モデルの適用範囲
モデルの適用領域
Applicability Domain (AD)

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

モデルの適用範囲（AD）とは？

✓モデルが十分な性能を発揮できるデータ領域

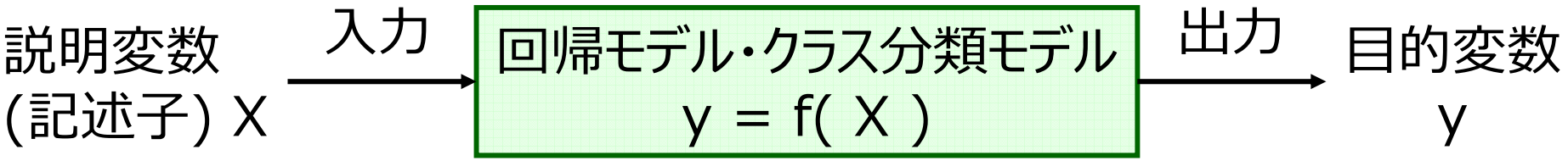
✓AD の設定方法

- トレーニングデータの範囲
- トレーニングデータの中心からの距離
- データ密度
- アンサンブル学習ベイズの定理を利用

✓モデルとの距離 (Distance to Model, DM) によって AD の広さを定量的にすることで、推定誤差を見積もることができる

✓クラス分類でアンサンブル学習により AD を設定するときは注意する

どんな X の値でもモデルに入力してよいのか？



入力してはいけない説明変数 X の値がある！

モデルの適用範囲・適用領域
(Applicability Domain, AD) [1-3]

[1] I. V. Tetko, et al., J. Chem. Inf. Model. 2008, 48, 1733.
[2] D. Horvath, G. Marcou, A. Varnek, J. Chem. Inf. Model. 2009, 49, 1762.
[3] H. Kaneko, M. Arakawa, K. Funatsu, AIChE J. 2011, 57, 1506.

