AIエンジニア研修

機械学習の基礎

アウトライン

1. 機械学習の概要と種類
2. 教師あり学習
3. 教師なし学習
4. ライブラリとプラットフォーム

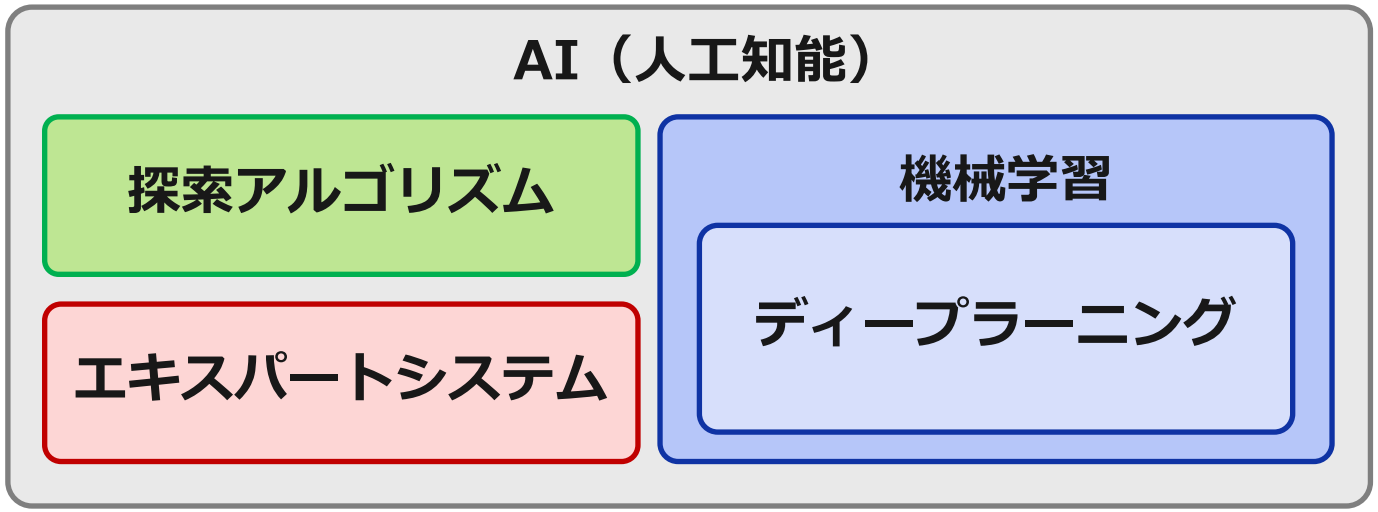
演習内容　Pythonを用いた機械学習実装

１．機械学習の概要と種類

機械学習の位置づけ

■AIとして開発される技術は主に次の3つのいずれかに属する

- **探索アルゴリズム**：事前に決めたアルゴリズムに基いて探索を行うAI
- **エキスパートシステム**：事前に決めたルールのみに基づくAI
- **機械学習**：結果から学習を行うAI　(ディープラーニングは機械学習の一種)



機械学習の位置づけ

名称	定義	メリット	デメリット
探索アルゴリズム	答えの候補の中からもっともらしいものを探して評価し、答えを返す。 （例：迷路のゴールを見つける）	処理過程が人に分かりやすい、プログラムが作りやすい。	数学的に表現できる問題しか扱えない。
エキスパートシステム	事前に用意された**ルール**と回答の候補を比較しながら評価し、もっともらしい答えを返す。	文字などの**記号**で表現される問題に適用できる。 (例：発話内容へ回答する)	**大量のルールを人が記述**しなければならない。 例外の多い問題が苦手。
従来の機械学習	大量のデータからパターンを抽出し、**基準を作る**。回答の候補を基準と比較しながら評価し、もっともらしい答えを返す。	**ルールの記述が不要。** **情報量が多いデータを扱う問題**に適用できる。 （例：画像認識や音声認識）	**大量のデータを用意**しなければならず、計算時間も長い。 **データで注目すべき範囲（特徴）を指定**しなければならず、複雑な指定が困難。
ディープラーニング	大量のデータから**様々な特徴を見つけ**、それに合わせてパターンを抽出し、基準を作る。 回答の候補を基準と比較しながら評価をし、もっともらしい答えを返す。	**特徴**は**粒度の細かいものから荒いものまでが自動的に多種作られることで、基準の汎用性**が飛躍的に高まる。	**機械学習よりも更に大量のデータを用意**しなければならず、計算時間も更に長い。 新しい技術であり、**扱い方はまだ研究段階。**

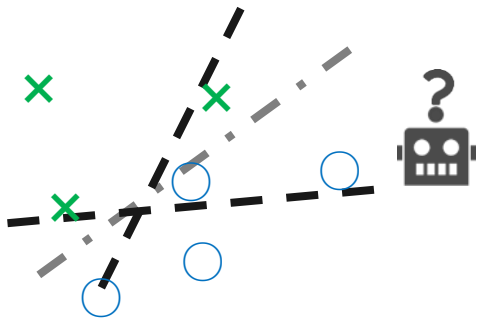
機械学習とは

機械学習: 広義には 「機械」 の 「学習」

- 与えられたデータ（入力）を元に対応する出力を推定すること
- 主に分類と回帰の２つ
- 入力から推定した出力の誤差が小さくなるように学習を行う
  - 下図の回帰の例では来客数から売上を推定し実測値とのズレが誤差になる

