機械学習とデータマイニングの基礎（大阪大学）

基本パーツを俯瞰する

Matthew J. Holland

大阪大学 産業科学研究所

大阪大学
OSAKA UNIVERSITY

学習問題から始める

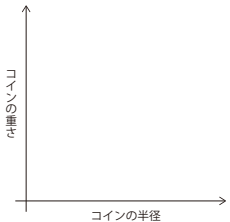
必ず，なんらかの「問い」から出発する．

（例） 自販機のセンサーだけで，
コインの識別はできないか？

まずは，具体例を図示しながら，重要な概念と用語をひと通り見ていく．[1]

[1] 色付けの凡例：用語，単なる強調，スライド中の区切り等．

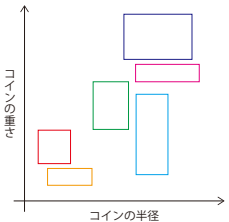
学習問題から始める



- 計測（記号化）ができる
- 学習問題とは関係がありそう

上記を満たせば特徴量と呼ぶ．

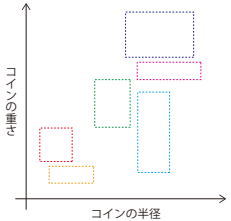
学習問題から始める



特徴量と学習問題の関係が
完全にわかっている．

→ そもそも学習は要らない．

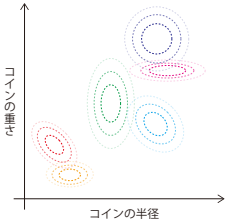
学習問題から始める



規則性があるのは知っているが，
正確にはわかっていない．

→ 学習する意義がある．

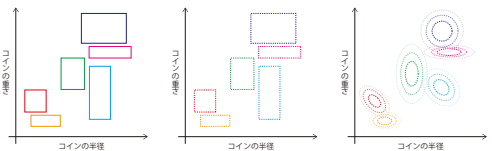
学習問題から始める



規則性はあるだろうけど，
ほとんど何もわかっていない．

→ 学習するしかない．

学習問題から始める



学習問題によって，生じるノイズが大きく異なる．

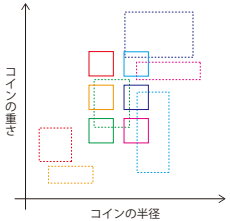
機械学習の基本的なパーツ

- モデル
- データ
- 学習アルゴリズム
- 評価
- 汎化性能

機械学習の基本的なパーツ

- モデル
- データ
- 学習アルゴリズム
- 評価
- 汎化性能

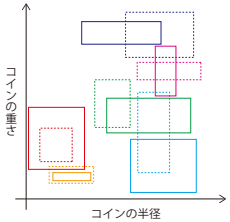
モデルの概念



識別課題における一つの候補.

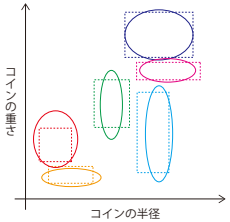
すべての候補の集合をモデルと呼ぶ.

モデルの概念



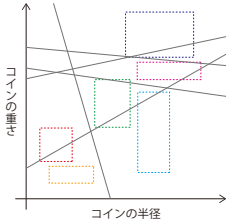
同じモデルから別の候補.

モデルの概念



今度は「長方形」ではなく,
「楕円形」のモデルからの一候補.

