ドッキングシミュレーションおよび
機械学習を活用した医薬品化学構造の設計

April. 28, 2016
ディスカッション

酒井研究室　　宮崎 大輝

回帰分析

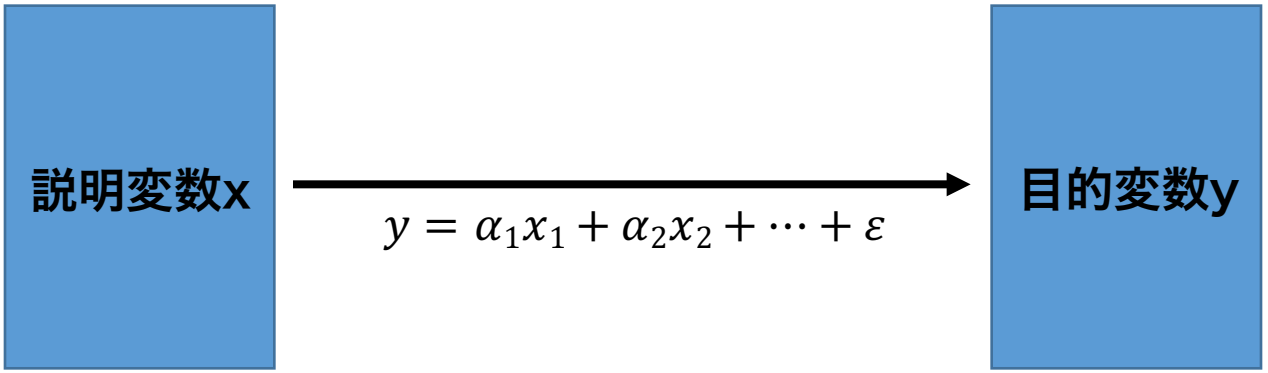
目的変数 y：測定が困難な量、物性値、特性など

説明変数 x：測定が容易な量、物性値、決まっているパラメタ、インデックスなど

目的変数 yを 説明変数 xの関数として予測することを目的とする。

線形回帰分析

目的変数を説明変数の線形関数として近似



線形回帰分析

- 線形重回帰分析(MLR)

モデル式：$y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \varepsilon \Rightarrow \boldsymbol{y = \alpha X + \varepsilon}$

モデル式の計算基準：

$\hat{\boldsymbol{y}} = \boldsymbol{\alpha X}$　　とするとき

$\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{y}} = \boldsymbol{y - \alpha X}$

の二乗和を最小とする

<u>特徴</u>

- 誤差の小さなモデル式が比較的簡単に得られる。
- 説明変数xが直接モデル式に組み込まれている。

<u>問題点</u>

- 多重共線性 (説明変数間の相関により、モデル化の有意性に問題)

⟶　主成分回帰分析

線形回帰分析

- 主成分回帰分析(PCR)

モデル式：$y = \alpha T + \varepsilon$　T：主成分 (説明変数の集約として抽出)

モデル式の計算基準：

$\hat{y} = \alpha T$　　とするとき

$\mathbf{e} = \mathbf{y} - \hat{y} = y - \alpha T$

の二乗和を最小とする

<u>特徴</u>

・主成分を入力変数として、線形回帰式を構築する

・主成分によって入力変数間の相関関係を捉えることが可能

<u>問題点</u>

・<u>主要な主成分が出力変数の推定に寄与するとは限らない</u>
(出力変数との相関が強い変数を入力変数として採用すべき)

⟶　Partial Least Squares 回帰法(PLS)