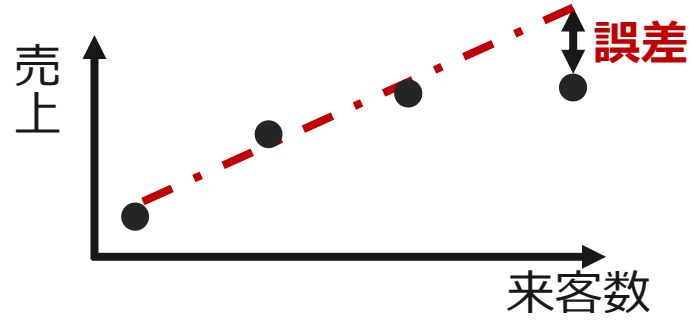
分類

上手く○と×を分ける線（式）を見つける



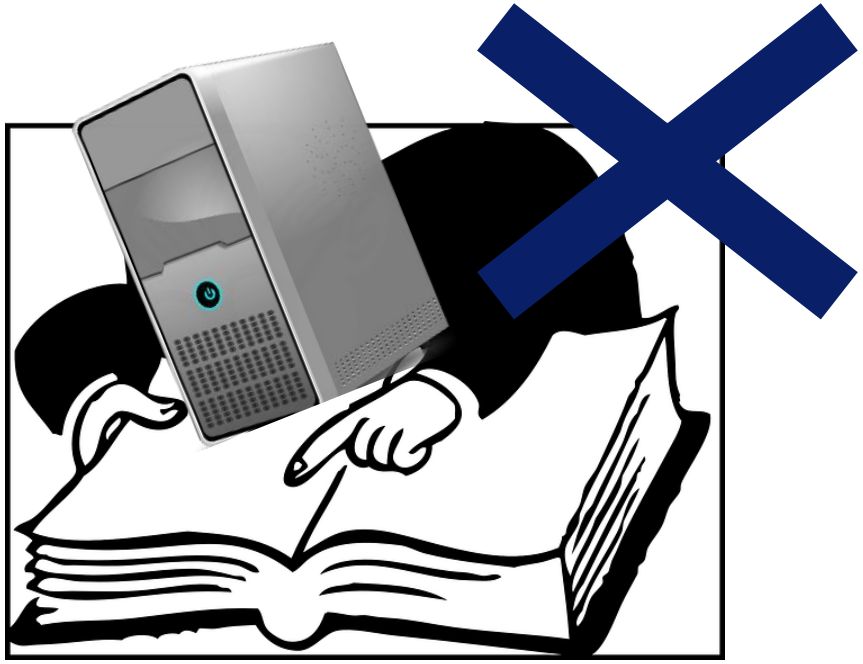
回帰

全ての点に最も近い線（式）を見つける

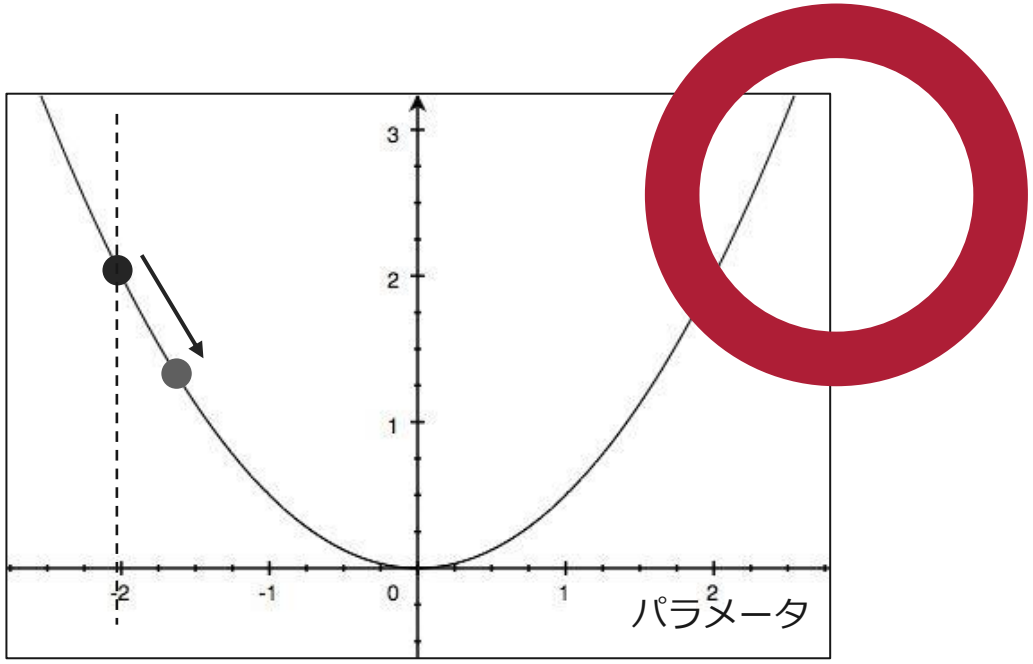


機械学習とは

機械のもつパラメータと呼ばれる数値を
より良い推定が行えるように決定、更新を行うこと



機械が一生懸命本を読んだり
囲碁を打ったりするわけではない



今よりも良い結果が求められるように
計算モデルの数値（パラメータ）を更新

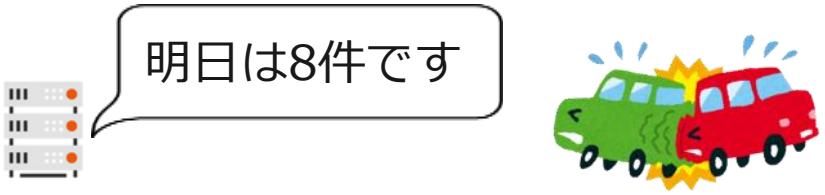
教師あり学習と教師なし学習

教師あり学習

正解のあるデータを用いて上手く答えを当てられるように学習する機械学習

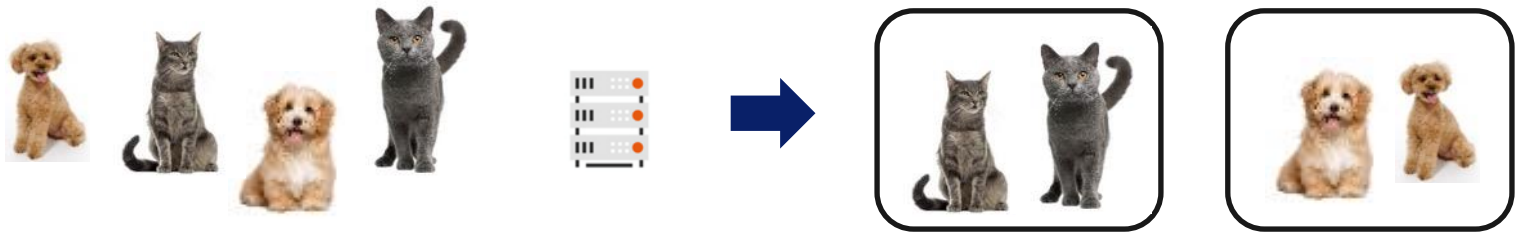
▷画像に写っている動物を識別する

▷ある日の交通事故の件数を推定する



教師なし学習

正解のないデータの中から類似性、規則性を見出そうとする機械学習



機械学習の種類

教師あり学習と教師なし学習の２種類

教師あり学習

データのラベルに基づく学習

・データに「猫」「犬」など予め意味付けを行うことをラベリングという
・データが入力されたとき対応するラベリングをデータの特徴量等に基づいて
　正しく推定できるようにする学習。

メリット

・学習や推定の結果が直感的に理解しやすく、チューニングもしやすい

デメリット

・学習データへのラベリングが事前に必要

教師なし学習

データの特徴を表す数値(特徴量)に基づく学習

・特徴を表す数値をもとに各データの近さを推定できるようにする学習
　　- 特徴とは猫の画像だと「目の間の距離」「耳の長さ」など

メリット

・学習データへのラベリングが不要（ラベリングが不要なだけでデータ自体は必要）

デメリット

・学習や推定の結果が直感的でないことがある

機械学習の「学習」と「テスト」

機械学習の処理は学習とテストの２つのフェーズに分かれる

学習

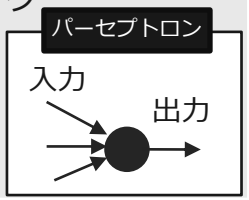
- 学習用のデータを用いてパラメータ更新を行い精度の向上を図る
- 入力 → 出力の計算に加え、パラメータ更新の計算を行うため
  テストより時間がかかる

テスト

- 学習で獲得されたパラメータ(モデルともいう)を使い精度検証を行う
- パラメータ更新を行わないため短時間で行える
- 実運用もテストと同じ計算のみを行うのが普通

機械学習の種類

教師あり学習

名称	定義	特徴	主な用途
線形回帰	全てのデータに最も近い点を通る直線を見つけてデータの予測をする手法	• 最も基本的な機械学習手法 • Excelの「線形近似」と同じ	比例関係にある数値データの予測、可視化
ニューラルネットワーク	脳の神経細胞を模した人工ニューロンをいくつも結びつけて学習させる手法	• パーセプトロンと呼ばれる人工の神経細胞ユニットを結びつけたネットワーク • 複数の入力を重み付けして足し合わせ出力を決める 	様々なデータの予測、分類やディープラーニング（後述）

教師なし学習

名称	定義	特徴	主な用途
クラスタリング	似通ったデータ同士をクラスターと呼ばれるグループにまとめる手法	• データさえあれば似たもののグループを見つけるのに便利 • クラスターへの意味付けまでは行えない • k-means法が広く使われている	ラベルの無い数値データやテキストデータの集約、レコメンドなど

2．教師あり学習

主な教師あり学習

- 回帰：数値を予測する
  - 線形回帰
  - ベイズ線形回帰

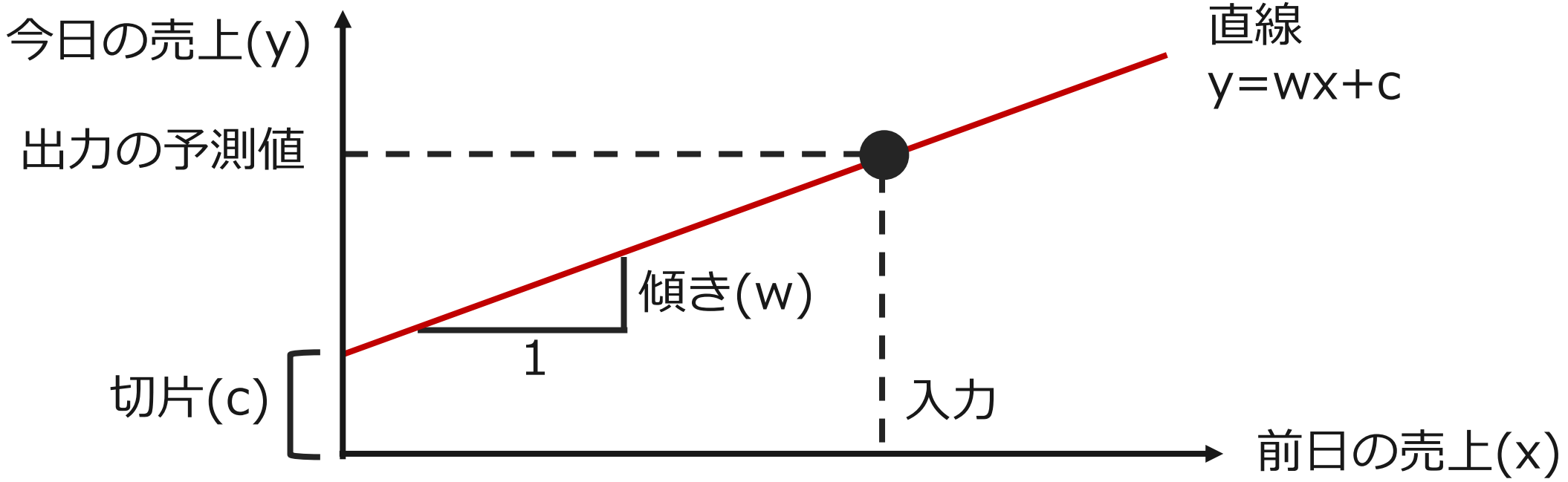
- 分類：クラスを判定する
  - ロジスティック回帰
  - 単純ベイズ分類器
  - 決定木
  - サポートベクターマシン（SVM）
  - ニューラルネットワーク

