k最近傍法
k-Nearest Neighbor
k-NN

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

k-NN とは？

✓各サンプルについて、最も距離の近い k 個のサンプルに基づく方法

✓クラス分類、回帰分析、モデルの適用範囲に利用可能

✓k の値を事前に決める必要がある

✓距離として、ユークリッド距離だけでなく様々な距離を利用できる

✓距離を非類似度と考えて、いろいろな類似度の指標も利用できる

k-NN によるクラス分類

✓クラスを推定したいサンプル $\mathbf{x}_{\text{new}}$ について、すべてのモデル構築用サンプルとの間でユークリッド距離を計算する

✓最も距離の近い k 個のサンプルを選択する

✓k 個のクラスで多数決をとった結果を、$\mathbf{x}_{\text{new}}$ の推定されたクラスとする

✓k 個のクラスにおける、推定されたクラスの割合で信頼度を検討できる

- たとえば $k = 7$ のとき、

① 4 サンプルがクラス A、3 サンプルがクラスBであった $\mathbf{x}_{\text{new}}$
② 7 サンプルすべてがクラスAであった $\mathbf{x}_{\text{new}}$

があれば、② の方が推定結果を信頼できる

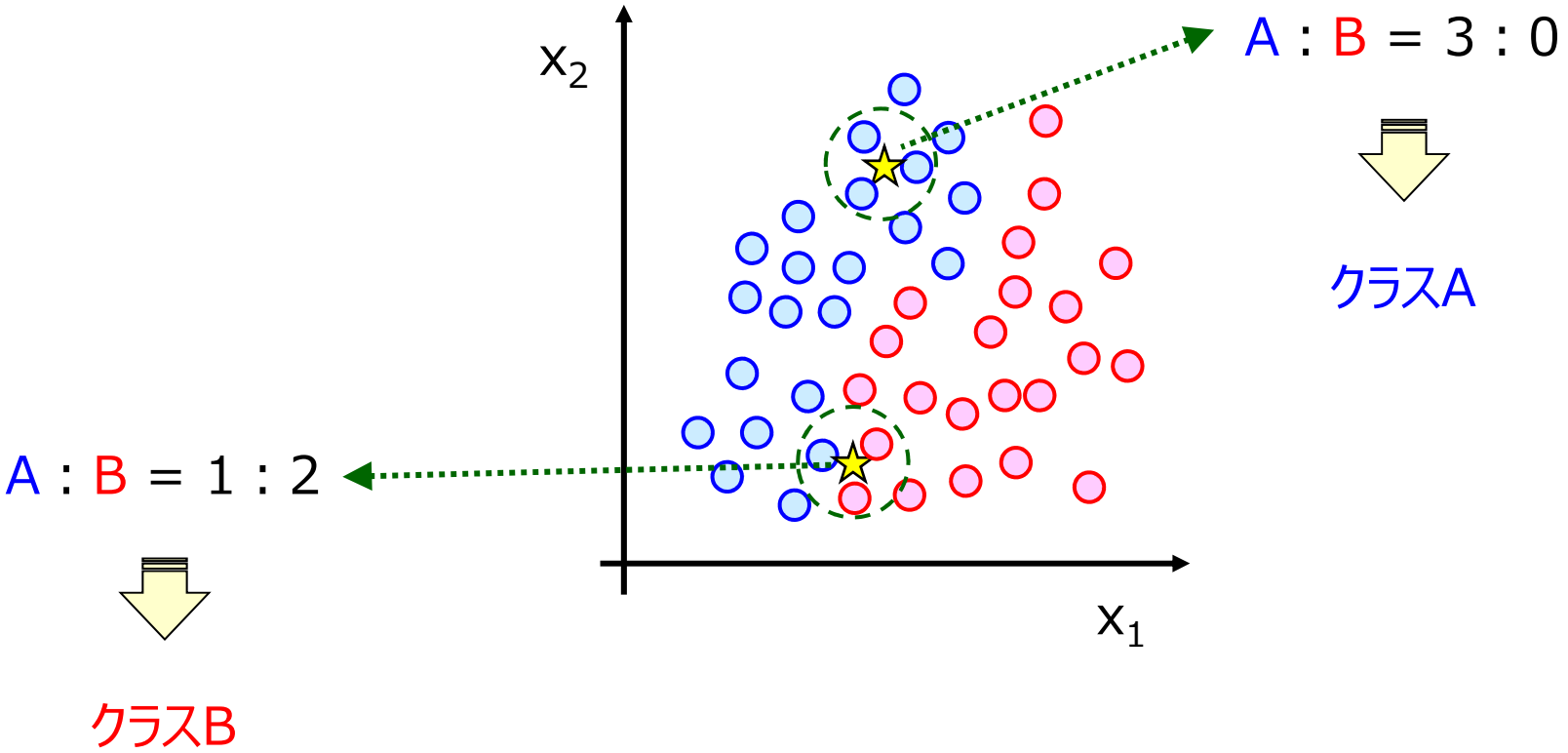
k-NN によるクラス分類　図

例) $k = 3$

○：クラスAのサンプル

○：クラスBのサンプル

☆：クラスを推定したいサンプル



k-NN による回帰分析

✓目的変数の値を推定したいサンプル $\mathbf{x}_{\text{new}}$ について、すべての
モデル構築用サンプルとの間でユークリッド距離を計算する

✓最も距離の近い k 個のサンプルを選択する

✓k 個の目的変数の値の平均値を、$\mathbf{x}_{\text{new}}$ の推定された値とする

✓k 個の目的変数の値の標準偏差で推定値の信頼度を検討できる

- 標準偏差が小さい (k 個の値がばらついていない) 方が、
標準偏差が大きい (k 個の値がばらついている) 方より
目的変数の推定値を信頼できる

k をどう決めるか？(クラス分類・回帰分析)