モデルの概念

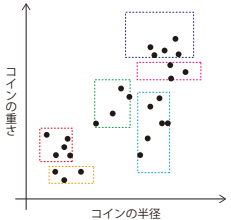


次は「楕円形」ではなく，
「平面」のモデルからの一候補．

機械学習の基本的なパーツ

- モデル
- データ
- 学習アルゴリズム
- 評価
- 汎化性能

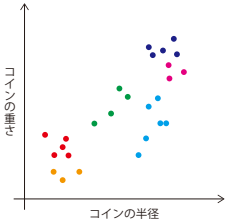
データの概念



データがないと何もできない．

少なくとも，
特徴量の観測値は前提とする．

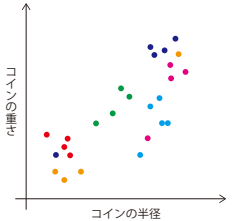
データの概念



データに「ラベル」が付く場合もある．
（教師ありともいう）

左図：規則性のノイズがない状況．

データの概念



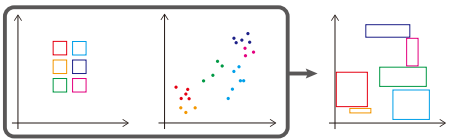
データに「ラベル」が付く場合もある．
（教師ありともいう）

左図：規則性のノイズがある状況．

機械学習の基本的なパーツ

- モデル
- データ
- 学習アルゴリズム
- 評価
- 汎化性能

学習アルゴリズムの概念

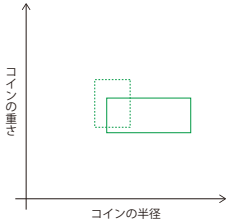


- 学習アルゴリズムはデータを受け取り，モデルから候補を選び取る．
- アルゴリズムの作業中は学習時．それが終われば実行時（もしくは評価時）．

機械学習の基本的なパーツ

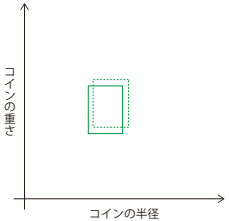
- モデル
- データ
- 学習アルゴリズム
- 評価
- 汎化性能

評価の概念



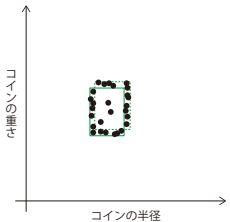
「関数近似」の評価が悪い一例.

評価の概念



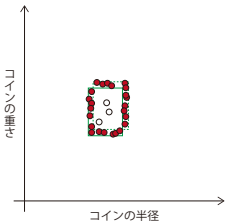
「関数近似」の評価が良い一例.

評価の概念



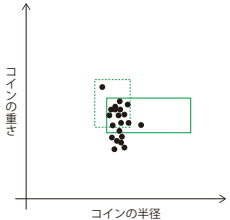
「予測能力」の評価が悪い一例.

評価の概念



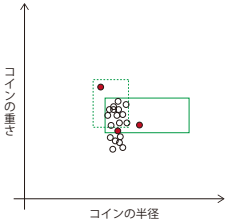
「予測能力」の評価が悪い一例.

評価の概念



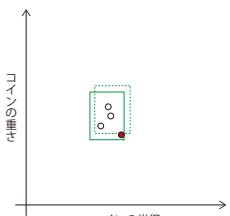
「予測能力」の評価が良い一例．

評価の概念



「予測能力」の評価が良い一例．

評価の概念



左図の例では，

- 不正解の数
- 不正解の割合
- 不正解が多いほど増える関数
- など...

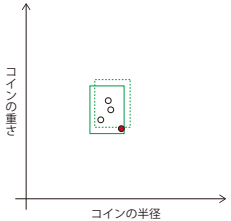
「小さい方が嬉しい」定量評価は
いろいろと考えられる．

総じて損失と呼ぶ．
（データと候補の関数）

機械学習の基本的なパーツ

- モデル
- データ
- 学習アルゴリズム
- 評価
- 汎化性能

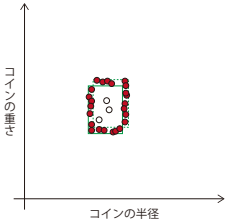
汎化性能の概念



汎化誤差とは，
学習時と実行時の評価のギャップ．

学習アルゴリズムに対して，
その出力の汎化誤差が小さいなら，
汎化性能が良いという．

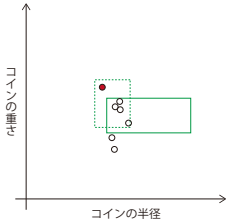
汎化性能の概念



汎化誤差とは，
学習時と実行時の評価のギャップ．

学習アルゴリズムに対して，
その出力の汎化誤差が小さいなら，
汎化性能が良いという．

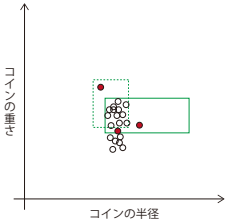
汎化性能の概念



汎化誤差とは，
学習時と実行時の評価のギャップ．

学習アルゴリズムに対して，
その出力の汎化誤差が小さいなら，
汎化性能が良いという．

汎化性能の概念



汎化誤差とは，
学習時と実行時の評価のギャップ．

学習アルゴリズムに対して，
その出力の汎化誤差が小さいなら，
汎化性能が良いという．

まとめ

要点

- 予測能力の評価は損失が担う．
- 学習時と評価時に分けて，その「出来栄え」の違いが「汎化誤差」．
- 学習アルゴリズムの役割は，モデルから「良い」候補を選ぶことである．

疑問点など

- 損失関数の設定はどのような意味を持つ？
- ほかに「学習」と言える問題設定はないのか？
- 学習アルゴリズムそのものが「良い」かどうか，どうやって定量化する？