モデルの適用範囲・適用領域のイメージ

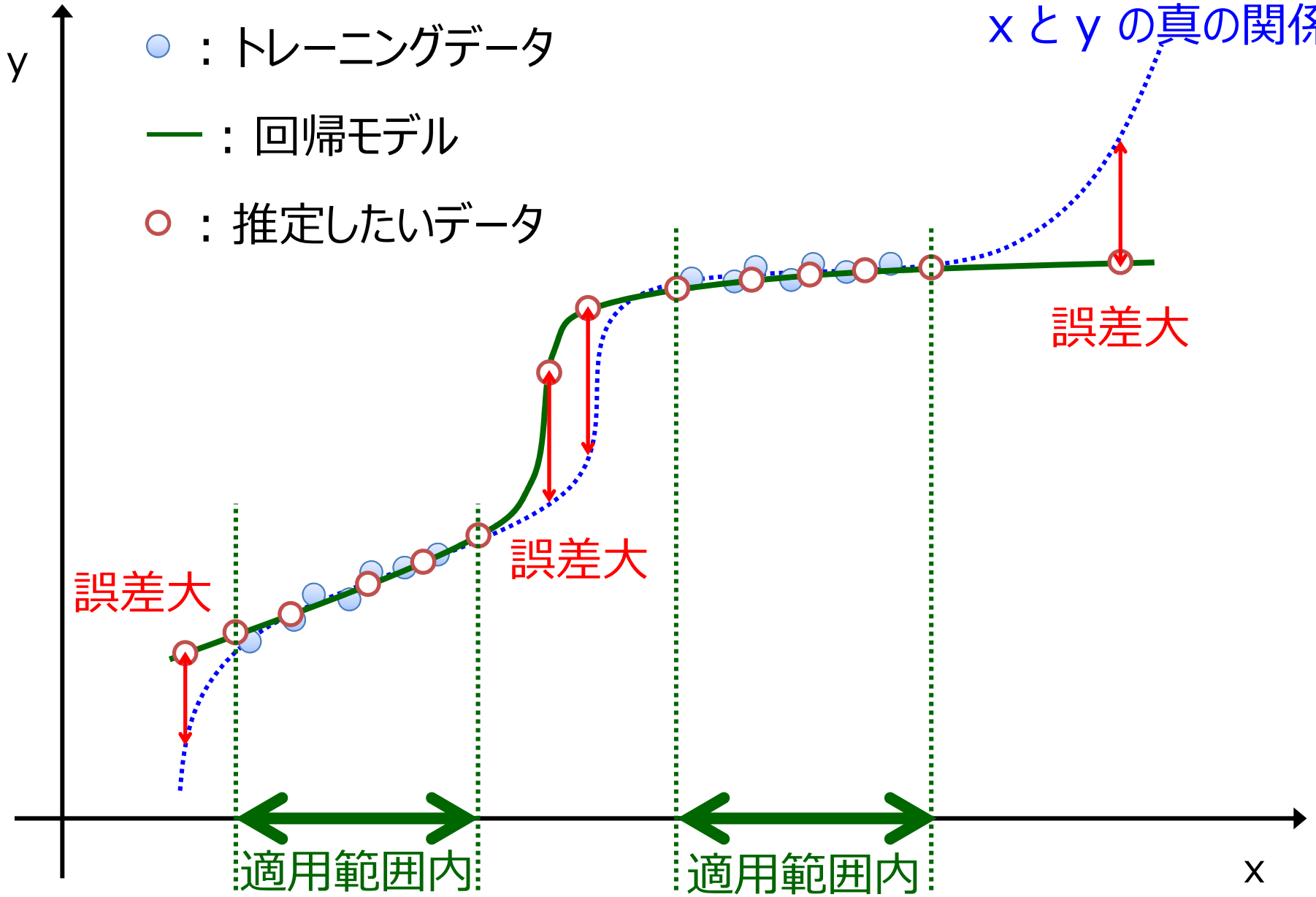
✓夏は暑い

- 夏であっても涼しい日はある
- クーラーの効いた部屋にいれば暑くない
- 北極や南極などでは夏でも寒い

✓炭化水素のデータで構築した水溶解度を推定するモデル

- アルコールの水溶解度を正しく推定できるか？

モデルの適用範囲・適用領域のイメージ



モデルの適用範囲・適用領域

予測したいデータによって、モデルの信頼性は異なる

『モデルが十分な性能を発揮できるデータ領域』を定めよう！

⇨ モデルの適用範囲・適用領域 (Applicability Domain, AD)

- QSAR [1-3]
- QSPR [4-6]

適切にモデルの適用範囲を設定し、
推定するときは適用範囲内かどうか判断する必要がある

[1] R.P. Sheridan, et al., *J. Chem. Inf. Comput. Sci.,* **44**, 1912-1928 (2004)
[2] I.V. Tetko, et al., *Drug Discov. Today,* **11**, 700-707 (2006)
[3] D. Horvath, et al., *J. Chem. Inf. Model*, **49**, 1762-1776 (2009)
[4] P. Bruneau, N.R. McElroy, *J. Chem. Inf. Model*, **46**, 1379-1387(2006)
[5] A. Schwaighofer, et al., *J. Chem. Inf. Model*, **47**, 407-424 (2007)
[6] Alexandre V., Igor B., *J. Chem. Inf. Model*, **52**, 1413-1437 (2012)

AD の設定

✓トレーニングデータの範囲 [1]

✓トレーニングデータの中心からの距離 [2,3]

✓データ密度 [4,5]

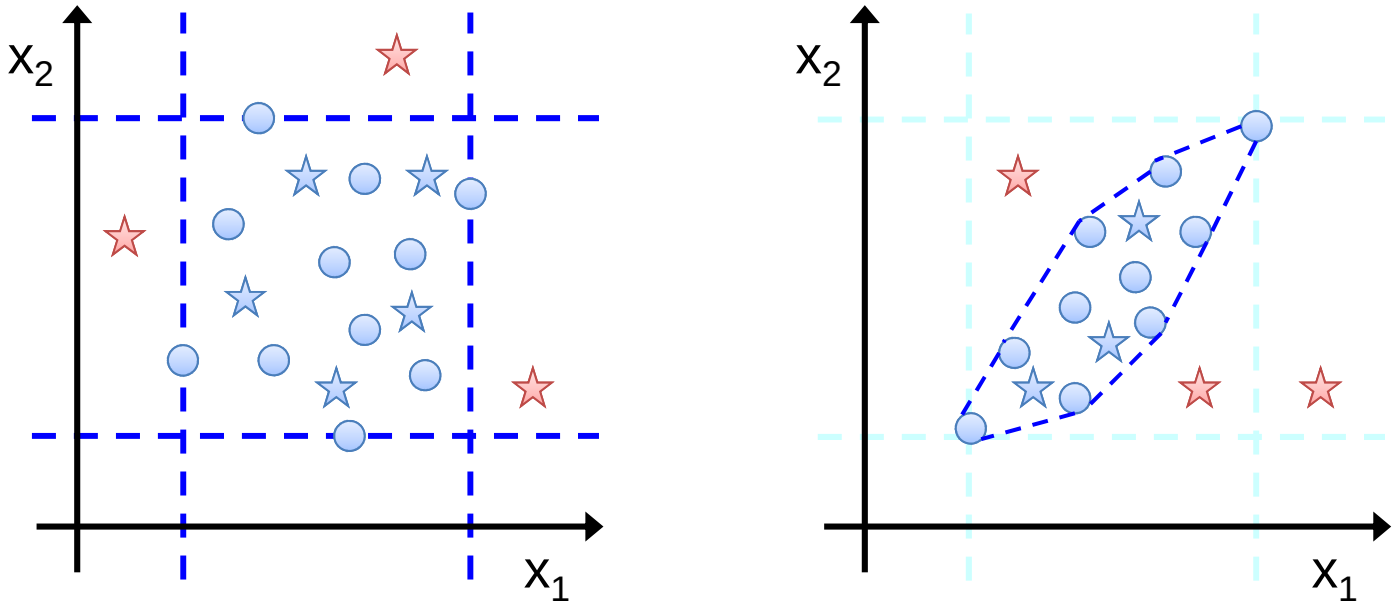
✓アンサンブル学習 [6]