線形回帰

下記のような式を用いて、入力に対する出力の数値を予測する手法

$$y = w_1 x_1 + w_2 x_2 + \cdots + c$$

- xが入力、yが出力の予測値（x_1は前日の売上、x_2はその日が晴れかどうか等）
  - xを説明変数、yを目的変数と呼ぶ
- wとcを学習によって調整する

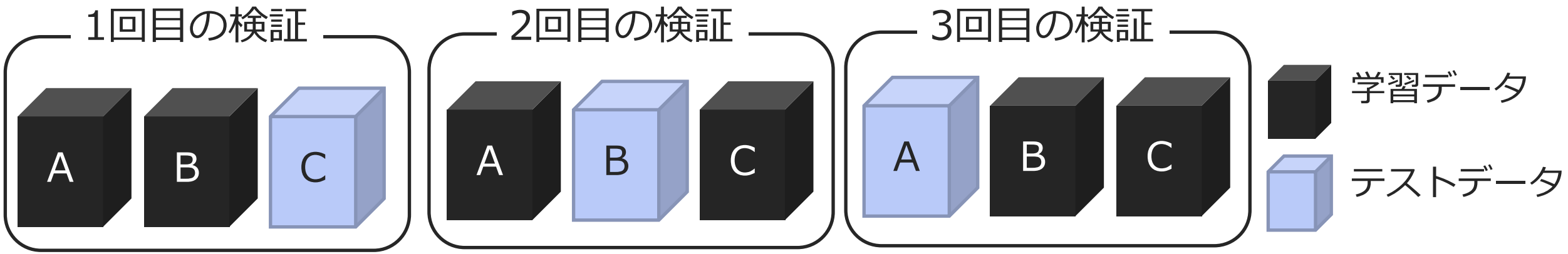


精度の検証

- 学習に用いる学習データと精度評価に用いるテストデータを分けて検証
  - 学習データで学習とテストを行うと過学習と呼ばれる問題が発生する（次項にて詳解）

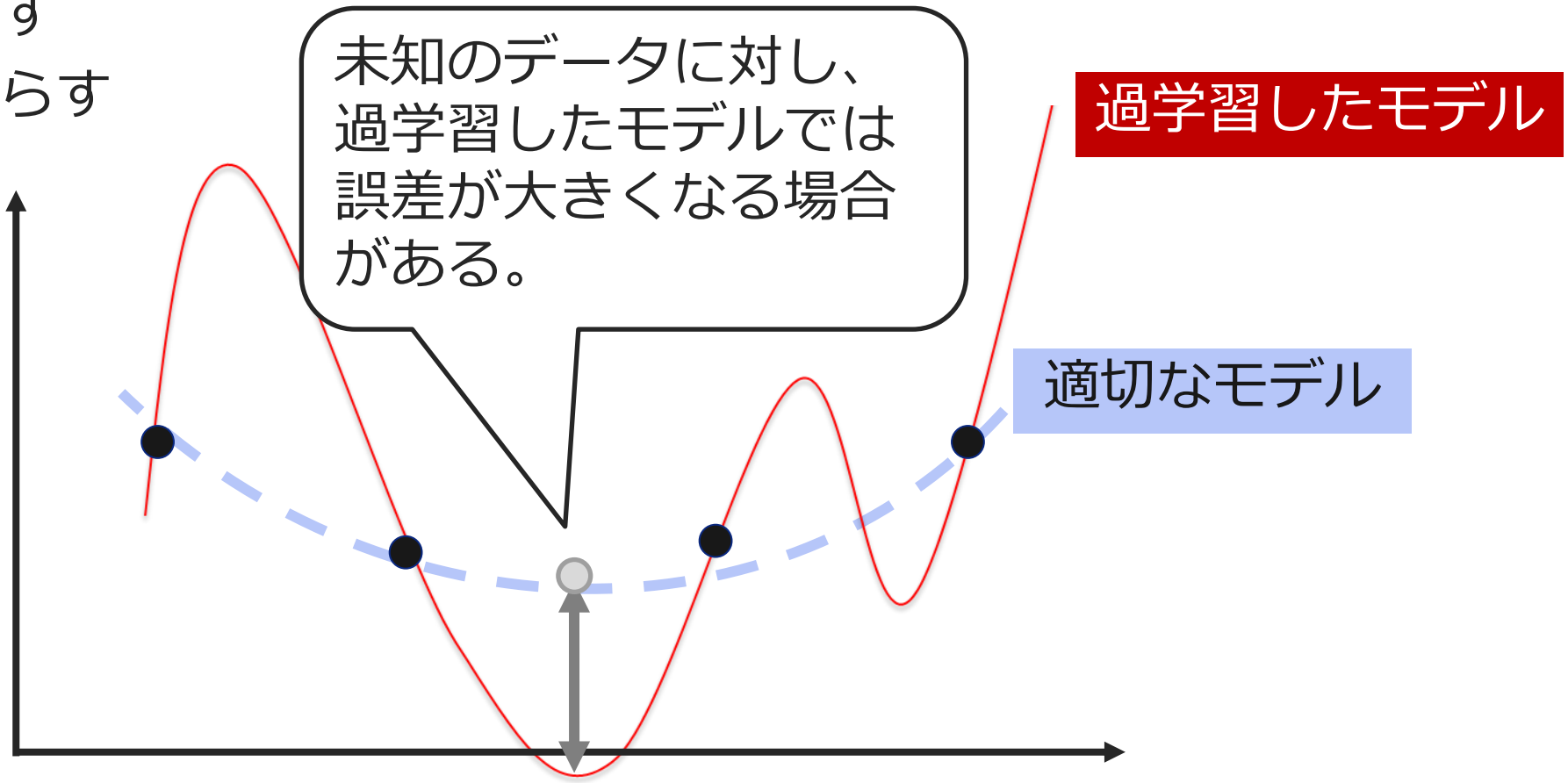
参考：交差検証

データをk個にグループ分けし、グループ毎に学習・テストデータを入れ替える



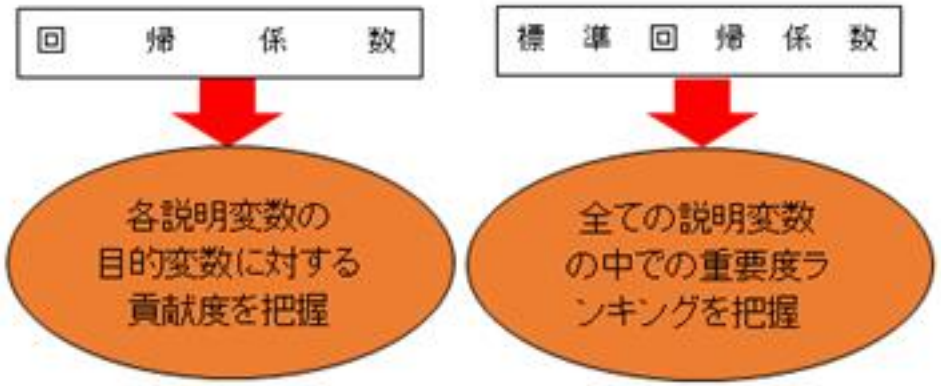
過学習

- 学習データに含まれていない未知のデータに対する精度が上がりにくいようなパラメータを学習した状態
- 主な原因として、説明変数に対して学習データが少なすぎる場合が考えられる
  - データを増やす
  - 説明変数を減らす

