モデルの逆解析

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

モデルの逆解析とは？

✓Y (物性・活性など) の値を回帰モデルやクラス分類モデルに入力して、
X (記述子・特徴量・パラメータ・入力変数) の値を推定すること

✓大きく分けて２つの方法がある

- 順解析を繰り返す
- ベイズの定理を利用する

順解析と逆解析

✓順解析：回帰モデルやクラス分類モデル $y = f(X)$ に、
X を入力して、y の推定値を得る

- 多変数(X) → 1 変数(y) なので問題ない

✓逆解析：回帰モデルやクラス分類モデル $y = f(X)$ に、
y を入力して、X の推定値を得る

- 1 変数(y) → 多変数(X) なので一般的には解析解が得られない

モデルの逆解析のやり方

✓順解析を繰り返し、目標の y になる X の値を選択する

- 全通りの X の候補を用いる (グリッドサーチ)
- ランダムに X の値を生成する
- 遺伝的アルゴリズム (Genetic Algorithm, GA) などの最適化手法を用いる

✓ベイズの定理を利用する

全通りの X の候補を用いる (グリッドサーチ)

- ✓X の変数それぞれに候補を設定し、それらのすべての組み合わせを X のデータとする
- ✓X のデータのうち、モデルの適用範囲 (Applicability Domain, AD) の中のサンプルのみ回帰モデルやクラス分類モデルに入力して、y の値を推定する
  - https://datachemeng.com/applicabilitydomain/
- ✓y の推定値の中で、目標の y を満たす X 変数の値の組み合わせのみ選択する

- ✓変数の数や、候補の数が多くなると、すべての組み合わせの数 (グリッドサーチする数) が膨大になってしまう
  - 20 変数で、それぞれ 10 候補とすると、10^{20} 通り
- ✓設定した候補の中からしか探索されないので注意

ランダムに X の値を生成する

- ✓X の変数それぞれに上限 (最大値) と下限 (最小値) を設定し、それらの間の中で、一様乱数で X のデータを生成する
- ✓生成された X のデータのうち、AD中のサンプルのみを回帰モデルやクラス分類モデルに入力し、y の値を推定する
- ✓y の推定値の中で、目標の y を満たす X 変数の値の組み合わせのみ選択する
- ✓ランダムに生成する X のデータ (サンプル) の数を、できるだけ多くしたほうがよい

最適化手法を用いる

✓X のデータを回帰モデルやクラス分類モデルに入力して推定された y の値を目的関数とする

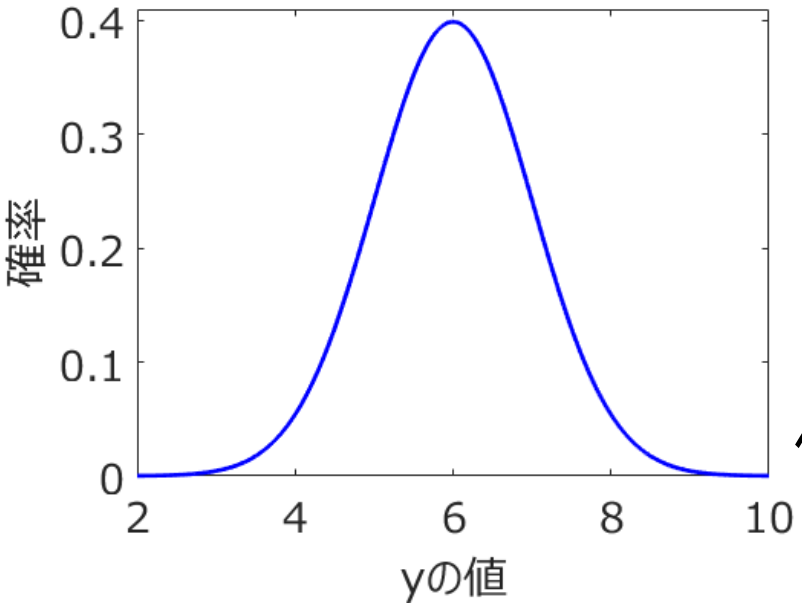
✓目的関数が最大 (もしくは最小) となるように、GA などの最適化手法により X の変数の値の組み合わせを最適化する

- AD 内のサンプルのみ考慮
- y の最大化、最小化ではなく、ある範囲に入れたい場合でも対応可能

✓解に初期値依存性があるため、最適化計算を何回かするとよい

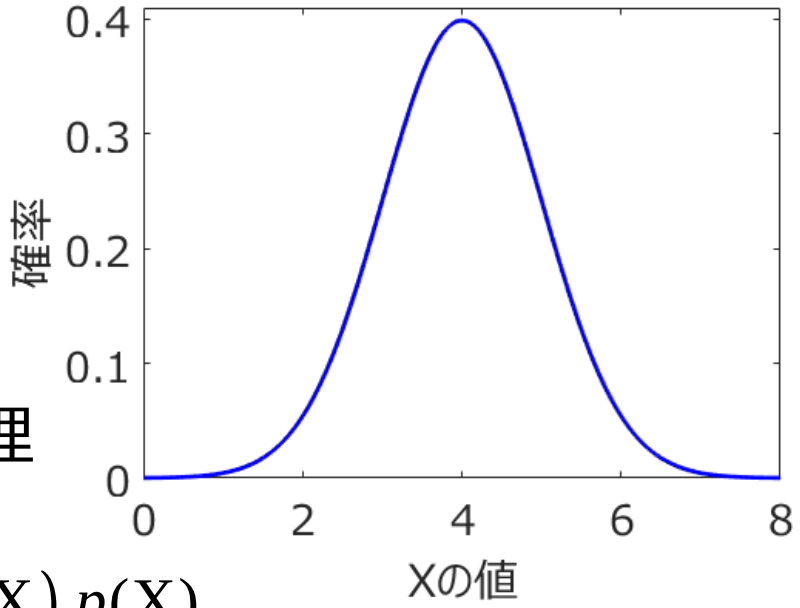
ベイズの定理を利用する

✓X, y, 回帰モデル・クラス分類モデルの出力が確率分布で与えられるときに有効



回帰分析

ベイズの定理



$$p(\mathrm{X}\,|\,\mathrm{y}) = \frac{p(\mathrm{y}\,|\,\mathrm{X})\,p(\mathrm{X})}{p(\mathrm{y})}$$

目標の y の値が得られる確率の高い、X の値の範囲が得られる

[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, Heidelberg 2006.

ベイズの定理を利用した逆解析

✓回帰モデル・クラス分類モデル：X が与えられたときの、y の確率分布 (事後分布) p(y|X)

✓求めたいもの：y が与えられたときの、X の確率分布 (事後分布) p(X|y)

ベイズの定理　$p(\mathrm{X}\,|\,\mathrm{y}) = \dfrac{p(\mathrm{y}\,|\,\mathrm{X})\,p(\mathrm{X})}{p(\mathrm{y})}$

p(y)：y の事前確率・・・正規分布と仮定
p(X)：X の事前確率 ・・・Gaussian mixture models [1] などで計算

AD は自動的に考慮される

[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, Heidelberg 2006.