線形回帰分析
- PLS (Partial Least Squares) 回帰法

目的変数と潜在変数（説明変数の線形結合）との内積が最大になるように、
潜在変数を決定

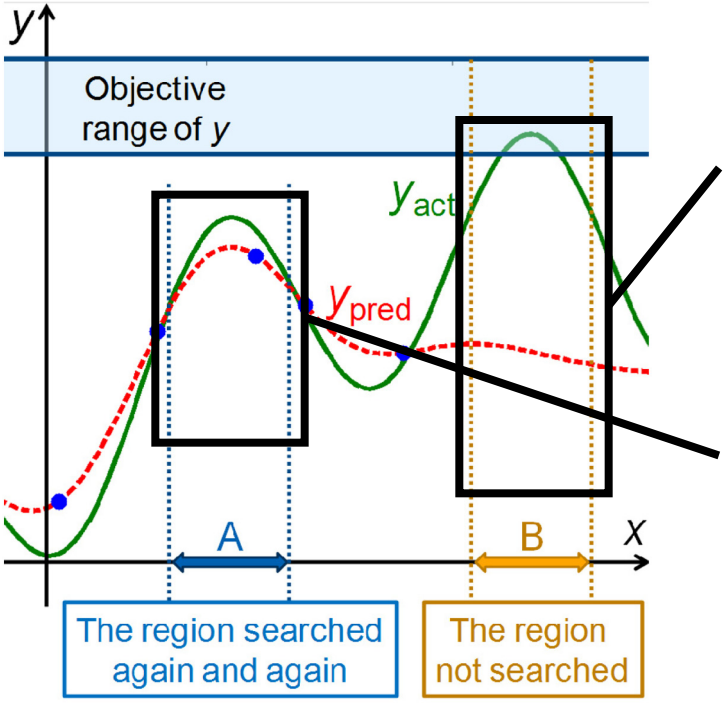
<u>特徴</u>

・計量化学において、サンプルサイズに比べて圧倒的に変量が多い場合や
　変数間の共線性が高い場合に有用

・回帰分析の精度向上だけでなく、次元削減、関連因子の抽出などの用法も

回帰モデルによる探索の問題点

・データ量の大小による予測誤差のために、適切な外挿領域の探索が行われない



データ量：小
予測誤差：大

探索すべき領域であるが、
予測値が低く、探索は行われない

データ量：大
予測誤差：小

目標値に達していないが、
探索され続けてしまう

・モデル構築に用いた既知データの密度が低いと、予測値の信頼性が低い

⟶ 予測誤差の大きさとデータ密度の考慮により効率的な探索が可能

予測誤差の大きさの推定
- GP(Gaussian Process)法

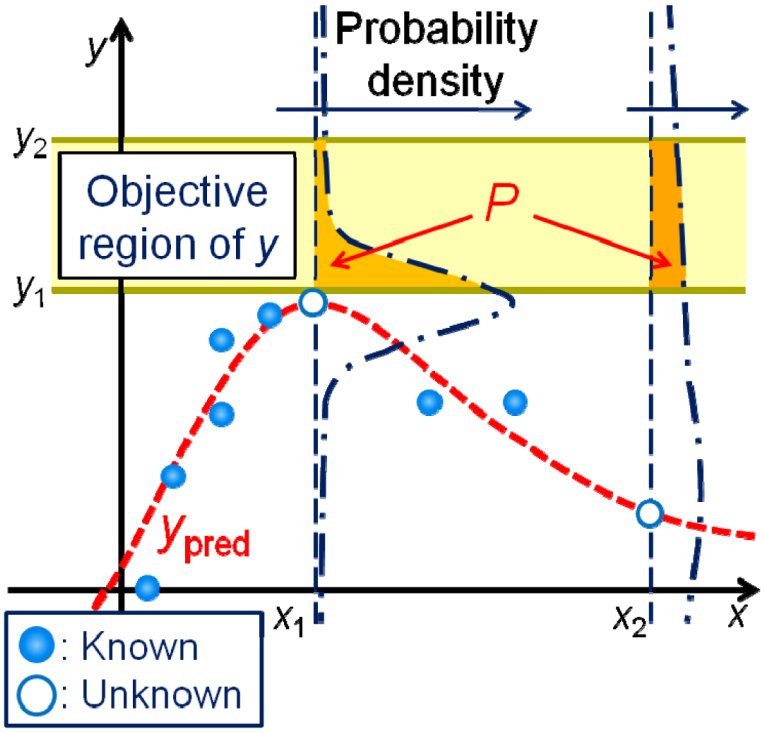
ある説明変数xが与えられた時に、
目的変数yを正規分布に従う確率モデルとする回帰手法

モデル式：$y = w^T \varphi(x)$
(φ：非線形関数、w: 回帰パラメタ)

- 予測誤差の分散s^2を求めることが可能
- s^2を用いて目的物性達成確率Pを算出

$$P = \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi}s} exp\left\{\frac{(y - y_{pred})^2}{2s^2}\right\} dy$$

予測誤差の分散が大きいx_2のような候補においてPが大きい値をとる



⟶ 探索の効率化