基本的なデータの前処理

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

どうしてデータの前処理をするの？

✓単位系が異なる場合など、各変数(記述子)が同等に扱われない
- 長さ: km, m, cm, mm, nm など
- 温度: ℃, K など

✓データ分布の中心が 0 であると、何かとうれしい

⇨ オートスケーリング (標準化)

✓情報量のない変数はいらない (かえって邪魔になるときもある)
- ほぼすべてのサンプルで値が同じ変数

⇨ 分散が0の変数の削除、同じ値を多くもつ変数の削除

- 似た変数の組の１つ

⇨ 相関係数の高い変数の組の１つを削除

オートスケーリング (標準化)

✓データ解析・ケモメトリックスにおける一般的な前処理の方法

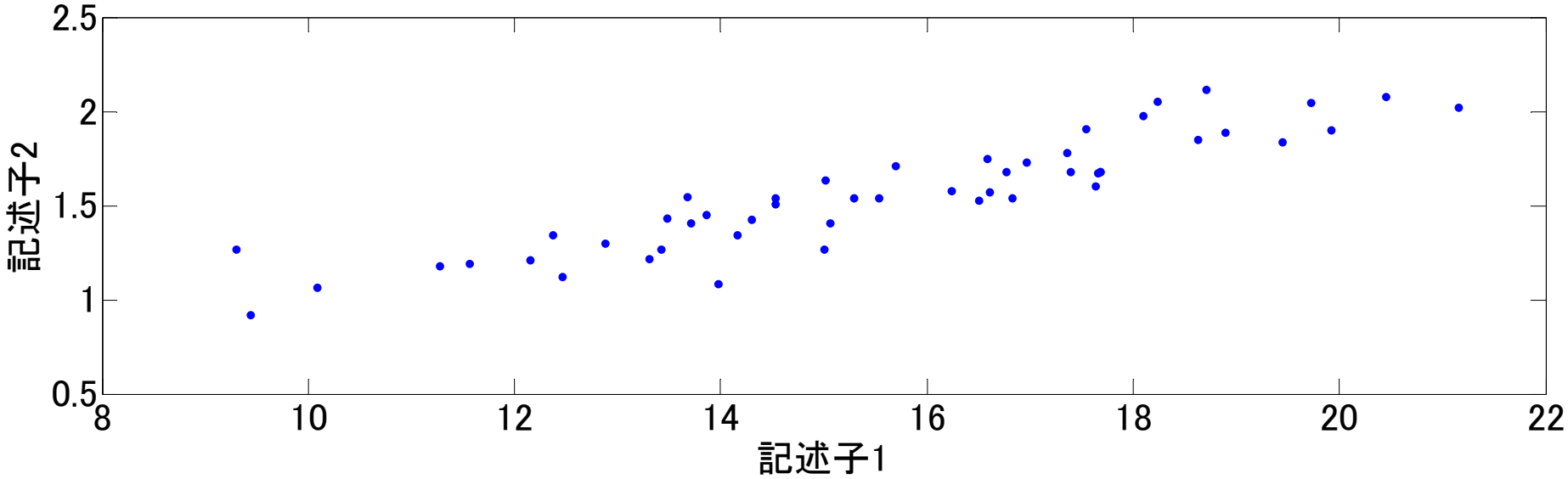
✓オートスケーリング ＝ センタリング　＋　スケーリング

- センタリング: 変数(記述子)ごとにその平均を引き、
  平均を 0 にする

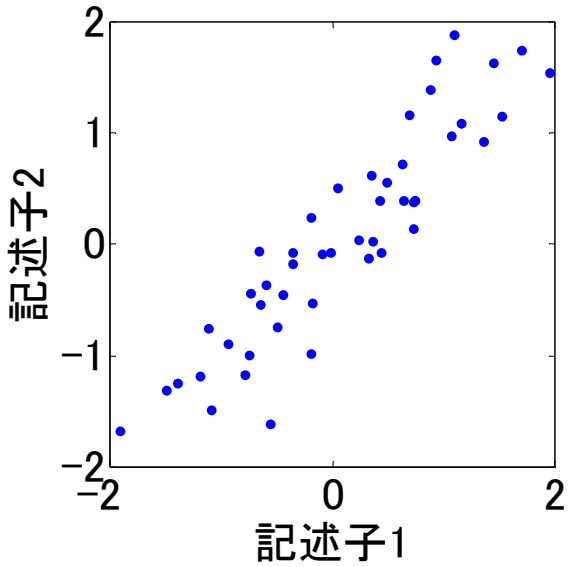
- スケーリング: 変数(記述子)ごとにその標準偏差で割り、
  標準偏差を 1 にする

