情報Ⅱ オンライン学習会

～機械学習によるデータ分析～

早稲田大学　蓮池隆

今日の学習会では…

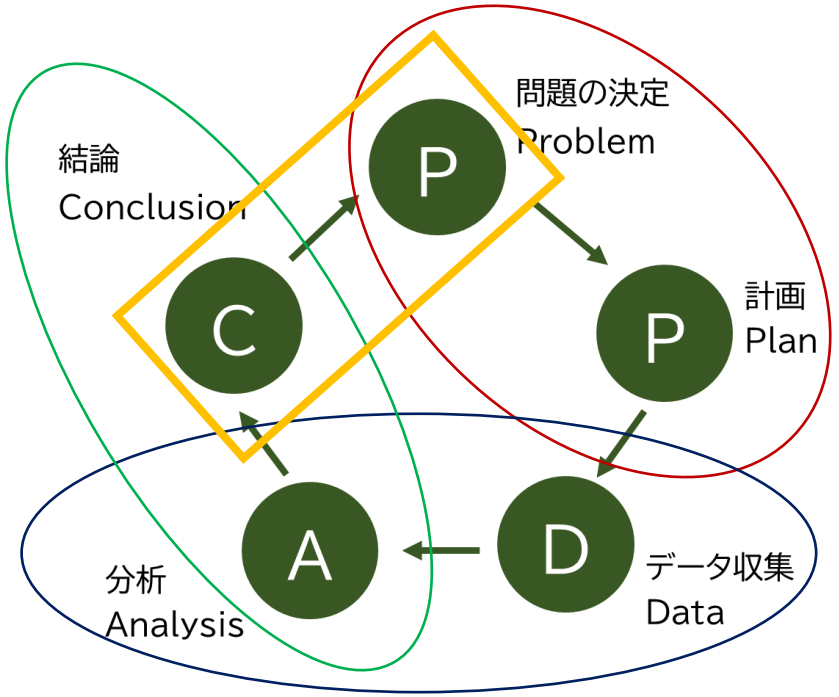
- 本日の学習会では，今後公開予定の情報IIに関する講義動画
  - 情報とデータサイエンス(3)：機械学習による分類『手書きの数字をコンピュータに認識させよう』
  - 情報とデータサイエンス(5)：ニューラルネットワークによる分類『より複雑な画像をコンピュータに認識させよう』
  - 情報とデータサイエンス(6)：モデルを用いた画像認識『自動で顔にぼかしを入れよう！』

  を基に，ポイントを解説していきます.
- 今回のポイントは「**機械学習をデータ分析に活かそう**」

データサイエンスとは…

- **データサイエンス**とは
  - さまざまな課題の解決や展望を予測するため，
  - 膨大に蓄積されているデータの内容やその分布を調べ，
  - 特定の傾向や性質に基づいた解析により，適切な解決方法を提示・評価する

(日本大百科全書(ニッポニカ))



一度で課題をすべて
解決できるのは稀

結果から新たな課題
に取り組む

昨日の学習会：統計的手法

- **統計的手法をデータ分析(何が重要？どれが効いてる？)に活かす**
  - 予測したい・目的(結果)に対して何が効いてるか知りたい　→　**回帰分析**
  - 変数を減らして(統合して)うまく説明したい　→　**主成分分析**
  - データをうまく分類したい　→　**クラスタリング**
  - 仮説が正しいかチェックしたい　→　**検定(推定)**

今日の学習会：機械学習

- **統計的手法をデータ分析(何が重要？どれが効いてる？)に活かす**
  - 予測したい・目的(結果)に対して何が効いてるか知りたい　→　**回帰分析**
  - 変数を減らして(統合して)うまく説明したい　→　**主成分分析**
  - データをうまく分類したい　→　**クラスタリング**
  - 仮説が正しいかチェックしたい　→　**検定(推定)**

- 統計的手法では複雑なモデルは扱いずらい(≒データが複雑になればなるほど精度を上げるには限界がある)
- 自然言語，画像などの非構造化データを扱う場合，統計的手法では困難
- どんな変数が効いているかを知るよりも，**とりあえず予測精度・分類精度を上げたい → 機械学習**

機械学習とプログラミング言語

- 機械学習を手っ取り早く始めるには，やっぱりPythonがよさそう
- Pythonだと機械学習のライブラリが豊富

（引用URL） https://www.codexa.net/machine-learning-python-library/

機械学習の進化は青天井!?

- 予測でも分類でも精度の高い「学習器」を作ることが重要
  - 学習器：データから機械学習モデルを学習するためのプログラム
  - 機械学習モデル：入力データに対して結果(=出力)を導き出す仕組み
- 現在でも**機械学習の進化・深化は止まらない！**(ChatGPTが最たる例)

機械学習の進化は青天井!?

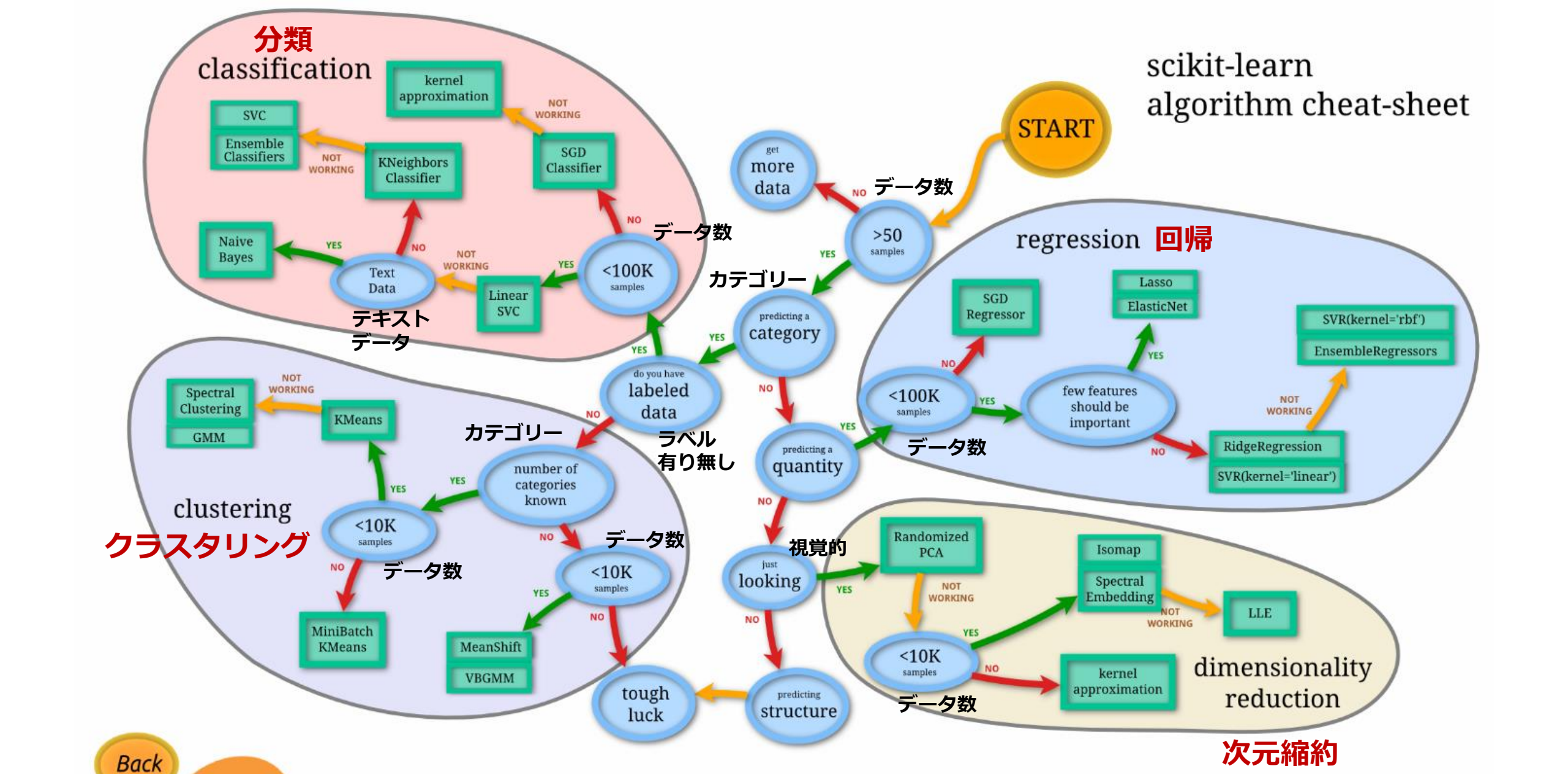
- 予測でも分類でも精度の高い「学習器」を作ることが重要
  - 学習器：データから機械学習モデルを学習するためのプログラム
  - 機械学習モデル：入力データに対して結果(=出力)を導き出す仕組み
- 現在でも**機械学習の進化・深化は止まらない！**(ChatGPTが最たる例)