[1] H. Kaneko, et al., Comput. Chem. Eng. 35 (2011) 1135–1142.
[2] S. Dimitrov, et al., J. Chem. Inf. Model. 45 (2005) 839–849.
[3] I. Sushko, et al., J. Chem. Inf. Model. 50 (2010) 2094–2111.
[4] I.I. Baskin, et al., Mol. Inf. 29 (2010) 581–587.
[5] H. Kaneko, et al., Chemometr. Intell. Lab. Syst. 58 (2001) 109–130.
[6] H. Kaneko, et al., J. Chem. Inf. Model. 54 (2014) 2469-2482.

トレーニングデータの範囲

✓モデル構築用データにおける各説明変数 X の範囲

✓変数間の相関が大きい場合は、凸包や主成分分析後のスコアを利用



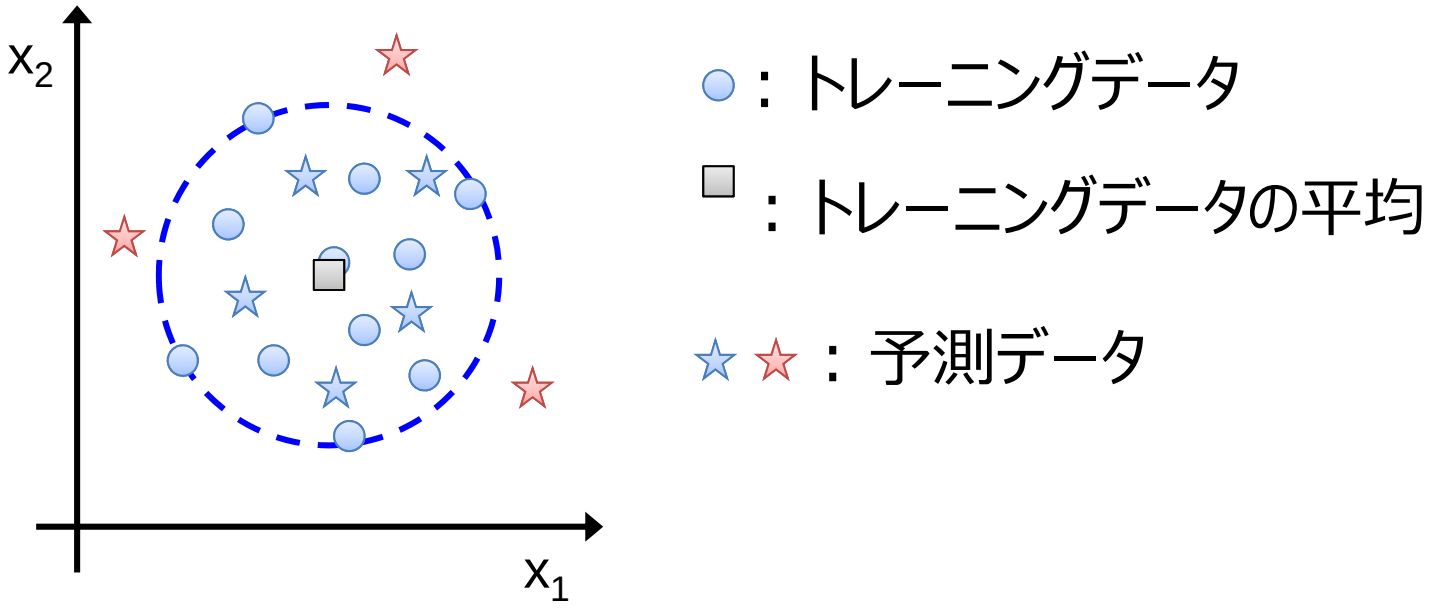
：トレーニングデータ

：予測データ

トレーニングデータの中心からの距離