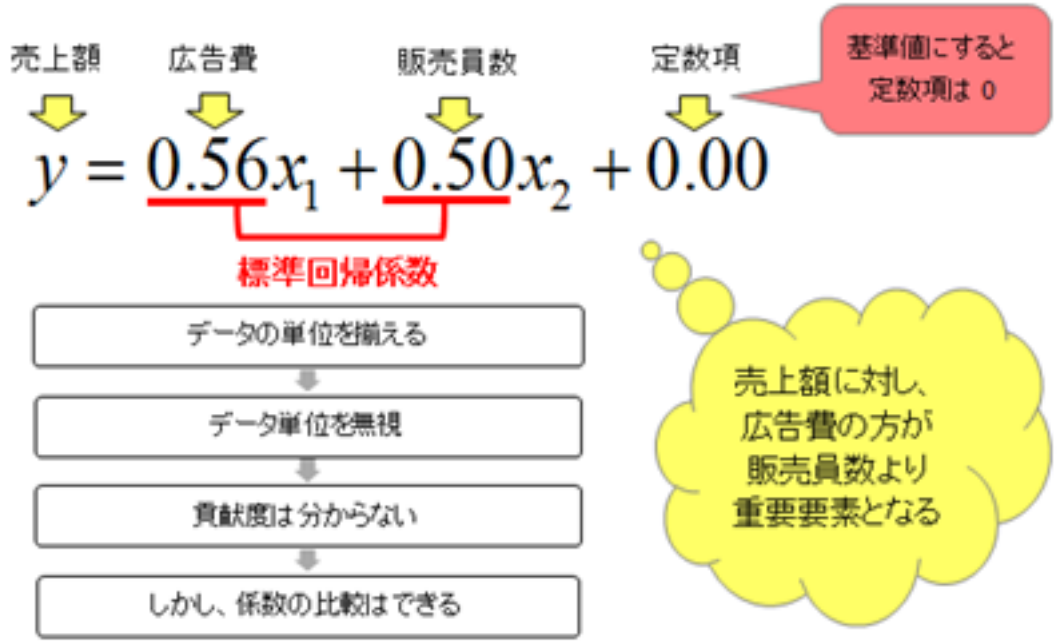


事例 | 商圏分析・売上分析

複数の説明変数を持つ線形回帰である「重回帰分析」は
商圏や売上の分析に広く使われている



	標準回帰係数	寄与率
広告費	0.56	52.8%
販売員数	0.50	47.2%
合計	1.06	100.0%



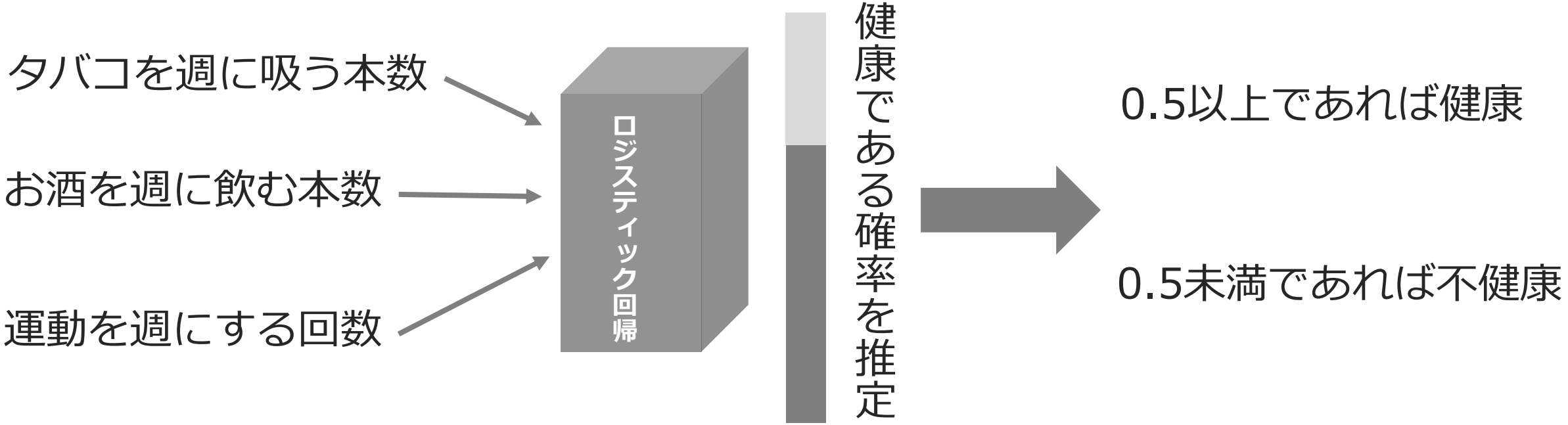
▶重回帰分析 - 株式会社アイスタット

ロジスティック回帰

線形回帰を確率として表現できるようにしたモデル

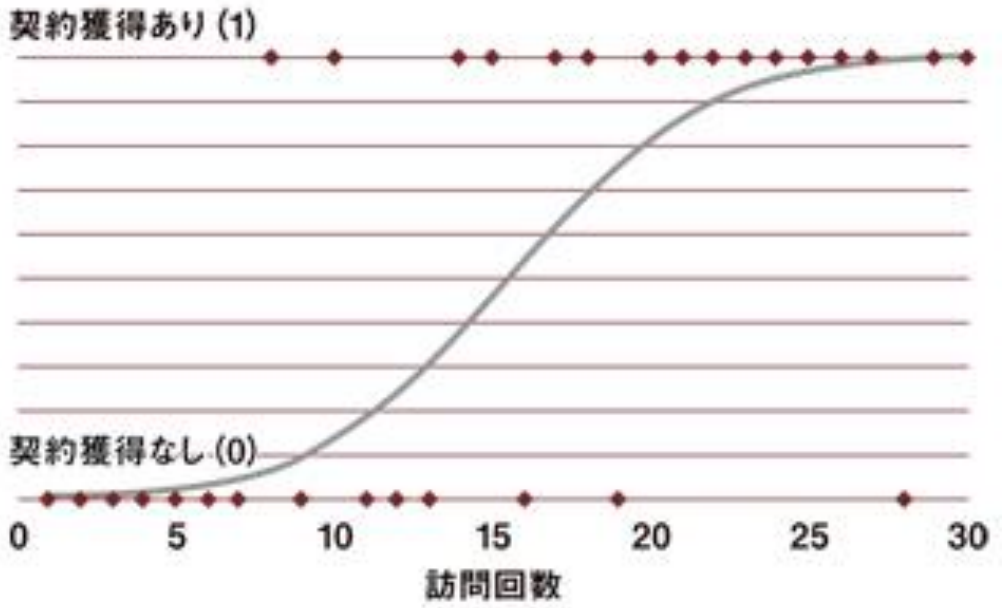
$$\ln\left(\frac{p}{1-p}\right) = w_1 x_1 + w_2 x_2 + \cdots + c$$

- 名前の通り回帰の手法だが、閾値を設けて分類に使われることが多い
- 手法としては後述のニューラルネットワークの元となる
  単純パーセプトロンと等価



事例 | 顧客の購買行動の推定

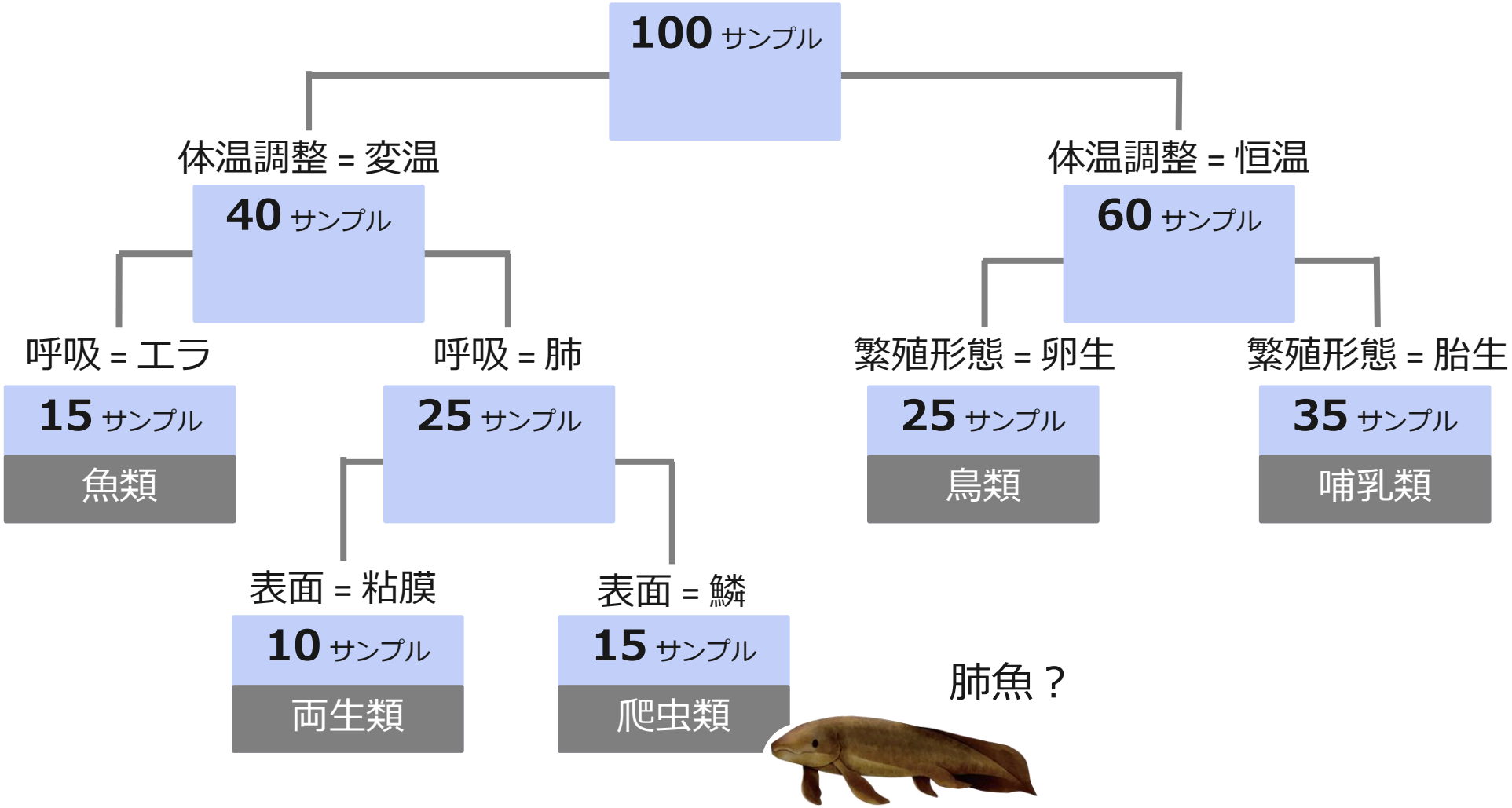
- 訪問回数に対して購買をする(1)かしないか(0)という入力に対して推定値が極端になるケースではロジスティック回帰が使われる
- 購買行動の他には「再入院するかどうか」や「画像の機械が故障しているかどうか」などの二値分類によく使われる