- **ベースとなる手法を知っておけば，普通のデータ分析には対応可能(この一部が情報IIの内容)**

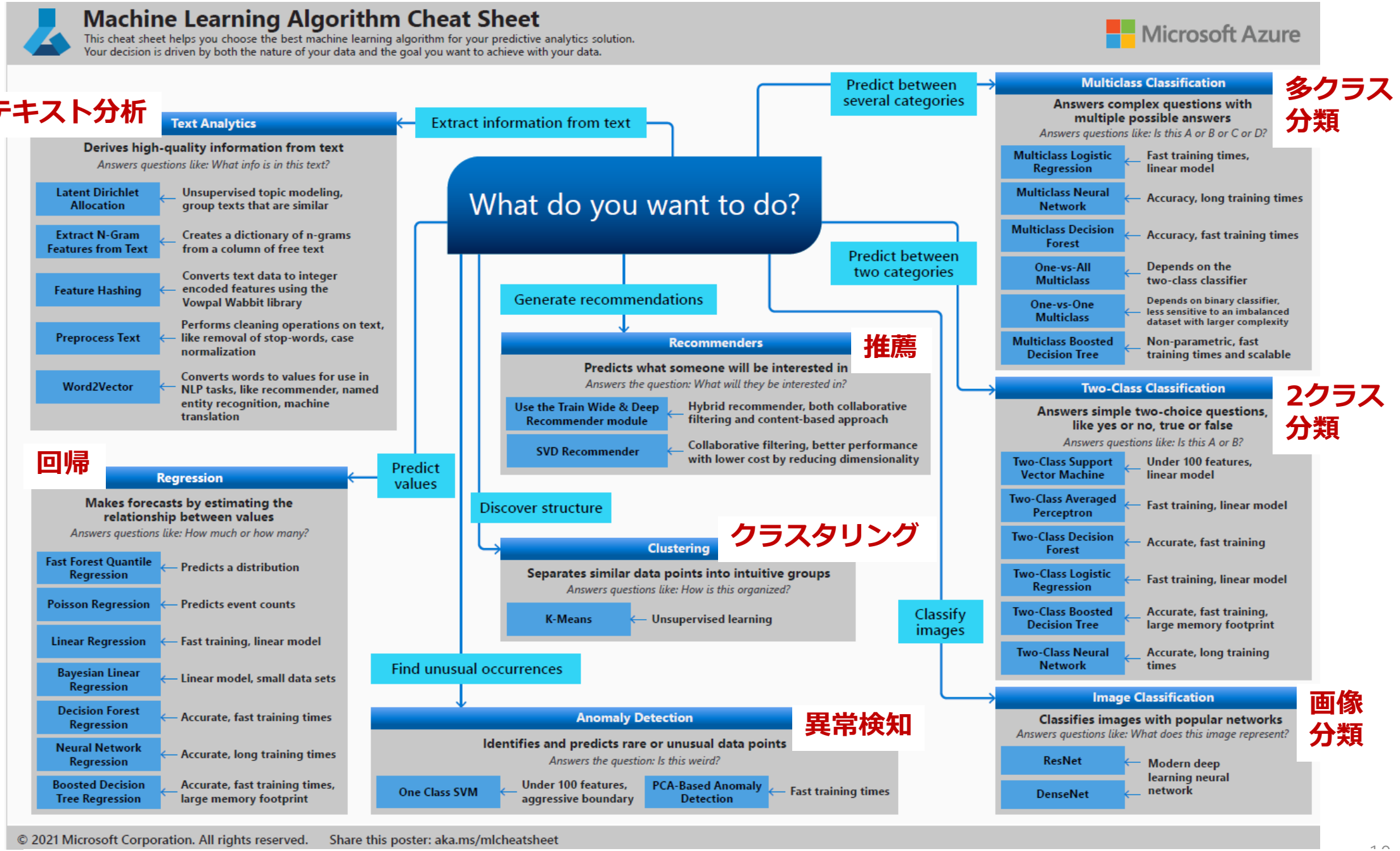
→ ベースとなる手法って何ですか？

ベースとなる手法も多いと，結局どう使えばよいかわかりません…

→ データ分析の目安となる「チートシート」と呼ばれるものが存在



(引用URL)　https://scikit-learn.org/stable/_static/ml_map.png



(引用URL)　https://learn.microsoft.com/ja-jp/azure/machine-learning/algorithm-cheat-sheet?view=azureml-api-1