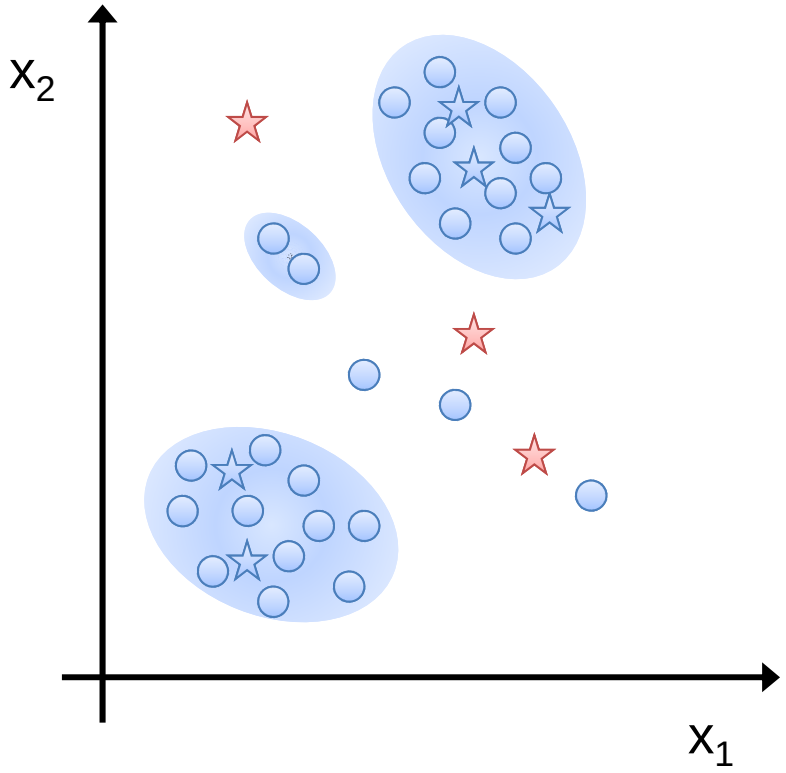
✓モデル構築用データの平均までの距離

✓変数間の相関が大きい場合は、マハラノビス距離を用いる



データ密度

✓モデル構築用データが密に存在する領域が適用範囲内



○：トレーニングデータ

☆☆：予測データ

データ密度

✓k最近傍法 (k-Nearest Neighbor method, k-NN)

- 詳細は https://datachemeng.com/knn/

✓One-Class Support Vector Machine (OCSVM)

- 詳細は https://datachemeng.com/ocsvm/

データ密度が高い ⇨ AD 内

アンサンブル学習

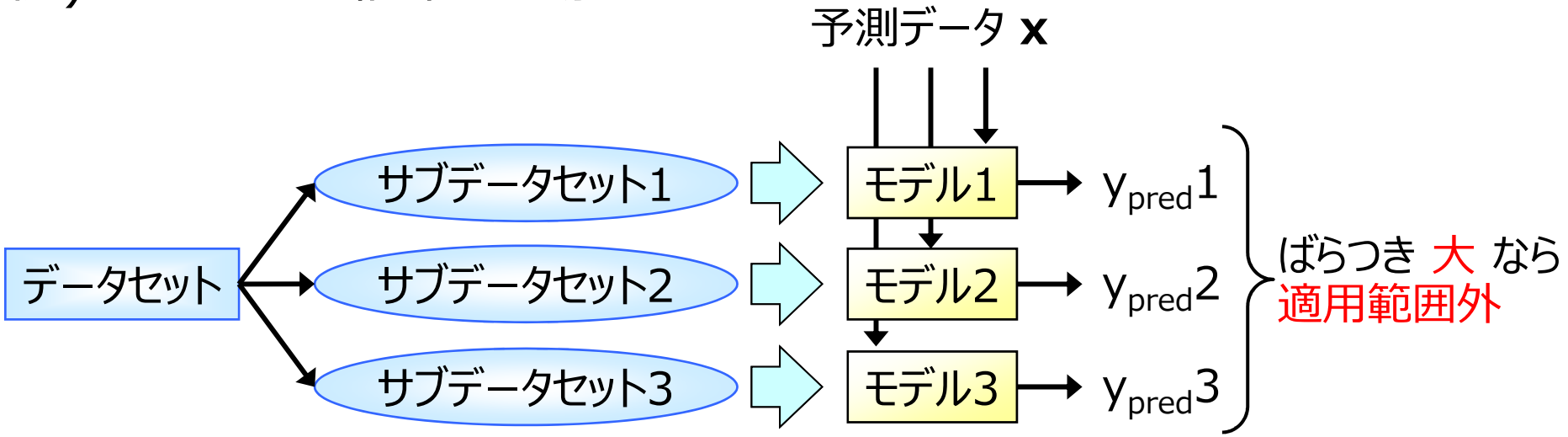
✓複数個のモデルを構築し、それらによる予測値の分散(ばらつき)を用いて適用範囲を評価

- 詳細は https://datachemeng.com/ensemblelearning/

✓例えば、

- トレーニングデータのサンプル
- 記述子
  をランダムに選択して複数モデルを構築

例) モデルを3つ構築した場合



モデルとの距離 (Distance to Model)