リピーターになる要因は何か？ ロジスティック回帰分析で探る – 日経ビッグデータ

決定木

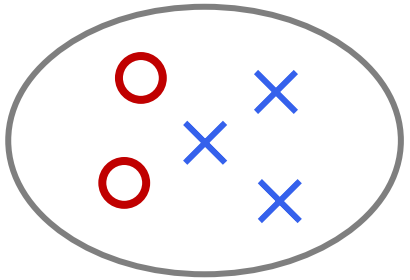
データを上手く分割する規則を見つけ出す手法



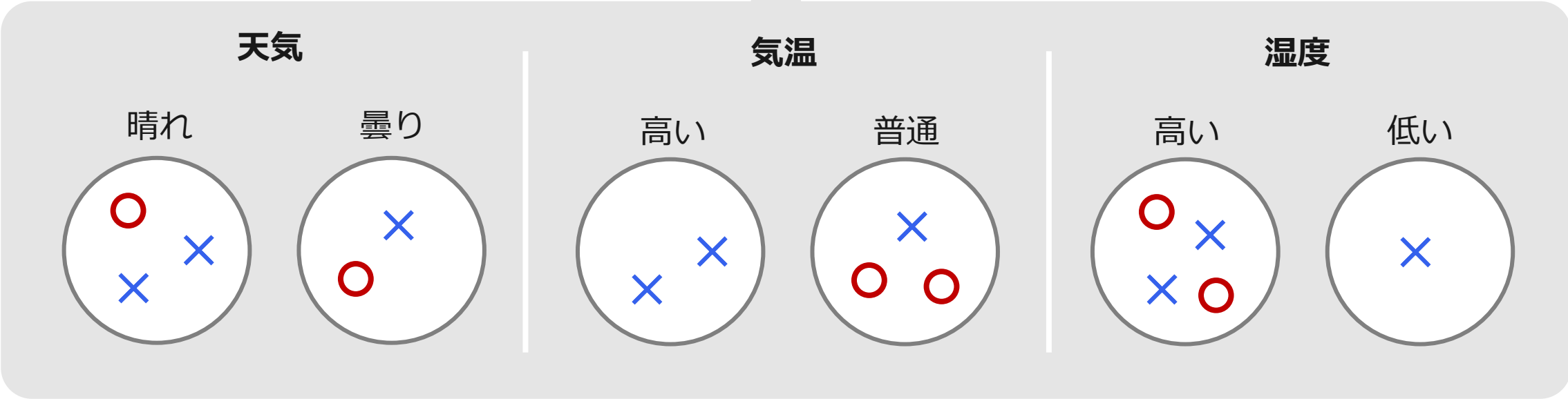
アルゴリズム

分岐条件に基づいて2分割する手続きを繰り返す。

例：
天気、気温、湿度を条件に、
ランニングする・しないを判定。

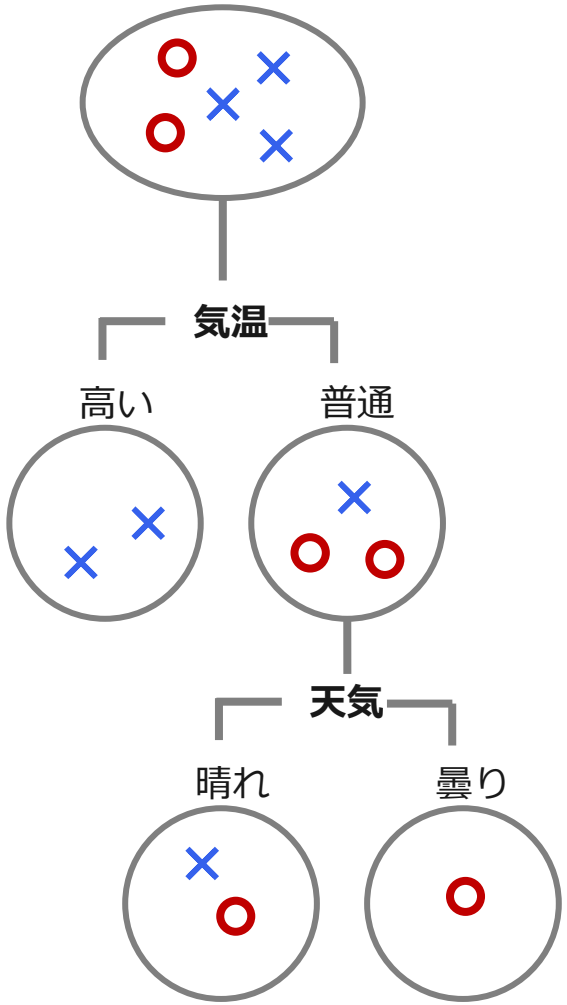


ランニング	天気	気温	湿度
しない	晴れ	高い	高い
する	曇り	普通	高い
しない	曇り	高い	低い
する	晴れ	普通	高い
しない	晴れ	普通	高い



アルゴリズム

最もきれいに分けられる基準からデータを分けていく



ランニング	天気	気温	湿度
しない	晴れ	高い	高い
する	曇り	普通	高い
しない	曇り	高い	低い
する	晴れ	普通	高い
しない	晴れ	普通	高い

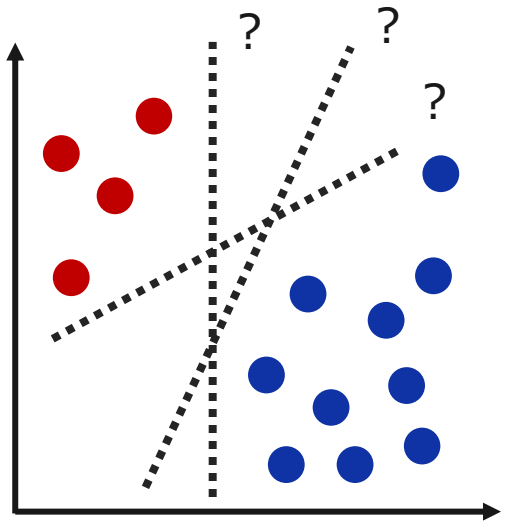
決定木が完成！

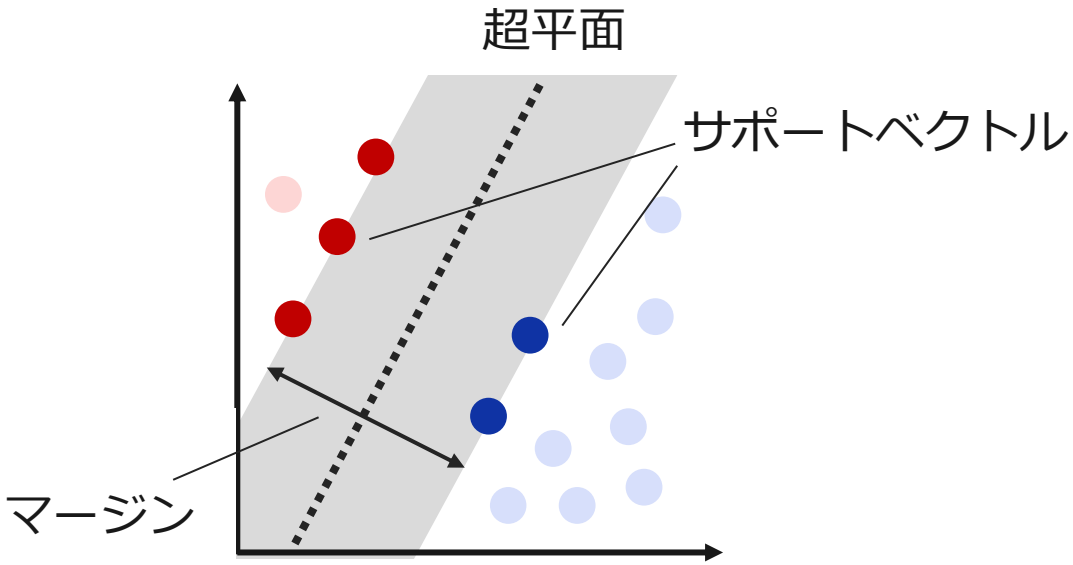
SVM（サポートベクターマシン）

境界線の近くにあるデータとの距離が最大となるよう、境界線を求めて分類する手法

特徴

- マージン最大化により汎化性能が高い

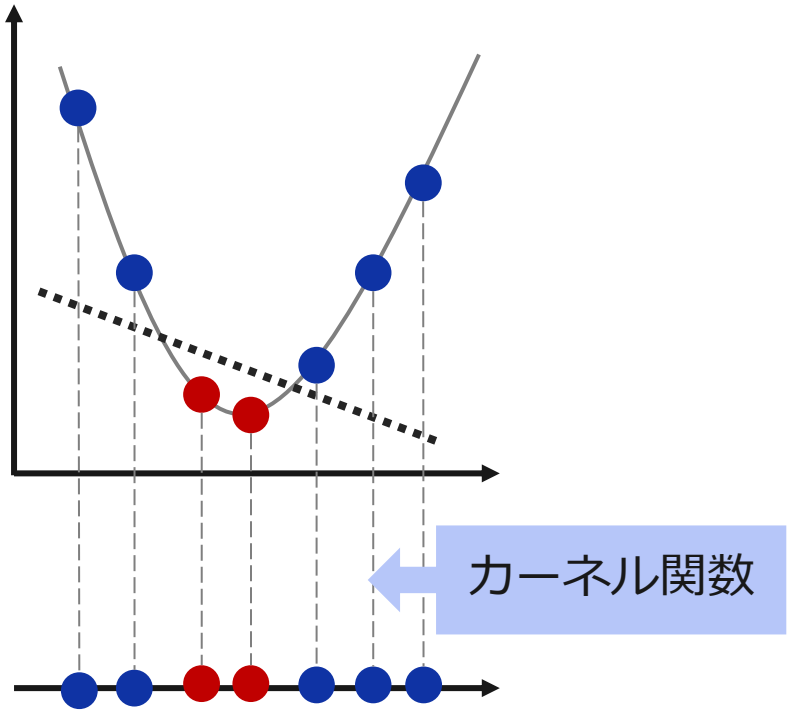


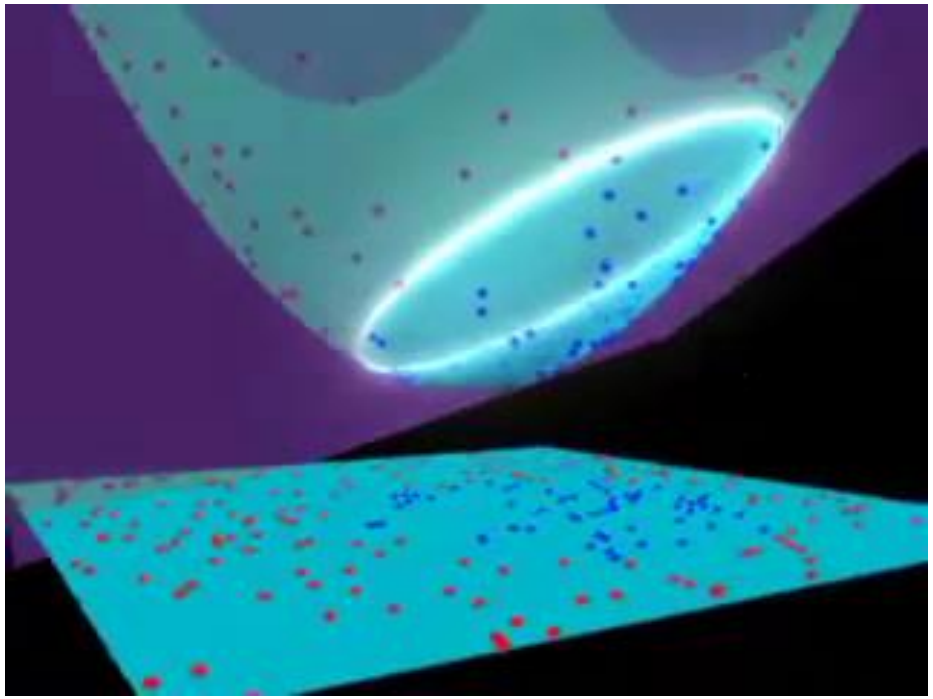


境界線の近くにあるデータ：サポートベクトル