情報とデータサイエンス

機械学習による分類

手書きの数字をコンピュータに認識させよう

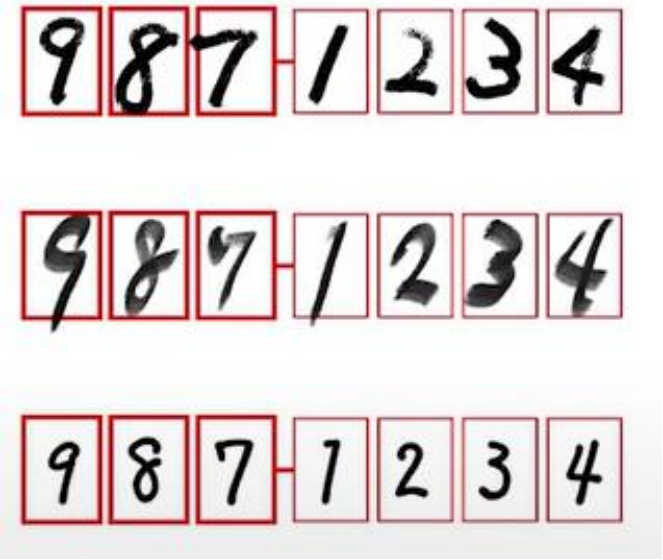
手書きの字を自動でコンピュータに読み込みたい

数字を画像として読み込むだけでなく，数値として扱うことができれば便利になる

例）手書きの郵便番号

→住所がわかり

配達先の仕分けができる

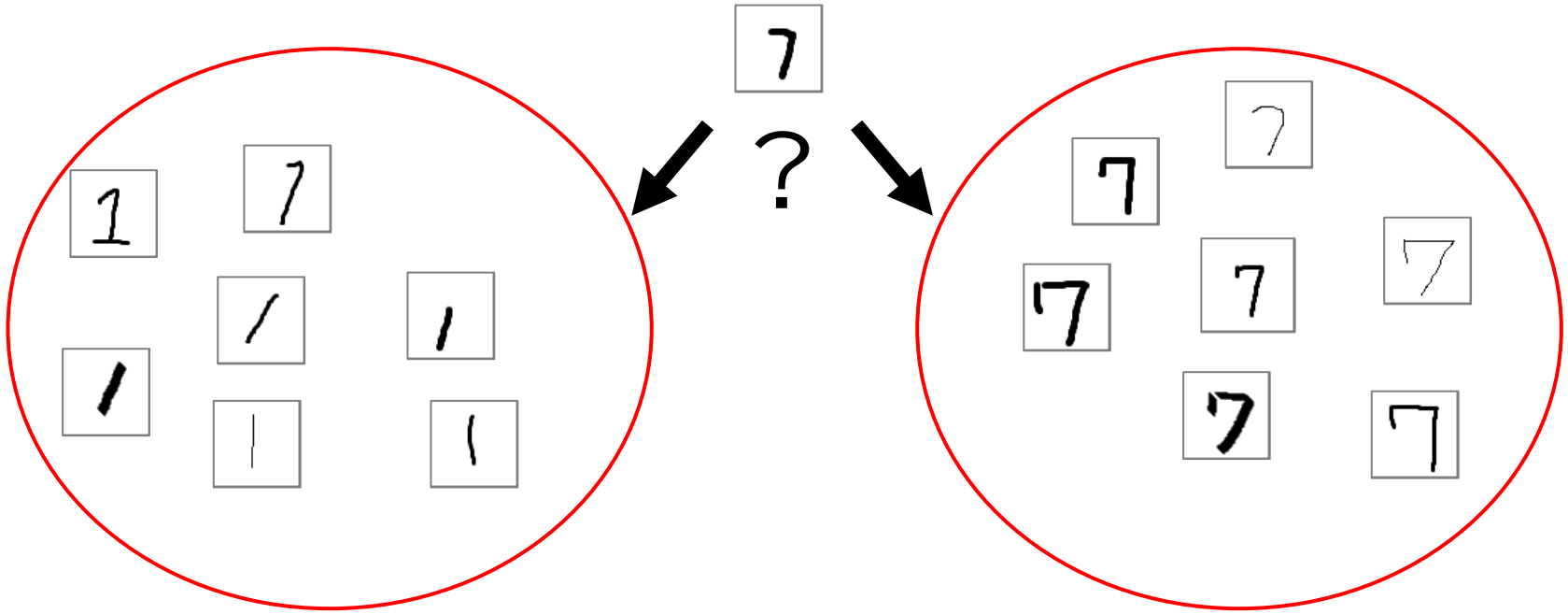


分類とは

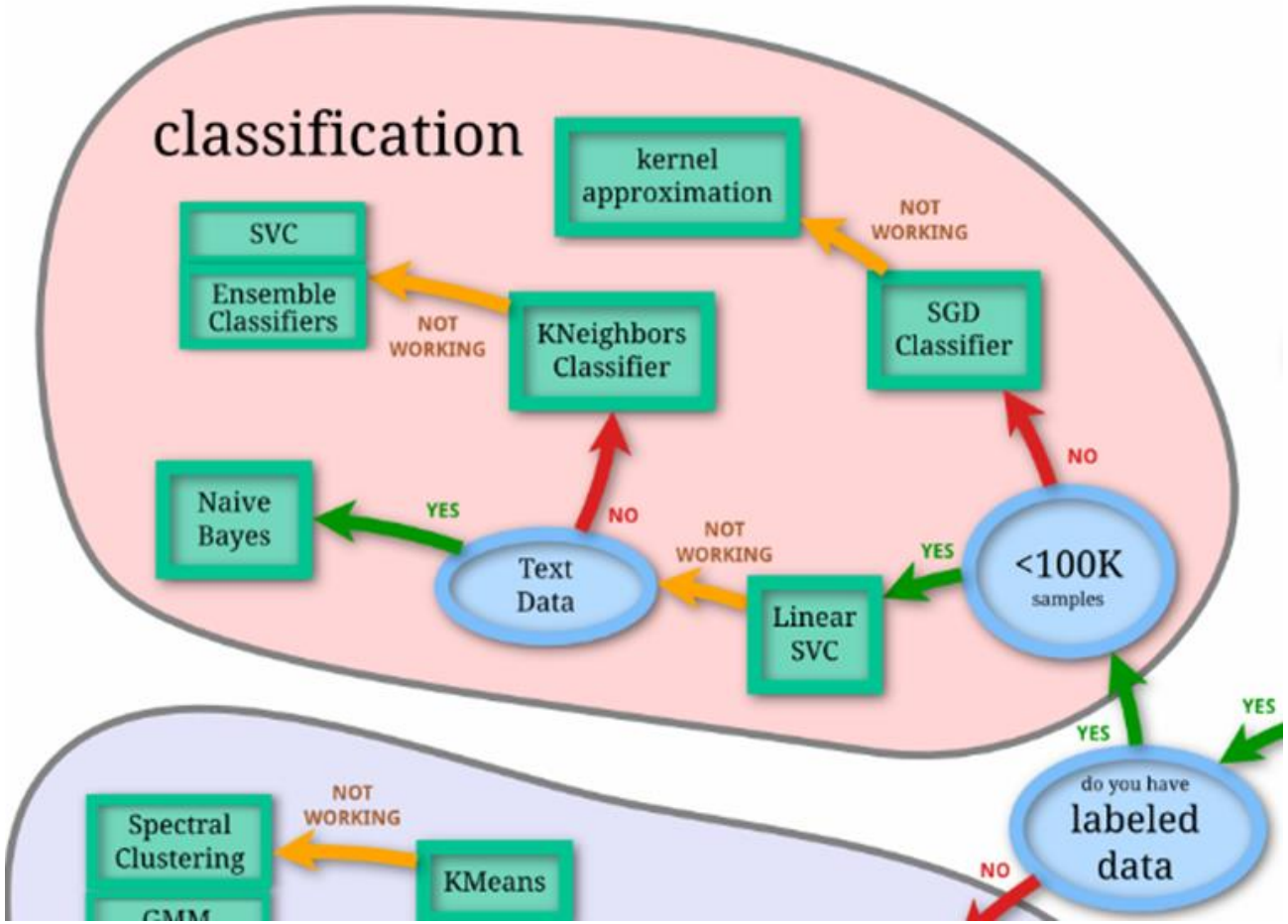
データがどのカテゴリーに属するかを予測する

→ そのために，事前に正解があるデータを使って学習しておく

（教師あり学習）



チートシートで確認



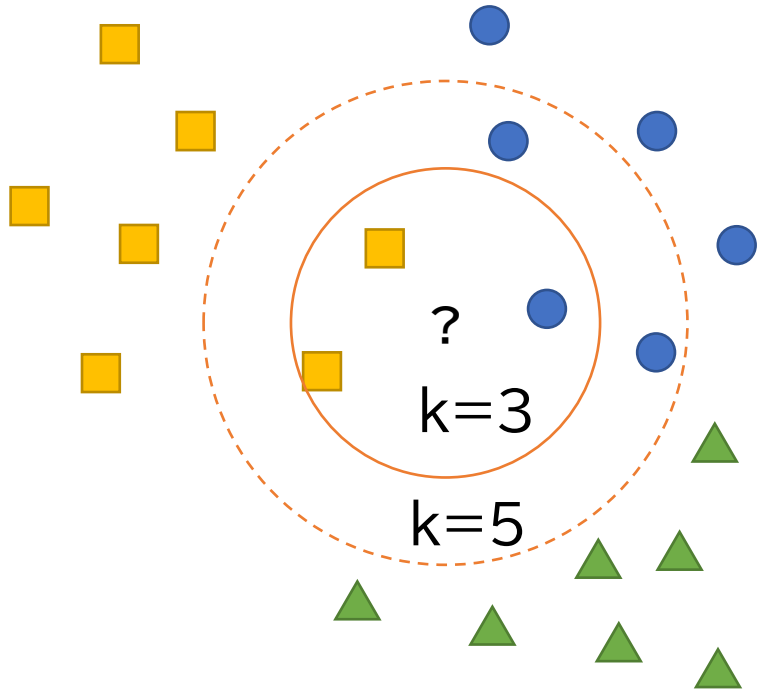
データ数が10万以下

- 線形の**サポートベクターマシン**による分類をまずやってみる
- うまくいかない時はテキストデータかどうか確認
  - テキストデータでなかったら，**k近傍法**を実施
  - テキストデータであれば，**ナイーブベイズ分類**を実施