モデルの適用範囲(AD)：モデルが十分な性能を発揮できるデータ領域

ある狭いデータ範囲　・・・　モデルは高い性能を発揮

広いデータ範囲　・・・　モデルはある程度の性能を発揮

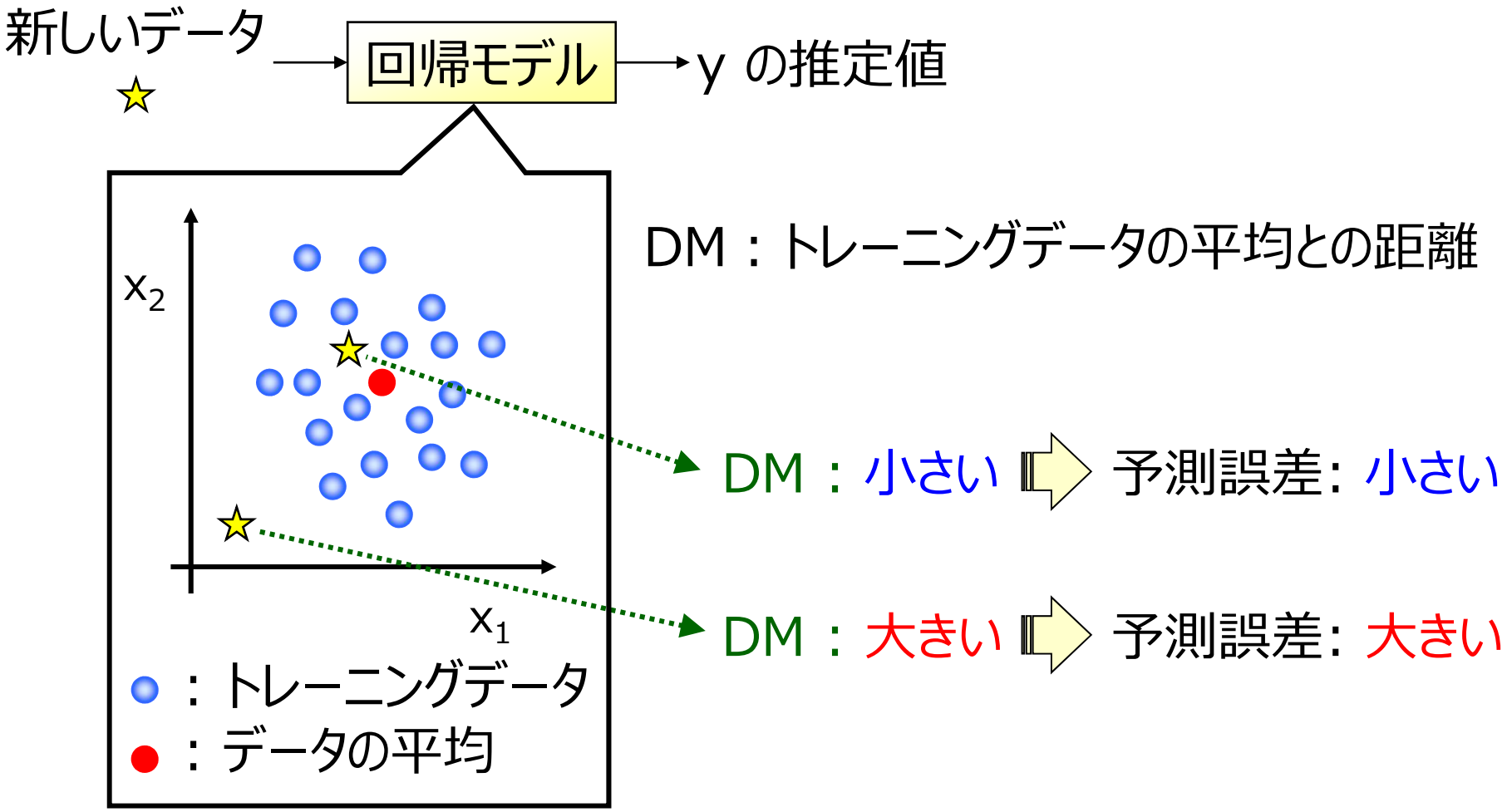
⇨ モデルとの距離 (Distance to Model, DM) [1,2] の導入

モデルとの距離が大きくなるにつれてモデルの性能(信頼性)は小さくなる

[1] Baskin II, Kireeva N, Varnek A.. *Mol. Inf.*, **29**, 581-587 (2010)
[2] Iurii S., et al., *J. Chem. Inf. Model*, **50**, 2094-2111 (2010)