境界線の近くにあるデータとの距離：マージン
境界線：超平面

SVM | イメージ

カーネルと呼ばれる関数を用いることによって、
データを分類しやすい形に変換し、境界線を求める。





▶ https://www.youtube.com/watch?v=3liCbRZPrZA

３．教師なし学習

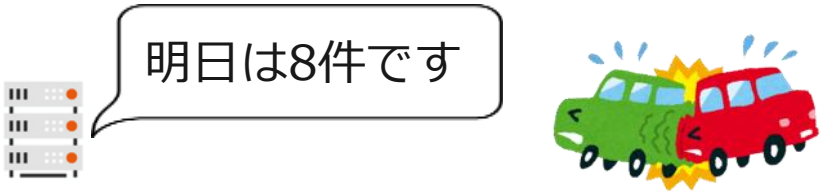
教師あり学習と教師なし（再掲）

教師あり学習

正解のあるデータを用いて上手く答えを当てられるように学習する機械学習

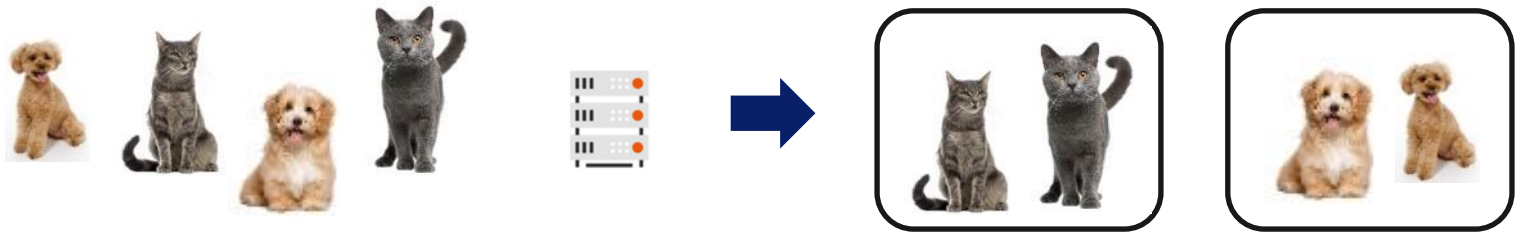
▷画像に写っている動物を識別する

▷ある日の交通事故の件数を推定する



教師なし学習

正解のないデータの中から類似性、規則性を見出そうとする機械学習

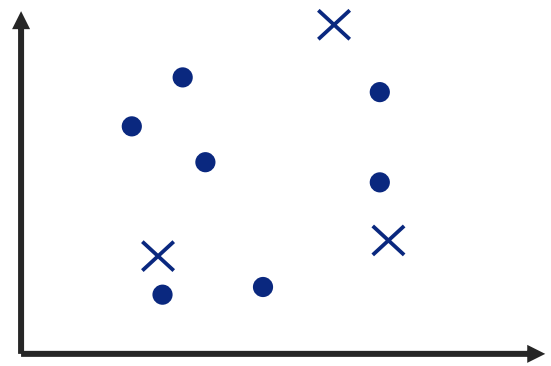


主な教師なし学習

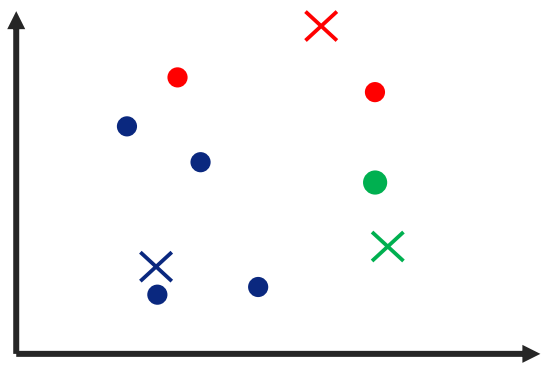
- k平均法（k-means、クラスタリング）
- 主成分分析（PCA: Principal Component Analysis）
- 自己組織化マップ（SOM: Self-organizing Map）
- Word2Vec
- t-SNE

k-means法

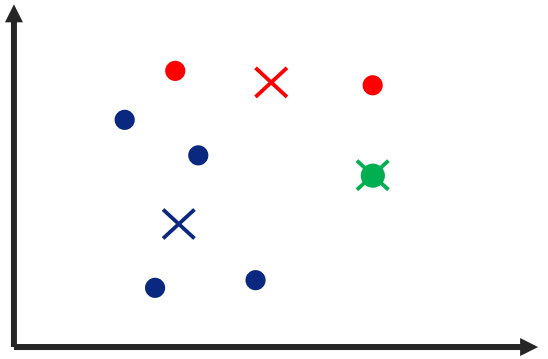
- データを複数のグループに分けるクラスタリング手法
- 「似ているもの同士をグループに分ける」という手法
  教師あり学習の分類と違って分けられたグループに意味があるとは限らない