分類のアルゴリズム①

☆　k近傍法

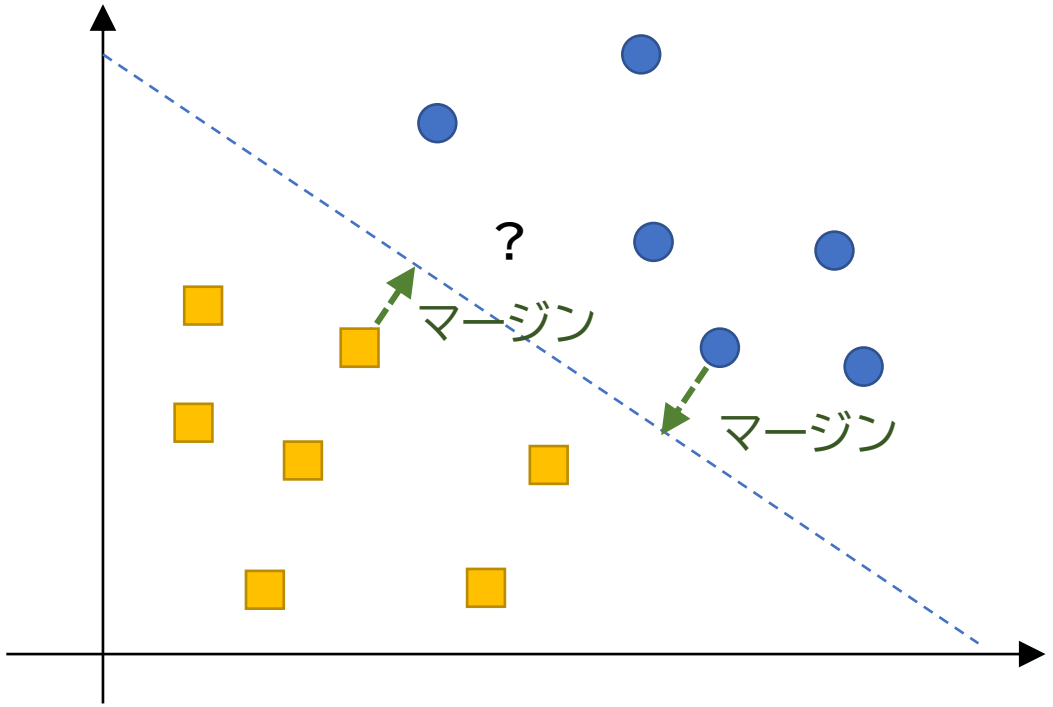
予測したいデータの近いk個のデータの中で
最も多いもののカテゴリーと推測する



分類のアルゴリズム②

☆　サポートベクターマシン(SVM)

カテゴリを可能な限り分割する境界線を見つけて、分類する。
このとき, マージンが最大になるように境界線を引く。

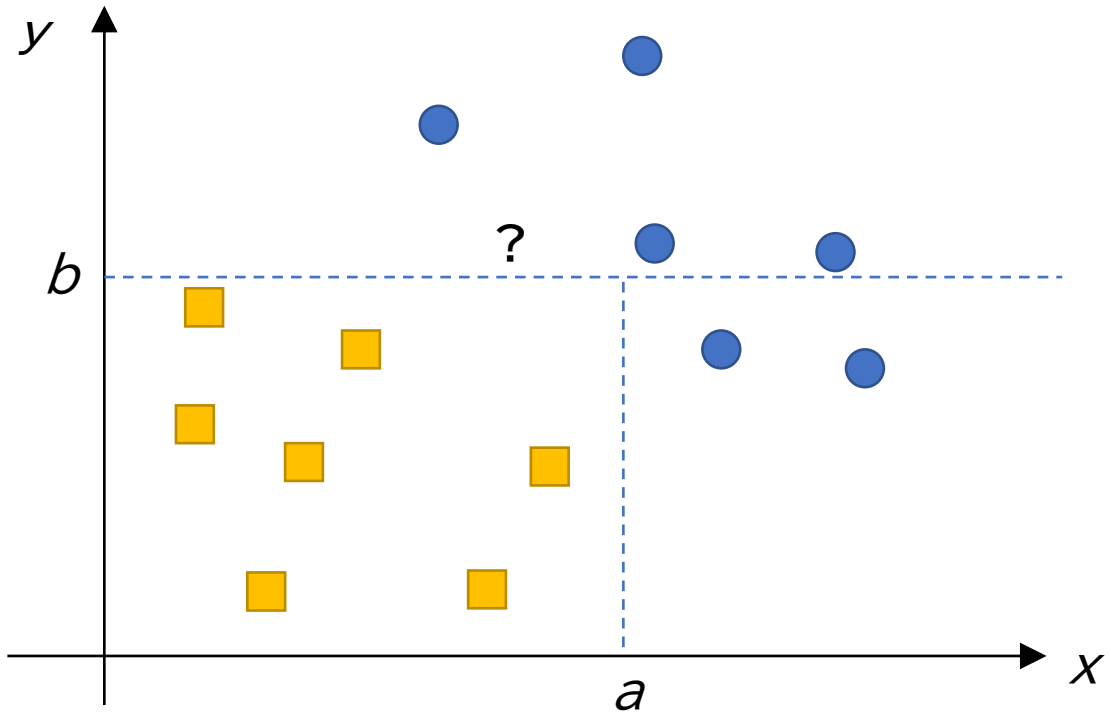


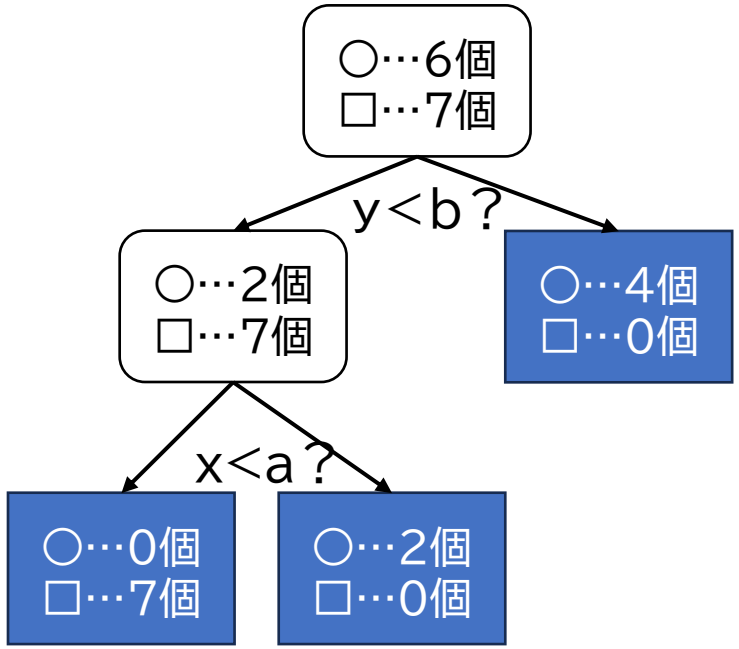
分類のアルゴリズム③

☆　決定木

段階的にデータを分割してデータを分類する

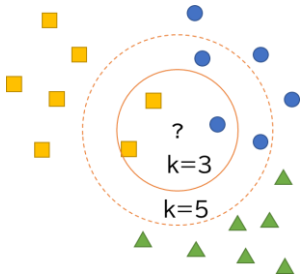
分析結果は木のような形状で示すことができる



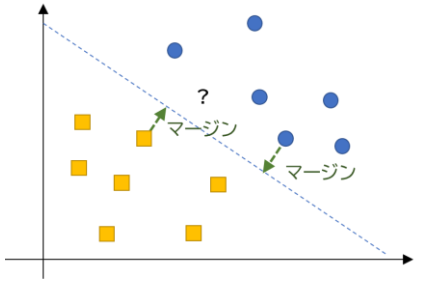


さまざまな分類手法

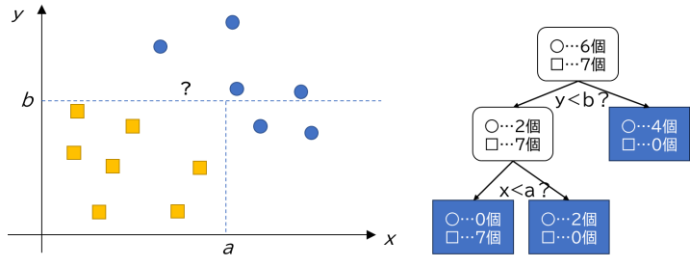
- **k近傍法**(k-NN)：予測したいデータの近いk個のデータの中で最も多いもののカテゴリーと推測する