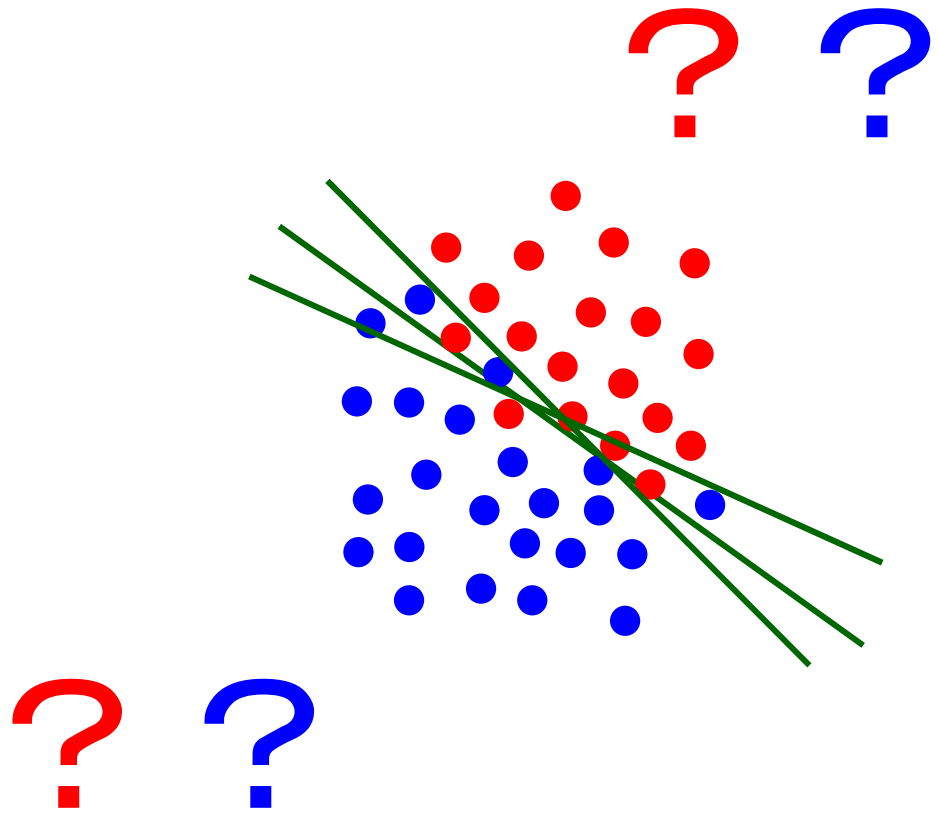
回帰モデルの予測誤差(信頼性)の推定

回帰モデルとの距離 (DM) と予測誤差との関係を定量化 [1]



注意！

✓クラス分類のとき、アンサンブル学習だけでモデルの適用範囲を設定すると、広くなりすぎてしまいます！



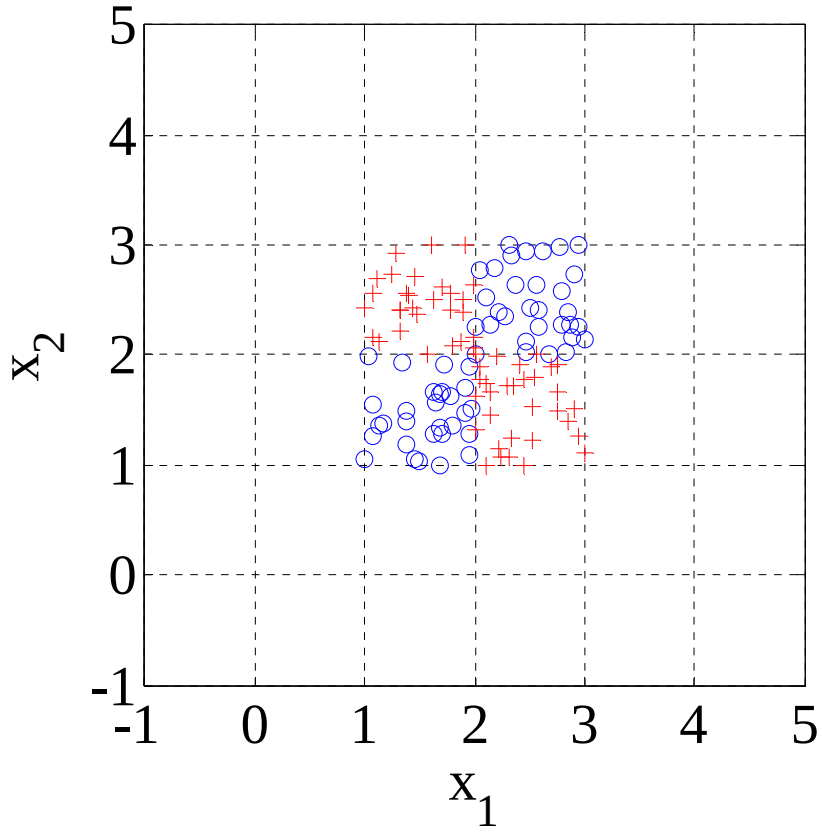
[1] H. Kaneko, K. Funatsu, J. Chem. Inf. Model., 54, 2469-2482, 2014

数値シミュレーションデータで確認

✓2クラス分類
- 各クラス 60データ

✓記述子: 2つ

✓クラス分類手法
- k-NN (k最近傍法)
- RF (Random Forest)
- SVM (Support Vector Machine)



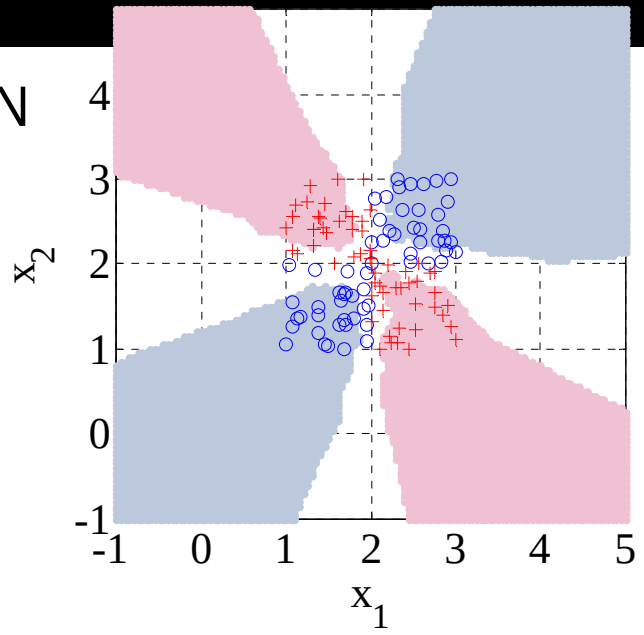
✓アンサンブル学習
- サンプルバギング (sample bagging, SB)
  - 重複を許してモデル構築用データを選択