与えられたk個の
ランダムな点を決める



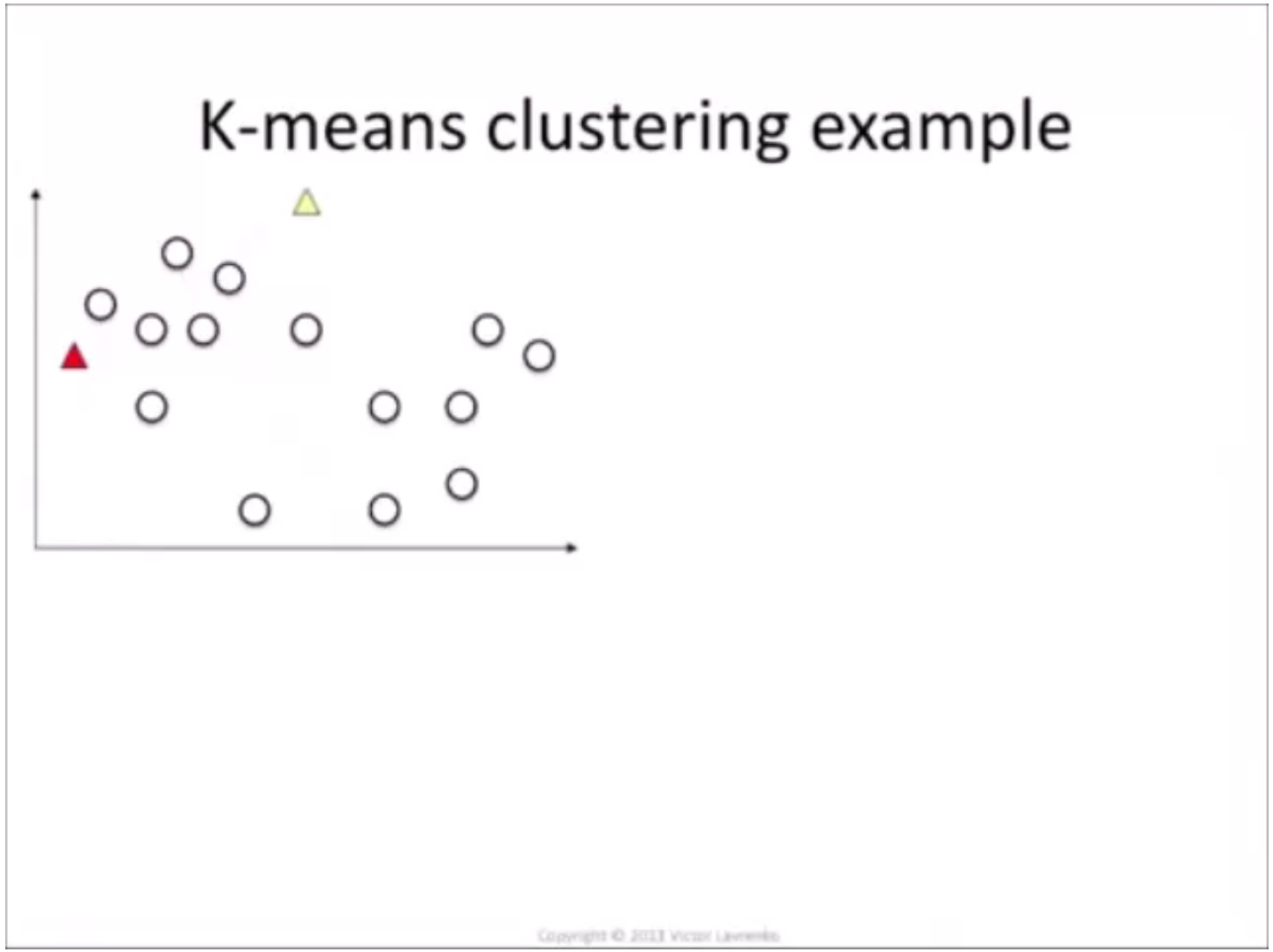
最も近い点の
クラスタに割り当てる



クラスタの重心に
点を更新する

割当が変わらなくなるまで繰り返す

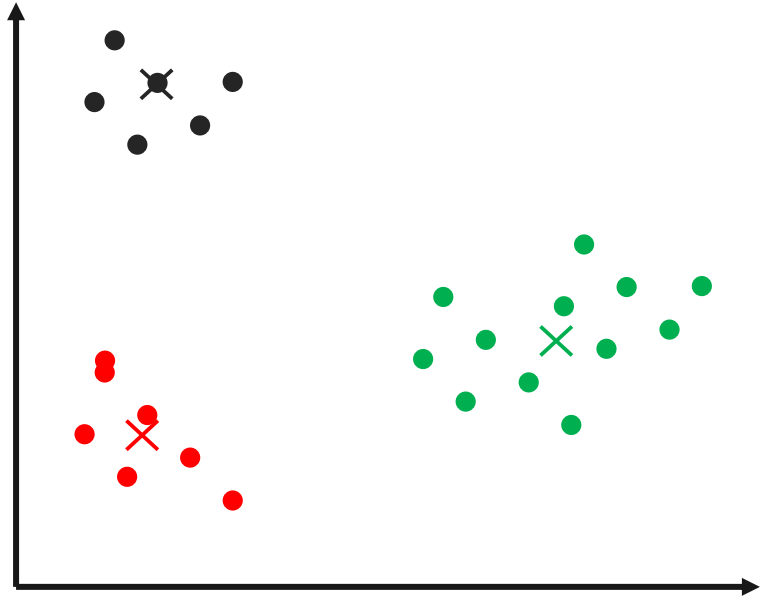
k-means学習の流れ



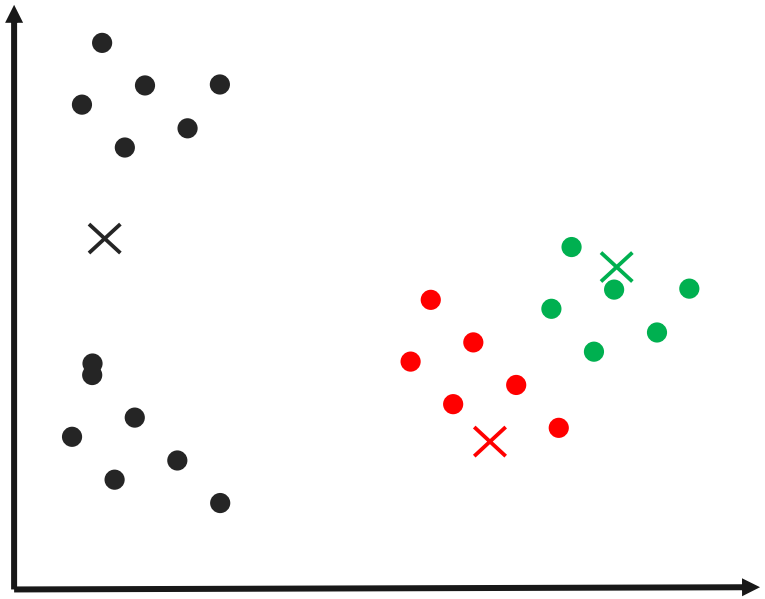
https://www.youtube.com/watch?v=_aWzGGNrcic
K-means clustering: how it worksより

k-means法の問題点

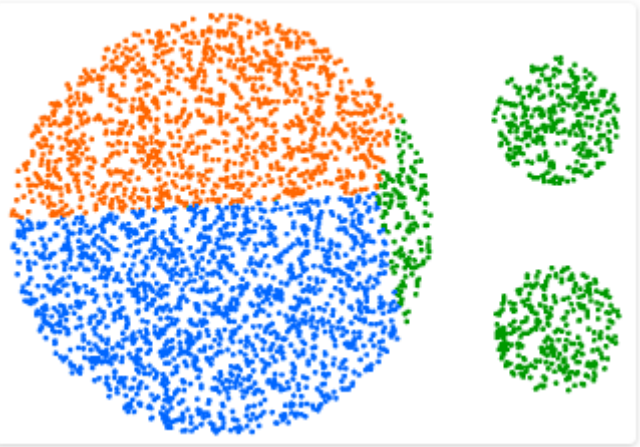
- 開始時に設定した重心次第でクラスタリングの結果が大きく異なる
- クラスタ内のデータ数に偏りが無いことを想定しているため
  データ自体に偏りがあると不適な場合がある



直感的には3つに分けるならこのような分け方が妥当と思われる



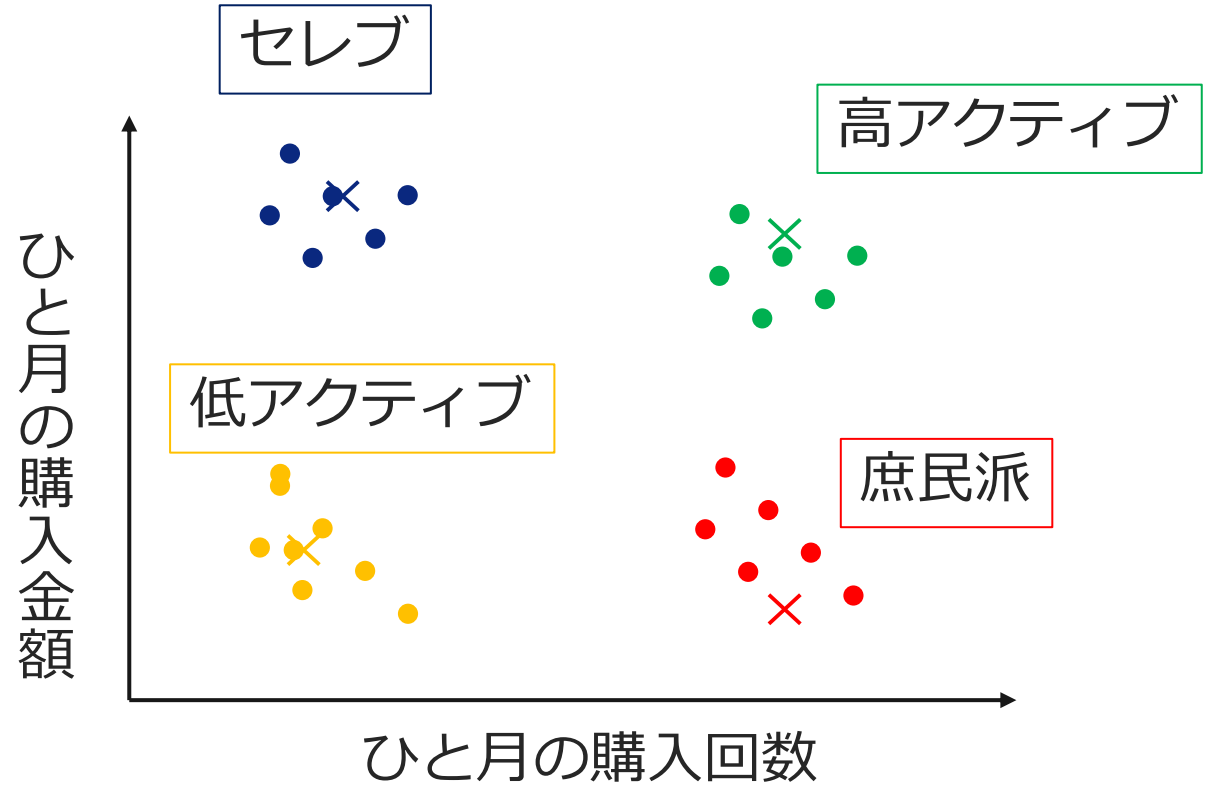
しかしランダムな点に偏りがあるとクラスタにも偏りができる



k-means 法に不適な事例[Guha 98]。データ数に偏りがあると不自然な分け方となる

クラスタリングにおける注意点

- クラスタの意味付けは人間が別途指標を決めて行う必要がある
  - 下図のようにクラスタに人が名前を付けるのも一例
  - データの特性から自動で付ける
    例えば文章データであれば頻出単語をラベルとするなど



kの決め方

- 重心の数(k)をいくつにするとよいか？
  - 答え：最も納得のいく結果を作り出すkが最適な値である
- クラスタリングはあくまでデータの傾向を一面から抽出する手法なので正解はない
  - 現実でも分け方がいくつも考えられる集団が存在する