- **サポートベクターマシン**(SVM)：カテゴリを可能な限り分割する境界線を見つけて分類する。このとき，マージンが最大になるように境界線を引く



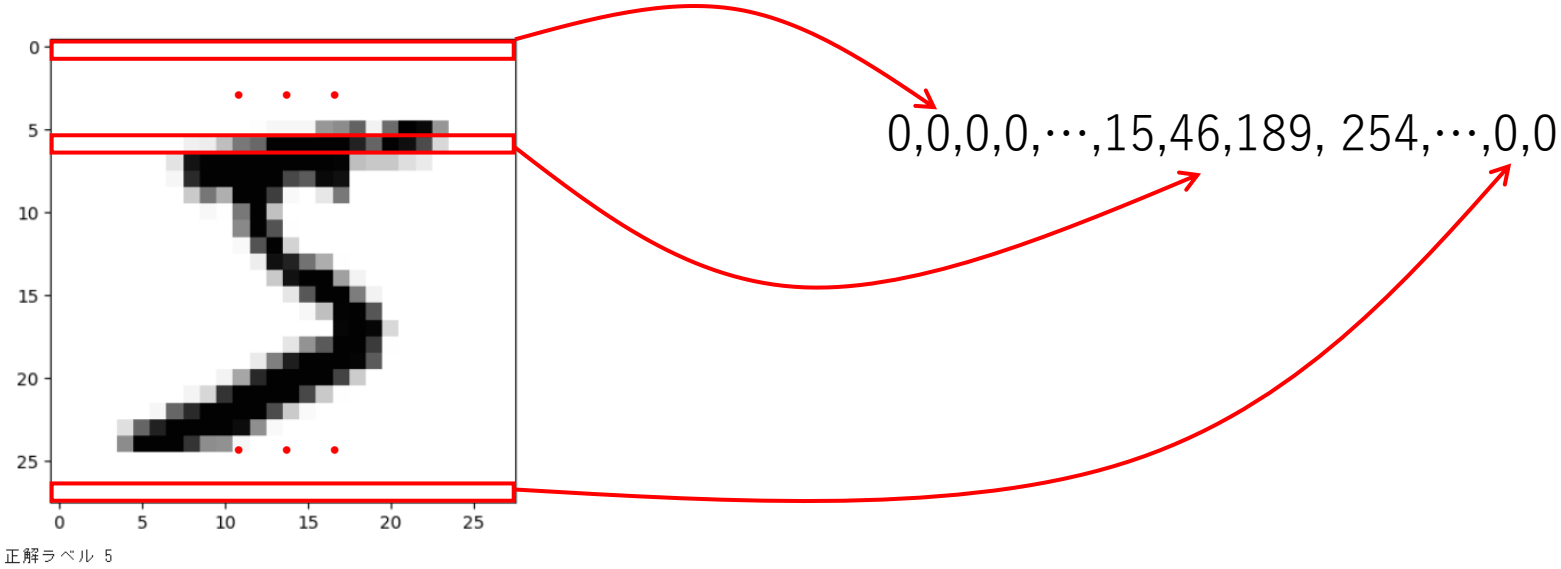
- **決定木**：段階的にデータを分割してデータを分類
  分析結果は木のような形状で表現可能



- 他にも，ナイーブベイズ分類器，ロジスティック回帰などたくさんある
  →学習データがどのような分布になっているかに合わせてアルゴリズムを選択

画像のままではうまく活用できない

- 画像を数値化(ベクトル化)する



- 数値化(ベクトル化)できれば，2つの画像の類似度を距離として表現できる
(類似度の計算(≒距離の計算)は数値であれば可能)

  → k近傍法などの距離を利用した分類が可能となる

分類のモデルの評価

正解率(Accuracy):正しく分類できた割合＝ $\frac{\square\text{の中の数の合計}}{\square\text{の中の数の合計}}$

右の表では $\frac{28+26+27}{28+4+5+0+26+3+1+2+27} \fallingdotseq 0.84$

		実際		
		A	B	C
予測	A	28	4	5
	B	0	26	3
	C	1	2	27

他にも

再現率(Recall):ある項目に着目し，真の正解(実際)のうち予測で当てられた割合

右の表でAに着目した場合，$\frac{28}{28+1} \fallingdotseq 0.97$

適合率(Precision):ある項目に着目し，予測結果のうち真の正解(実際)と一致した割合

右の表でAに着目した場合，$\frac{28}{28+4+5} \fallingdotseq 0.76$

音声や画像の解析が可能となれば…

- 文字や音声の認識　→　迷惑メールやオレオレ詐欺電話を推測
- 自動運転での交通状況の確認
  （信号、他の自動車、歩行者、車線などを区別して認識）
- 病気の発見　→　早期治療に結び付ける
- コンピュータウイルスの発見　→　情報セキュリティの向上

- いずれにしても，(精度を上げるとすれば)分類を行う際に相当数の学習用データが必要
- 文字や音声、画像を数値化することが必要
  (結局，計算機を用いた分類・予測において，音声や画像をそのまま扱うことはできず，数値化(数値を要素とするベクトル化)が必要

情報とデータサイエンス

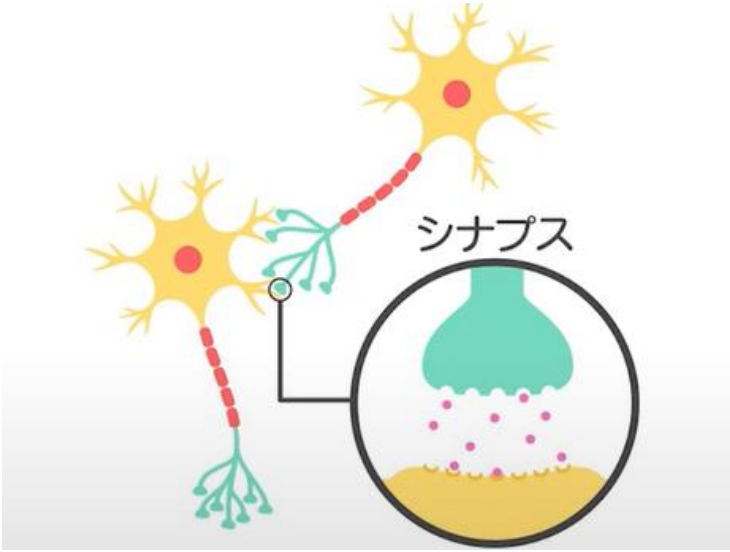
ニューラルネットワークによる分類

より複雑な画像をコンピュータに認識させよう

ニューラルネットワークとは

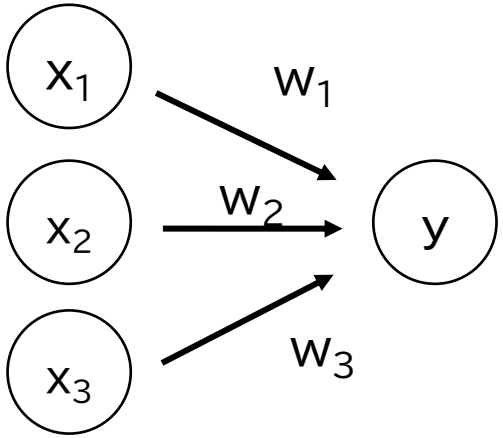
人間の脳内の神経細胞どうしのつながりを模した数理モデル

人間の脳内の神経細胞どうしのつながり



電気信号はシナプスで重みづけされて
一方向に伝播する

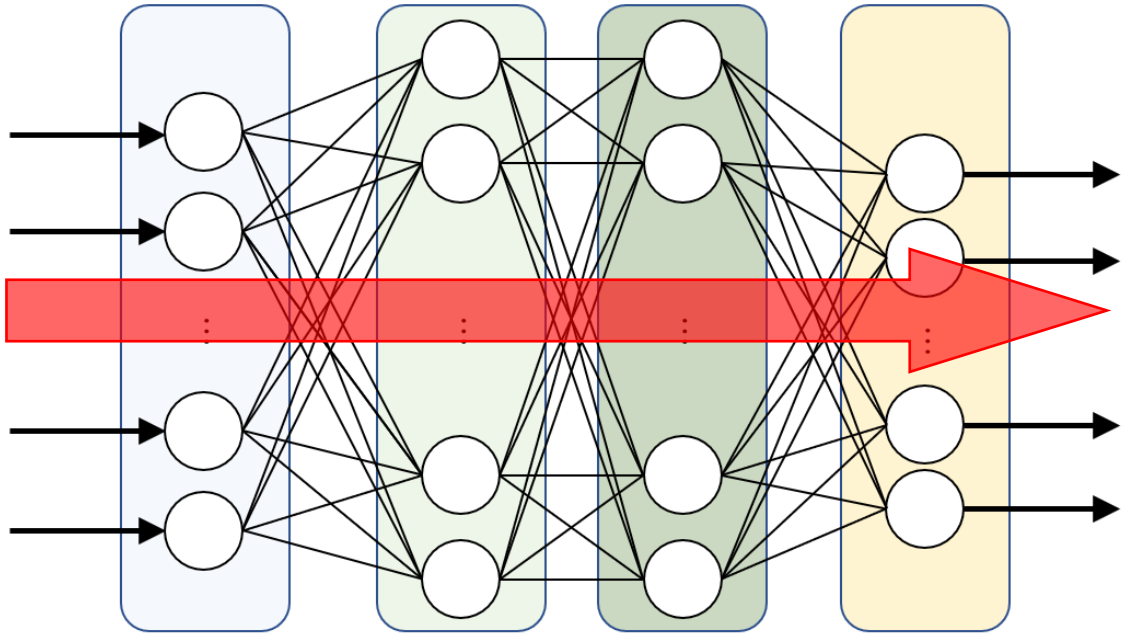
ニューラルネットワーク



信号が伝播されるときに、
乗算によりw_iの重みが付けられる

$$y = f(w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3)$$

ニューラルネットワークの構成



入力層　　中間層　　出力層

順伝播（フォワードプロパゲーション）で計算して分類する

データセットの例(fashion_mnistの読み込み)