✓モデルの数: 101

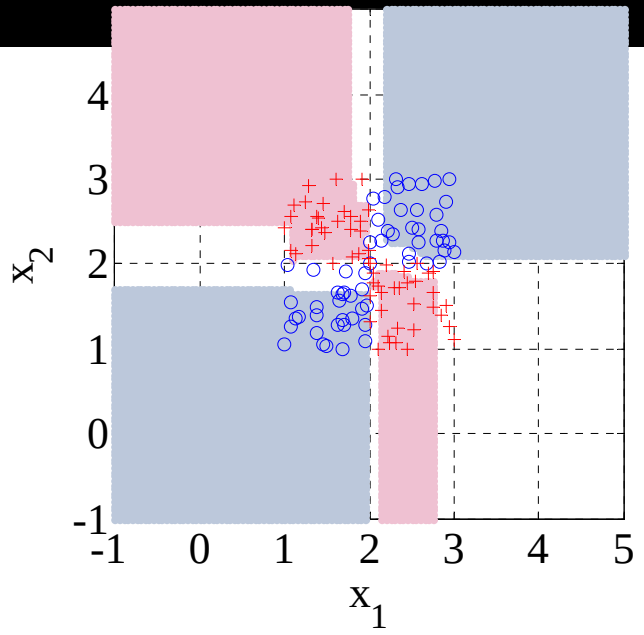
[1] H. Kaneko, K. Funatsu, J. Chem. Inf. Model., 54, 2469-2482, 2014

