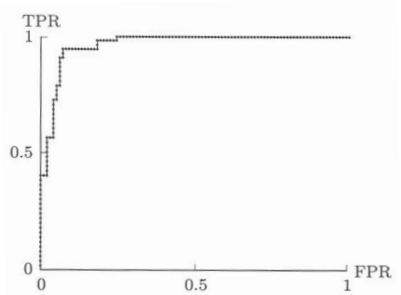
結果の可視化： ROC曲線

ROC曲線の描き方

- 2ページ前のスライドの識別器の閾値θについて考える
  - $\boldsymbol{\theta < 4.3}$ で正例とすると、**全てのテストデータが負**と判定される。
    - **TPR = FPR = 0** なので、**ROC曲線の原点（0, 0）**に対応する。
  - $\boldsymbol{\theta > 8.0}$ で正例とすると、**全てのテストデータが正**と判定される。
    - **TPR = FPR = 1** なので、**ROC曲線の（1,1）**に対応する。
  - θを4.3から8.0に小刻みに変化させると、右図のROC曲線となる。



（再掲）ある特徴量による識別

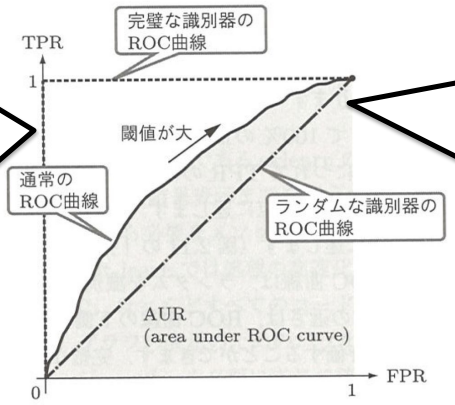


閾値θのみで判定する識別器のROC曲線

結果の可視化：ROC曲線

完璧な識別器のROC曲線

原点から出発し、θを大きくするとTPRの軸に沿って上昇し(0,1) の点(FPR = 0, TPR = 1)の理想的な点に達します。



ランダムに正負を出力する識別器のROC曲線

正負の出力の割合を変えることで、ROC曲線は原点と(1,1)を結ぶ直線となる。

通常の識別器に対するROC曲線

- 上記の「完璧な識別器」と「ランダムな識別器」の間に存在
- 完璧な識別器までの近さは、**ROC曲線の下側の面積**（**area under ROC curve; AUR**）で評価する。
  - 完璧な識別器はAUR = 1、ランダム識別器はAUR = 0.5なので AURが1に近いほど、よい識別器である。

結果の可視化：多クラス識別

- 多クラス識別の評価方法
  1. マクロ平均
     - **クラスごとの精度や再現率を求め、その平均**を計算する。
  2. マイクロ平均
     - **各クラスでの混同行列**を作成し、それらを集計（しゅうけい）する。

  - 各クラスの事例数の違いが大きいときは、マイクロ平均または**事例数で重（おも）みをつけた**マクロ平均を用いる。

演習問題2-2（10分間）

- ある画像認識技術でバナナを検出することを考える。
  - 100個のリンゴ、30個のバナナ、70個のオレンジの画像データに対して、以下の識別結果が得られた。
  1. 識別結果から混同行列を求めよ。
  2. 混同行列から正解率、精度、再現率、F値を計算せよ。また、この認識技術の性能について考察（こうさつ）せよ。

	識別結果	
	バナナ	バナナ以外
りんご	20	80
バナナ	15	15
オレンジ	10	60

	予測＋	予測－
正解＋		
正解－		

正解率

精度

再現率

F値