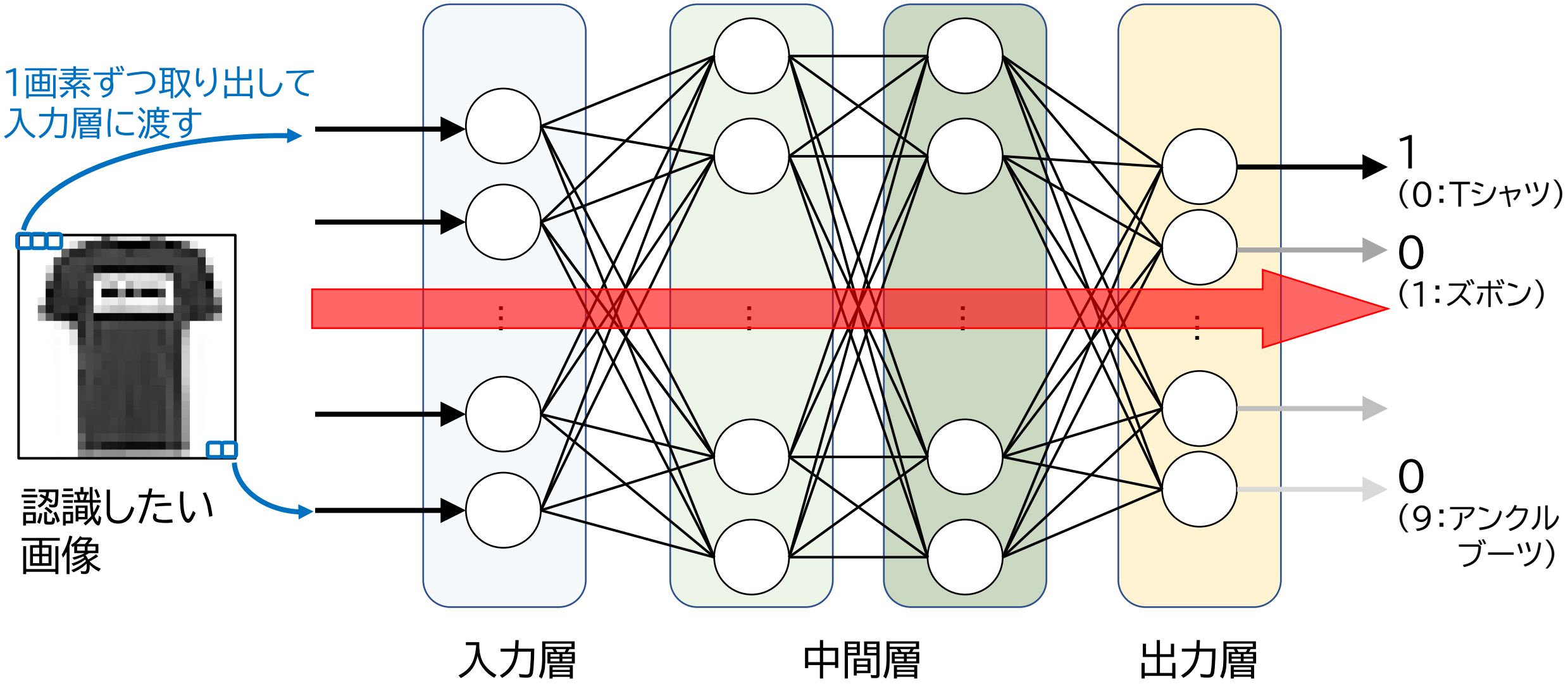
```
from keras.datasets import fashion_mnist
(X_train, y_train), (X_test, y_test) =
fashion_mnist.load_data()
```

fashion_mnistという
ファッションアイテムの画像データを
読み込む

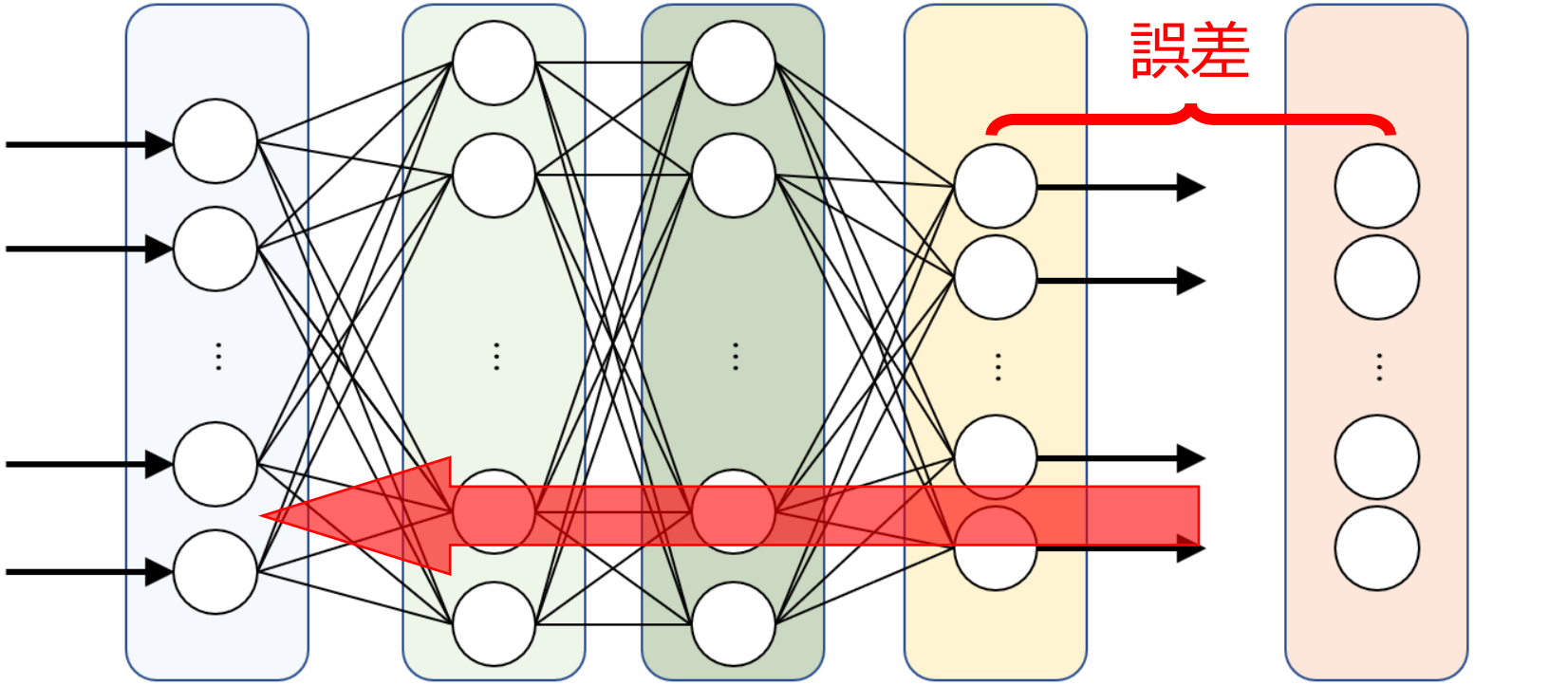
(詳細は講義動画を参考にしてください)