✓各変数(記述子)が同等の重みを持つようになる

オートスケーリングの例



オートスケーリング



センタリング

$x_i^{(k)}$: k 個目のサンプルにおける、i 番目の変数(記述子) の値

✓センタリング　各変数(記述子)の平均を0にする
（それぞれのサンプルから平均を引く）

$$x_i^{(k)\prime} = x_i^{(k)} - \mu_i$$

$$\mu_i = \frac{\sum_{k=1}^{n} x_i^{(k)}}{n}$$

n : サンプル数

スケーリング

$x_i^{(k)}$: k 個目のサンプルにおける、i 番目の変数(記述子) の値

✓スケーリング　各変数(記述子)の標準偏差を1にする
（それぞれのサンプルを標準偏差で割る）

$$x_i^{(k)\prime\prime} = \frac{x_i^{(k)\prime}}{\sigma_i}$$

$$\sigma_i = \sqrt{\frac{\sum_{k=1}^{n}\left(x_i^{(k)} - \mu_i\right)^2}{n-1}}$$

モデル検証用(テスト)データのオートスケーリング

✓モデル検証用データ(テストデータ)のオートスケーリングには、
モデル構築用データ(トレーニングデータ)の平均・標準偏差を使用

- テストデータの平均・標準偏差ではないので注意
- テストデータの平均・標準偏差を使うとトレーニングデータのスケールと変わってしまう

$$x_{\mathrm{test},i}^{(k)}{}'' = \frac{x_{\mathrm{test},i}^{(k)} - \mu_i}{\sigma_i}$$

$x_{\mathrm{test},i}^{(k)}$: テストデータのk 個目のサンプルにおける、i 番目の変数(記述子) の値

μ_i : トレーニングデータの i 番目の変数(記述子) の平均

σ_i : トレーニングデータの i 番目の変数(記述子) の標準偏差

分散が0の変数の削除

✓分散が0、つまりすべてのサンプルで同じ値をもつ変数は、意味がない

✓分散が0ということは、標準偏差が0なので、スケーリングができない
(0で割ることになってしまう)

✓最初に、分散 $\frac{\sum_{k=1}^{n}\left(x_i^{(k)}-\mu_i\right)^2}{n-1}$ が 0 の変数を削除しましょう！

$x_i^{(k)}$: k 個目のサンプルにおける、i 番目の変数(記述子) の値

n : サンプル数

$$\mu_i = \frac{\sum_{k=1}^{n} x_i^{(k)}}{n}$$

同じ値を多くもつ変数の削除

✓分散が 0 の変数を削除するだけで十分か？

✓１つのサンプルの値が１で、他のサンプルの値がすべて 0 のような変数もいらなそう

- (注意！) 分散の値が小さい、ということではない。
  分散の小さい、たとえば 0.01未満の、変数を削除してしまうと、すべて小さい値でばらつきは小さいが重要な変数を削除する危険性がある
- クロスバリデーション(交差検定)をするときに、サンプルを分割したあとに分散が０になってしまうとよくない
  (クロスバリデーションを知らない人は意味がわからなくてOKです)

✓同じ値を多くもつ変数も削除しましょう！

- わたし(金子)は、よく 5-fold クロスバリデーションを行うため、8割以上が同じ値である変数を削除しています

注意点

✓ １つのサンプルの値が１で、他のサンプルの値がすべて 0 のような変数
- ノイズで １ になった変数のときは、過学習してしまうため変数を削除すべき

- その変数で １ をとるサンプルが y に対して意味をもつときもある
  - ベンゼン環をもつ分子が一つだけあり、
  - yが毒性の有無で、ベンゼン環によって毒性が発生するとき

✓削除しないときと、削除するときの両方モデリングして比較するとよい
- クロスバリデーションでは注意が必要

相関係数の高い変数の組の１つの削除

✓同じ変数が２つあっても意味がない

✓ちょっとしか違わないが (誤差？というレベルで) 似ている変数も、
どちらか１つでOK

✓最初に変数の数を減らしておくことで、

- 次元の呪いを低減できる
- あとのデータ解析がやりやすくなる

✓相関係数が高い変数の組の１つを削除しましょう！

$$\frac{\sum_{k=1}^{n}\left(x_i^{(k)}-\mu_i\right)\left(x_j^{(k)}-\mu_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_i^{(k)}-\mu_i\right)^2\sum_{k=1}^{n}\left(x_j^{(k)}-\mu_j\right)^2}}$$

: i 番目の変数と j 番目の変数との相関係数

しきい値は？どちらを消す？

✓しきい値は？

- 0.8, 0.9, 0.95, 0.99など、いろいろな候補があります
- たとえば、0.99 のように思い切って決めてしまうか、
  細かく最適化したい場合は試行錯誤的に決めることになります

✓２つのうち どちらを消す？

- どちらでもあまり変わりませんが、その他の変数との相関係数を
  調べて、その絶対値の和の大きい方が他の変数との重複が
  大きいと考え、そちらを削除するようにしています

注意

- ✓主成分分析 (Principal Component Analysis, PCA) や部分的最小二乗法 (Partial Least Squares, PLS) をすれば、基本的に変数間の相関関係には対処できる

- ✓相関係数の高い変数の組の１つを削除したからといって、その後の解析結果があまり変わらないこともある

[発展] 変数間の非線形性を考えた相関係数

✓相関係数は、変数間に直線的な関係があるときに、
値が１や -1 付近になる

✓変数間に、指数関数的・対数関数的な関係など
非線形関係があるときには、相関係数の絶対値が小さくなってしまう

⇨ Maximum Information Coefficient (MIC)
[Reshef, D. N., et al., Science, 334, 1518–1524, 2011.]

✓MIC

- 変数間の非線形性を考慮した相関係数
- “A Correlation for the 21st Century” とのこと
  http://science.sciencemag.org/content/334/6062/1502.full

- MICの大きい変数の組の１つを削除するのもよいでしょう
- R言語でパッケージあり (minerva)