✓クラス分類、回帰分析ともに、k の値を 1, 2, 3, … として、
クロスバリデーションの結果が最も良好であった k の値とする

- クラス分類の例：クロスバリデーション後の正解率が最も高い k の値
- 回帰分析の例：r^2_{CV} が最も高い k の値

✓クロスバリデーションや r^2_{CV} についてはこちら
http://datachemeng.com/modelvalidation/

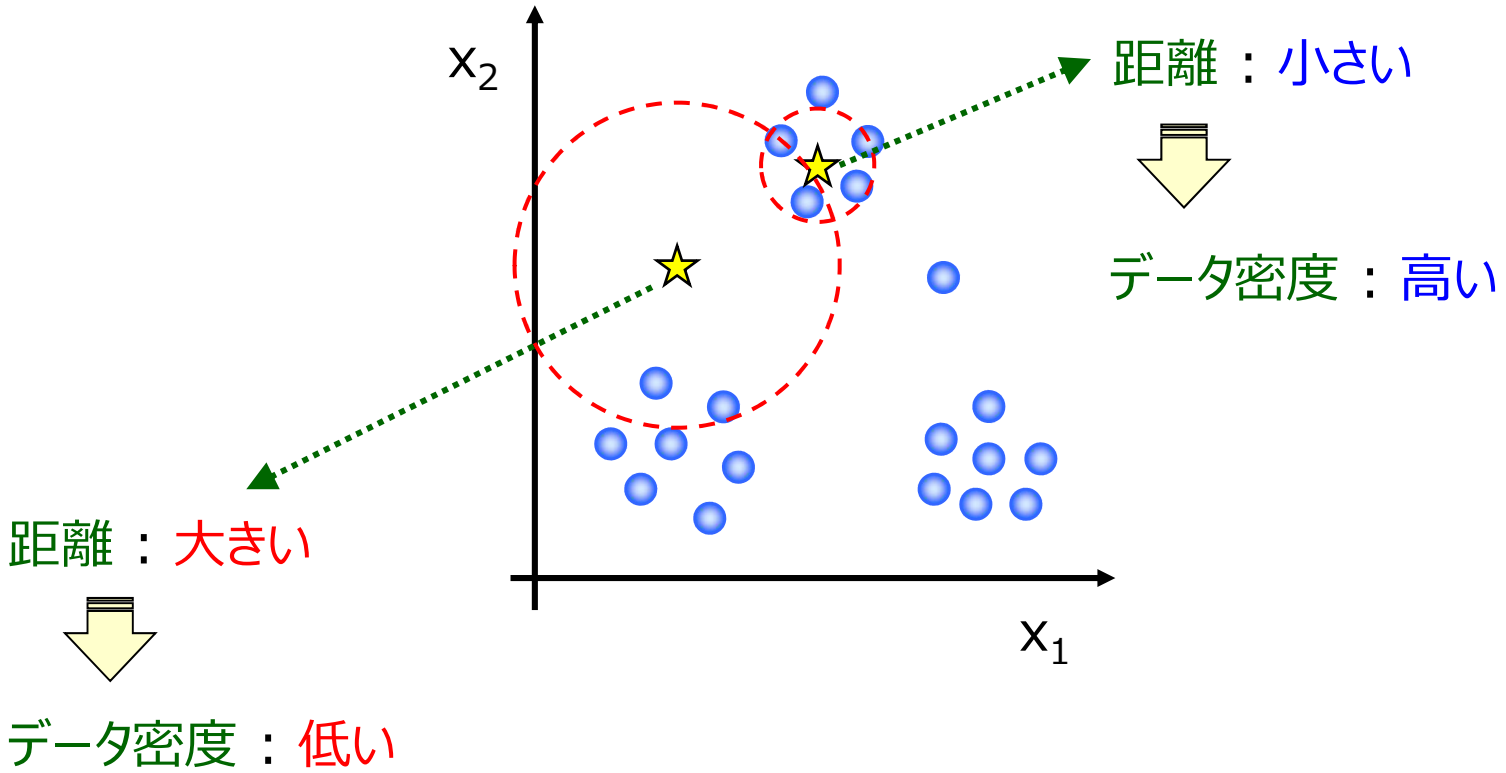
k-NN によるモデルの適用範囲の設定

データ密度が高い  モデルの適用範囲内

k-NN でデータ密度を推定する

k-NN によるモデルの適用範囲の指標

例) $k = 3$



k 個の距離の平均をモデルの適用範囲の指標とする

指標の値が小さいほど、適用範囲内

指標の閾値をどう決めるか？

- ✓モデル構築用サンプルにおいて、leave-one out クロスバリデーションで各サンプルの指標の値を計算し、その 99.7 % がモデルの適用範囲内となる値とする
  - 99.7 % は ３σ 法に由来
  - 99.7 % を小さくすると、適用範囲が狭くなる

k の値をどう決めるか？(モデルの適用範囲)

✓試行錯誤で決める

✓一般的には $k = 5$ とか $k = 10$

✓モデル構築用サンプルが少ないときは、k の値を小さくしたほうがよい

- 例：30サンプルのとき、$k = 1$

距離

✓ユークリッド距離だけでなく、

- マハラノビス距離
- チェビシェフ距離

など、いろいろな距離を利用できる

非類似度

✓距離を非類似度ととらえる

✓いろいろな類似度の指標を利用できる

- tanimoto 係数
- コサイン類似度

など