【訓練データ：60000枚】
【テストデータ：10000枚】

画像が認識される仕組み



ニューラルネットワークでの学習の仕組み



誤差逆伝播（バックプロパゲーション）により重みを調整

情報とデータサイエンス

機械学習・深層学習はここまでできる，進化・深化している

自動で顔にぼかしを入れよう！
（学習済みモデルを用いて実践してみよう）

人間の顔を認識することも可能に

顔って、
こんなものだ！

学習フェーズ

学習画像

特徴量抽出

学習

学習結果データ

検出フェーズ

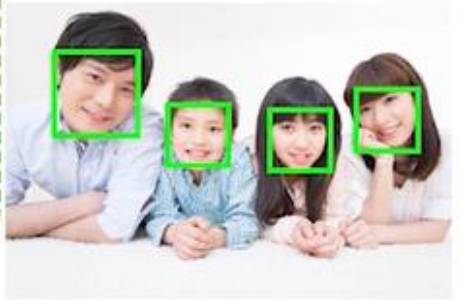
入力画像

特徴量抽出

検出

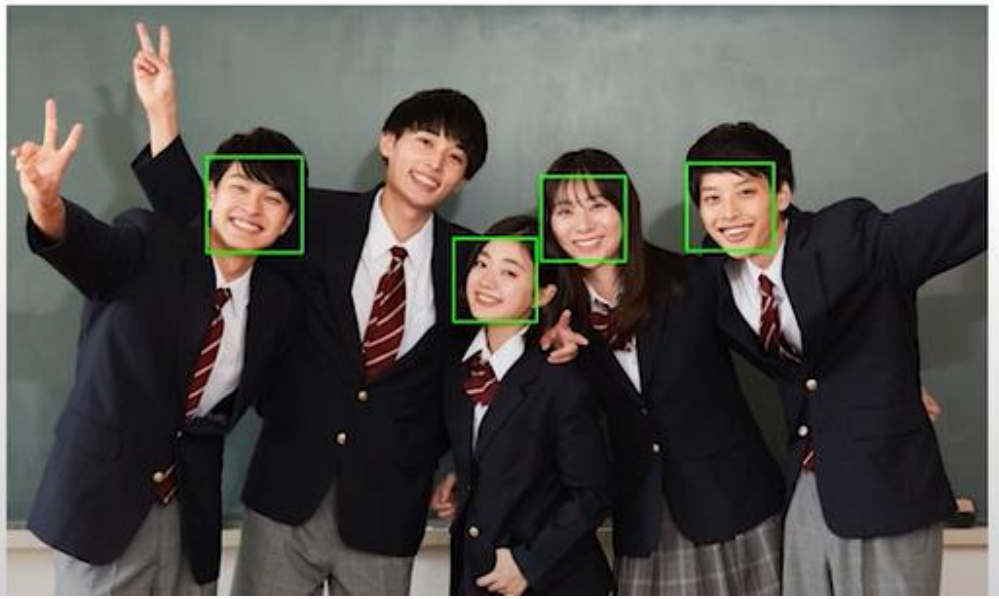
認識結果

顔っぽいのは、
ここだ！



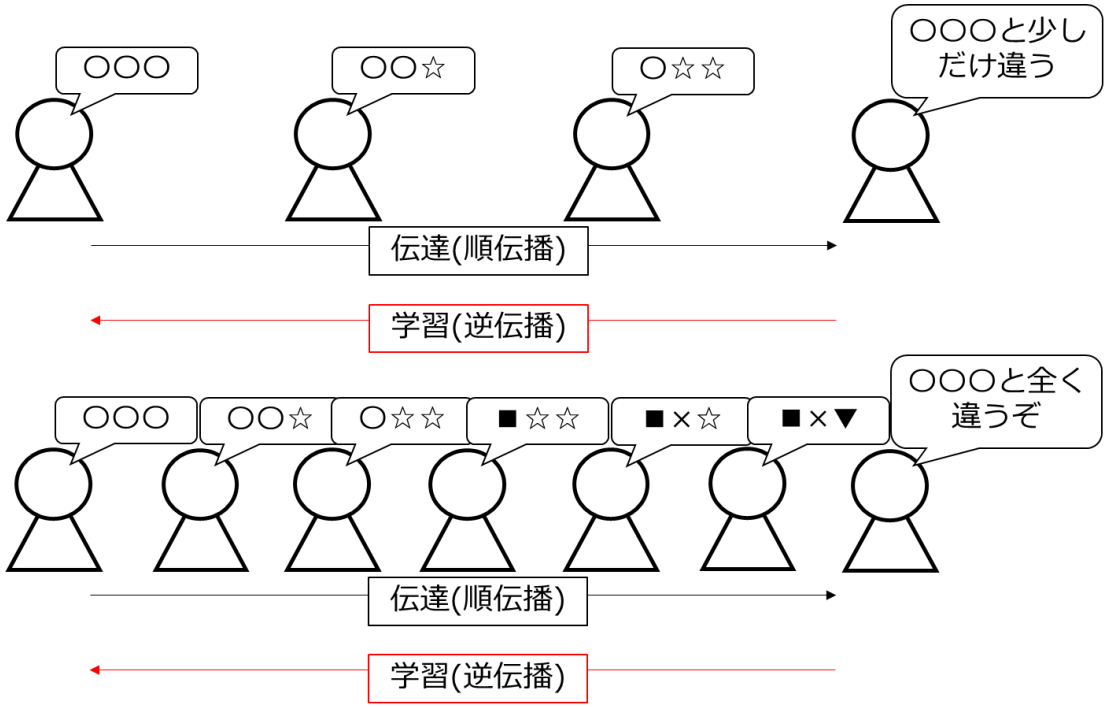
もちろん完璧ではない

- 検出モデルを変えてみる
  haarcascade_frontalface_alt.xml → haarcascade_eye.xml
- 精度を高めるためには？
  - 「誤検出」や「検出漏れ」もある →パラメータを調整してみる
  - カスケード型分類器以外を試してみる etc..
- 顔に間違われるものには何がある？
- 様々なシステムを作ってみよう！



(参考)ニューラルネットが発展したわけ

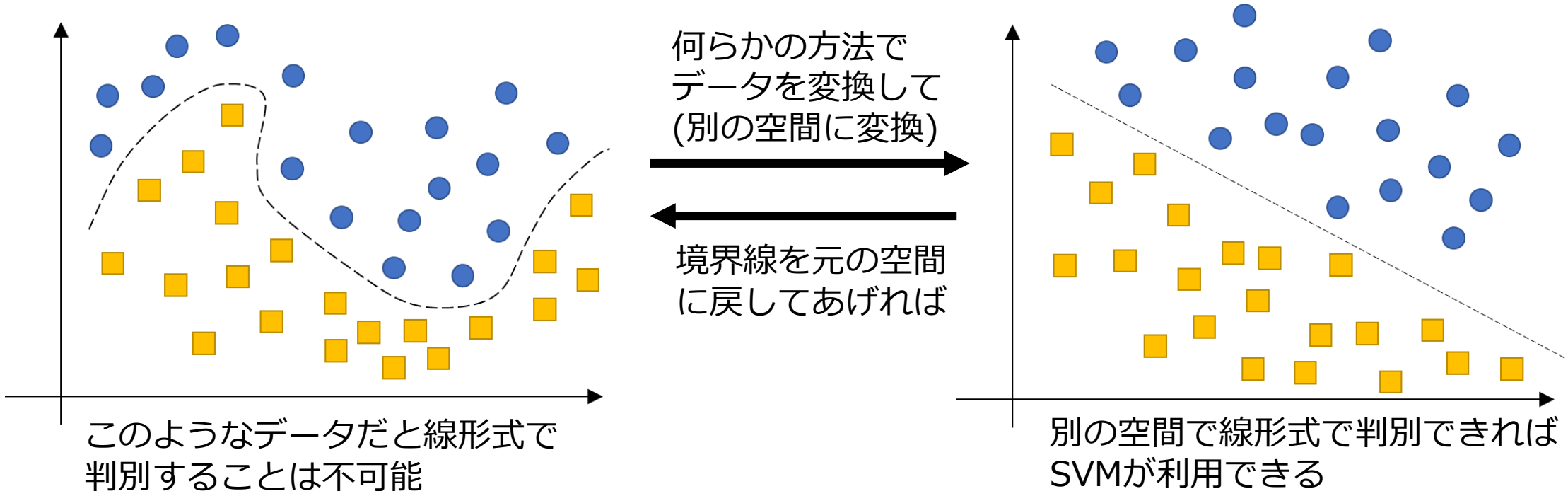
- 1986年に誤差逆伝播法が再認識され，ニューラルネットワークの学習がうまくできるはずだった…
- 誤差逆伝播法予測では正解との誤差のフィードバックが必要だが，中間層が多くなると，階層をさかのぼるごとに誤差が減少し、学習速度・学習精度が低下する現象が見られた(**勾配消失**)



- 2006年に発表されたオートエンコーダの出現により，勾配消失が解消
- 畳み込みニューラルネットワーク(CNN)やリカレントニューラルネットワーク(RNN)の開発により，従来の音声や画像解析手法を劇的に上回るパフォーマンスが達成される