すべてのサブモデルで分類結果が一致した領域

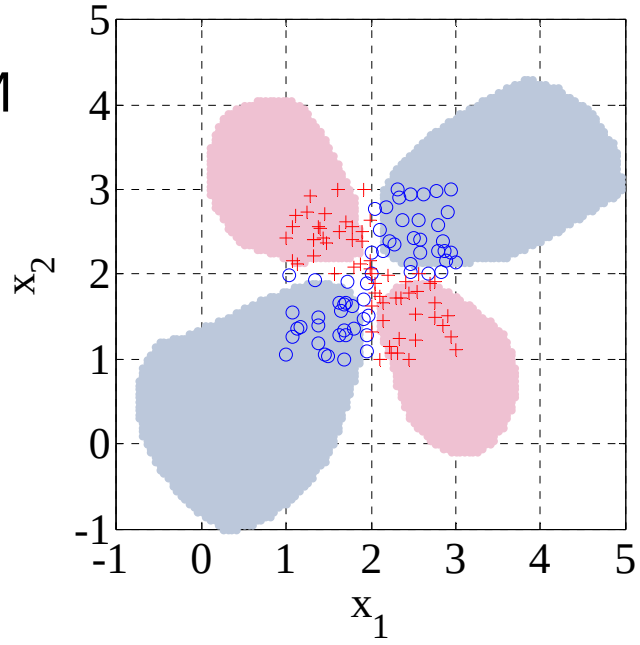
k-NN



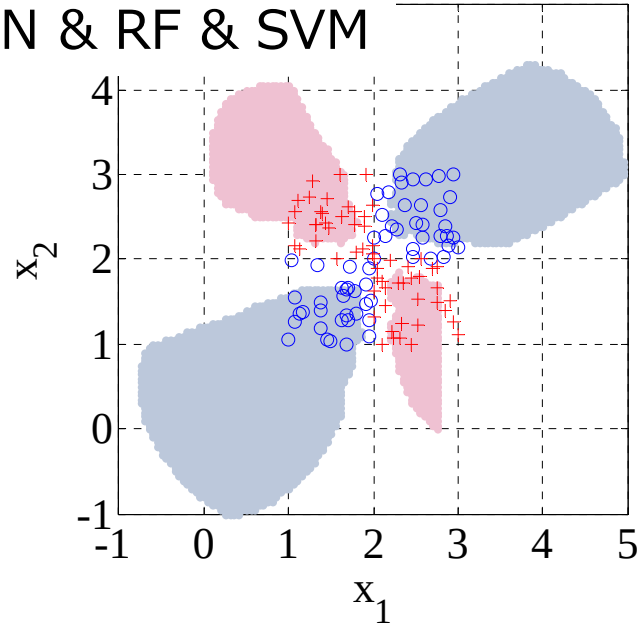
RF



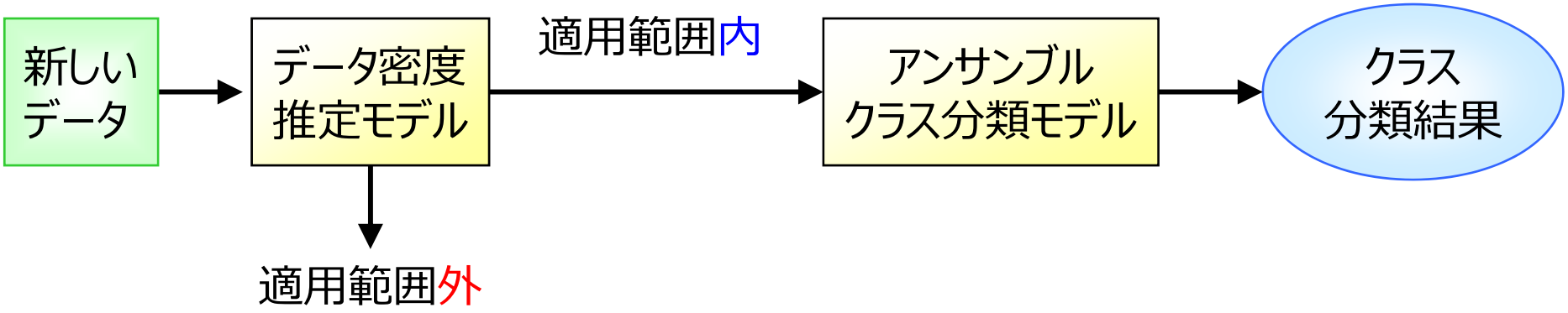
SVM

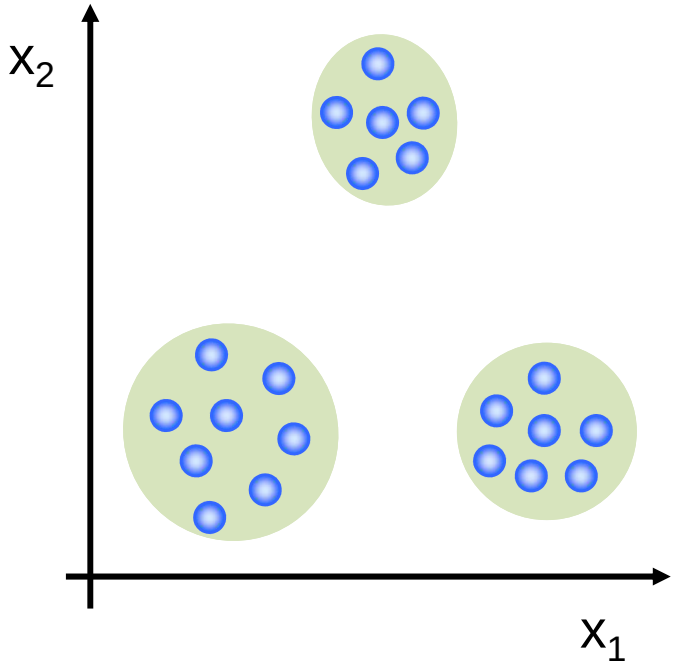


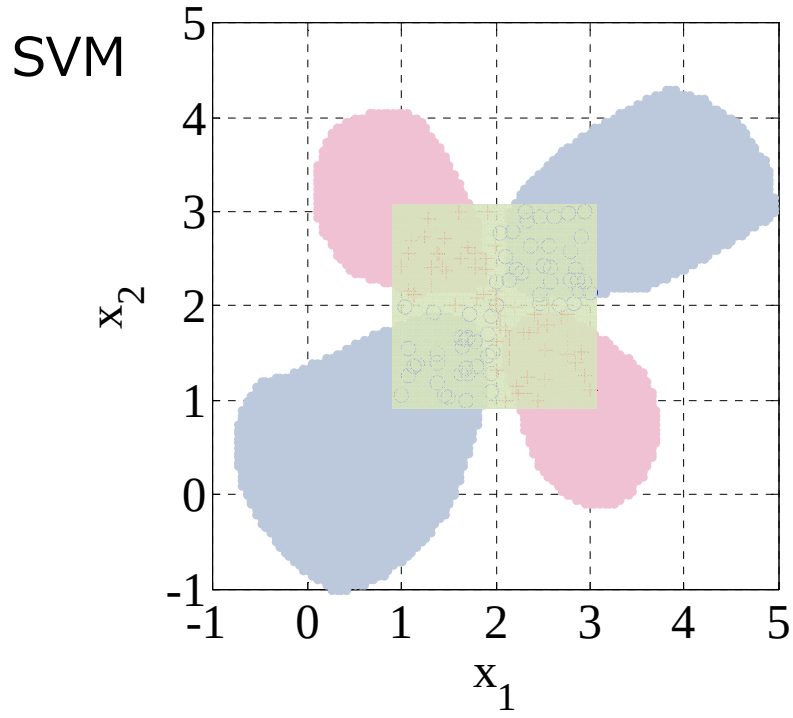
k-NN & RF & SVM



データ密度も使いましょう！







[1] H. Kaneko, K. Funatsu, J. Chem. Inf. Model., 54, 2469-2482, 2014