    例）トランプの分け方

    k=2: 色で分ける(赤、黒)

    k=4: スートで分ける（スペード、ハート、クラブ、ダイヤ）

    k=13: 数で分ける（1,2,…,J,Q,K）

- 用途によって適切なkは異なる
  - 上記の例でk=3,5などとするとクラスタに偏りがでるため
    そのような場合は偏りが小さくなるkを設定すると良い

事例 | 顧客クラスタリング（GEO）

- 顧客の購買履歴から趣味・趣向を判別しクラスタリング
- クラスタの属性に合わせたクーポンやメルマガを配信することで販促に繋げている

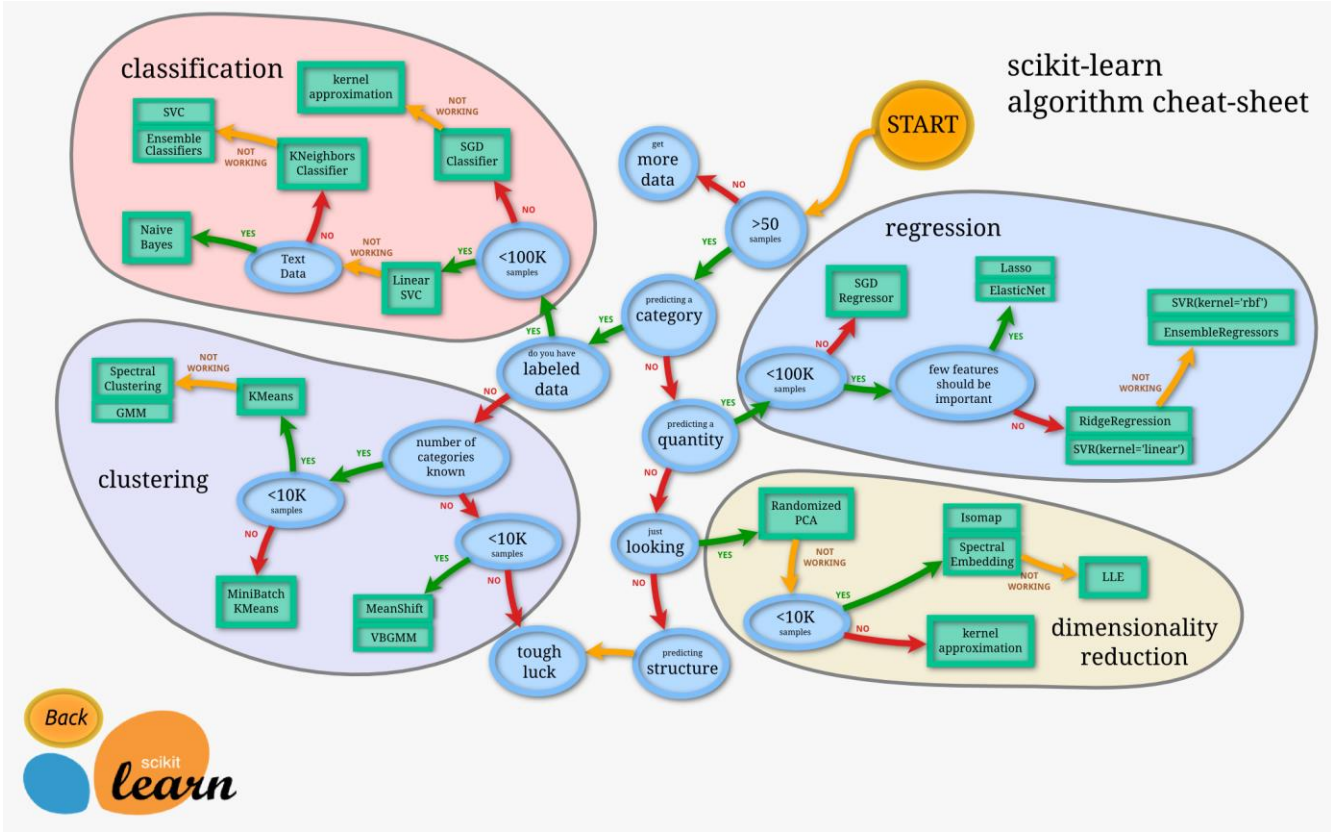
▶ゲオがビッグデータ活用へ、36歳社長の強烈な危機感 – 日経ビッグデータ

４．ライブラリとプラットフォーム

scikit-learn

- プログラミング言語Pythonで利用できるオープンソース機械学習ライブラリ
- Python上で高速に行列計算、科学計算が行えるライブラリNumpyやScipyと連携して計算を行う
- 回帰や分類、クラスタリング用の様々なアルゴリズムが用意されている

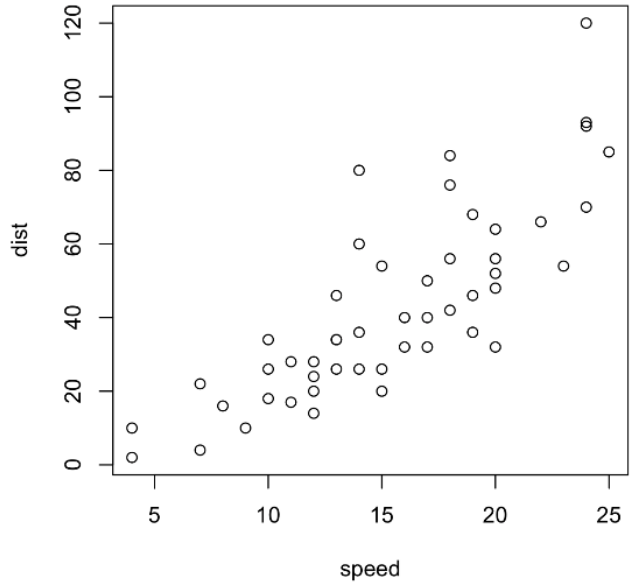
データの量やタスクの特性に応じて
適切な手法を選ぶ助けになる

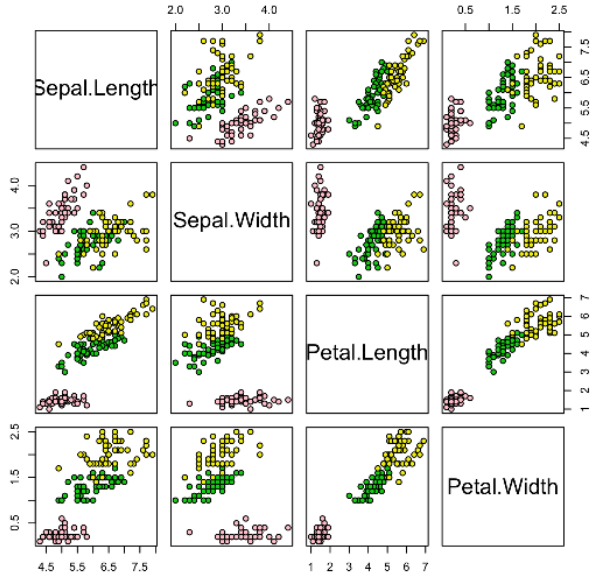


▷scikit-learn公式cheat-sheet（カンペ）
http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

R

- オープンソースの統計解析用プログラミング言語
- ベクトルや行列の表現や演算に長けており、機械学習用の組み込み関数も充実
- 結果の可視化も容易に行えるため学術、商用問わず幅広く使われている
  - 類似の商用ライブラリにMATLABがある





クラウドサービスの機械学習

Amazon Machine Learning

amazon web services™

二値分類、多クラス分類、回帰をサポートしているがアルゴリズムは基本的にロジスティック回帰のみ。
S3に格納されたデータを自動的に分析するバッチ機能を持つ。

Azure Machine Learning

Microsoft Azure

分類、回帰、クラスタリングに加えレコメンデーションなど幅広いアルゴリズムをサポートし、UIも充実。
プログラミング不要で複数のアルゴリズムを繋いで独自のモデルを作ることもできる。

Google Cloud Machine Learning

Google Cloud Platform

Azureと同じく様々な機械学習をサポート。画像認識や音声認識、翻訳のAPIなども含まれている上、TensorFlowやscikit-learnもサポートしており幅広い開発が可能。