(参考)SVMとその発展

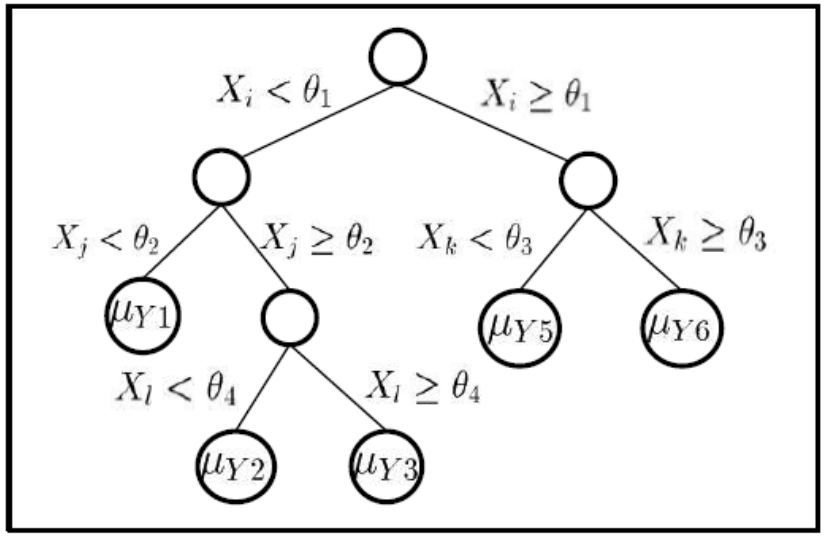
- SVM(特にLinear SVM)は線形式による判別



- 何らかの変換方法が存在する場合がある：カーネル法(カーネルトリック)
- 多少間違った判別をしてもよい：ソフトマージン

(参考)決定木の利用とその発展

- 決定木をベースとしたモデル



(メリット)
- モデルの構造が分かりやすい
- 非線形構造も(近似的に)表現可能

(デメリット)
- 過学習しやすい(学習データに合わせがち)
- 木が深くなりすぎる場合あり

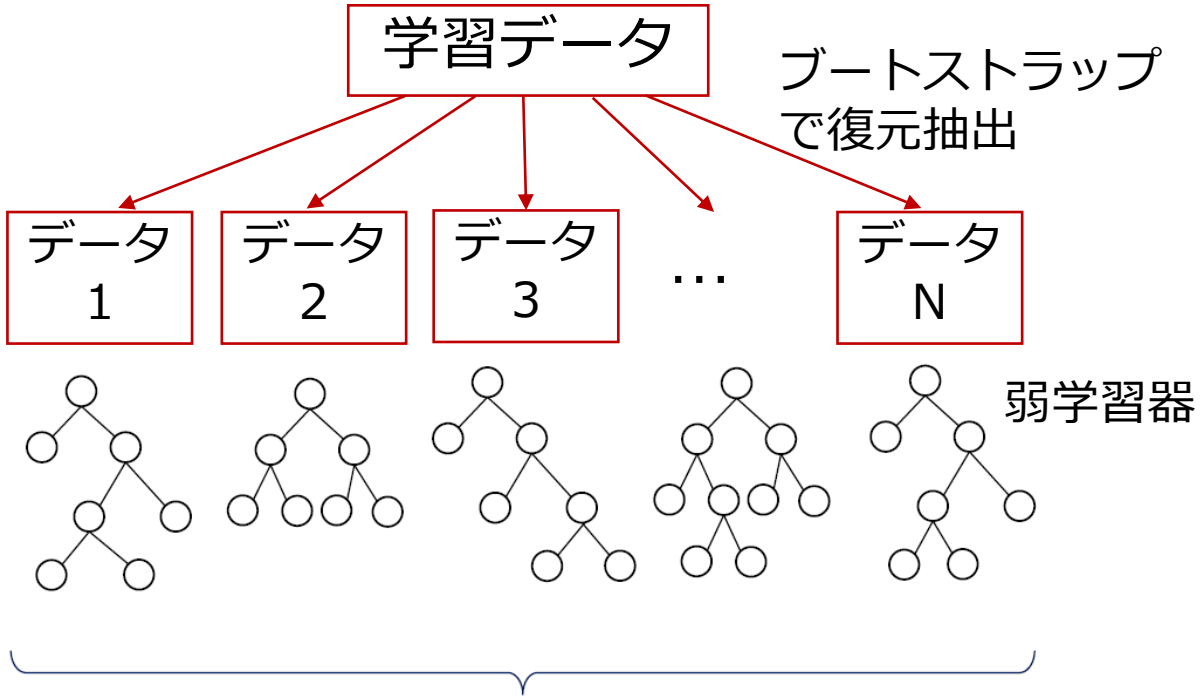
複数の決定木を用いた集団学習
(アンサンブル学習)

バギング(ブートストラップ)系
- **ランダムフォレスト**

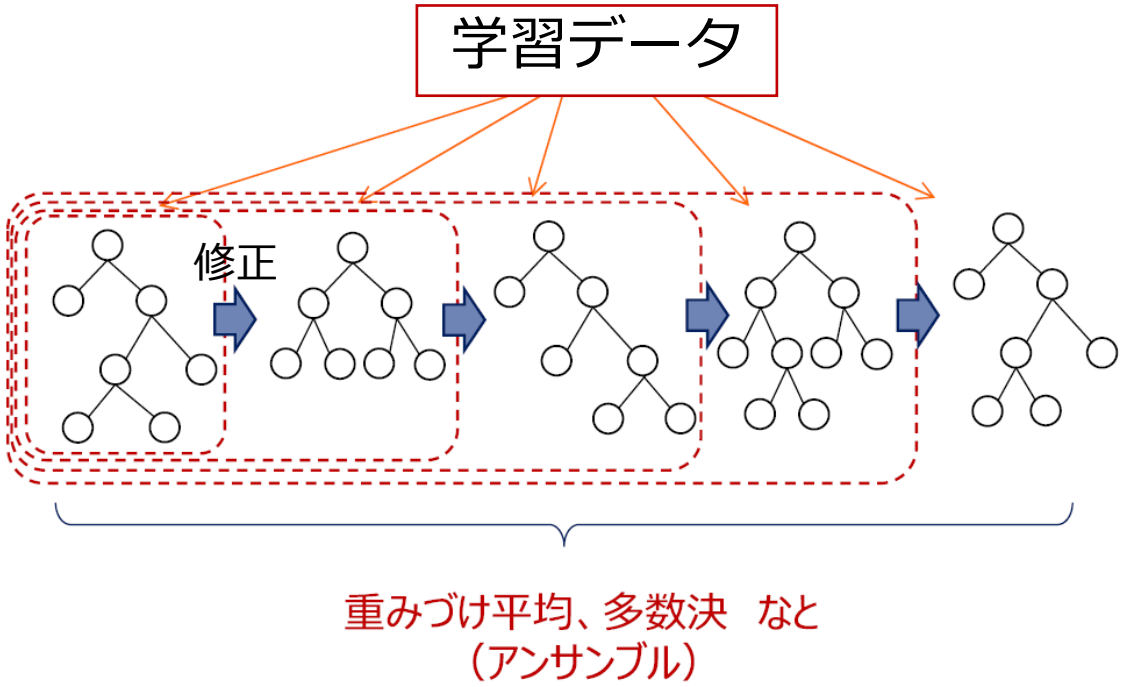
ブースティング系
- **勾配ブースティング**
  XGBoost，LightGBMなど

(参考)決定木の利用とその発展

ランダムフォレスト



勾配ブースティング



(統計的手法も含めた)2日間のまとめ

- データ分析を行うときに考えるべきこと
  - 何をしたいのか(目的は何か？仮説検定？予測？分類？要約(集約)？)
  - どのようなデータで分析するのか？
  - 分析でどのようなモデルが必要なのか(高い精度？高い解釈性？)

- 目的や収集できたデータの特徴(数値，文書など)や量，変数の数などから手法を選択する
- Web上で最先端手法の解説やプログラム等を掲載してくれているため，手法を「使う」面では，ある程度のプログラミング力があれば，今からでもチャレンジできる(はず)
- 「なぜその手法を使うのか」を理解できることも，今後データサイエンスに関わっていくとすれば重要

結果をしっかりと分析し，次の問いにつなげることが最も重要！