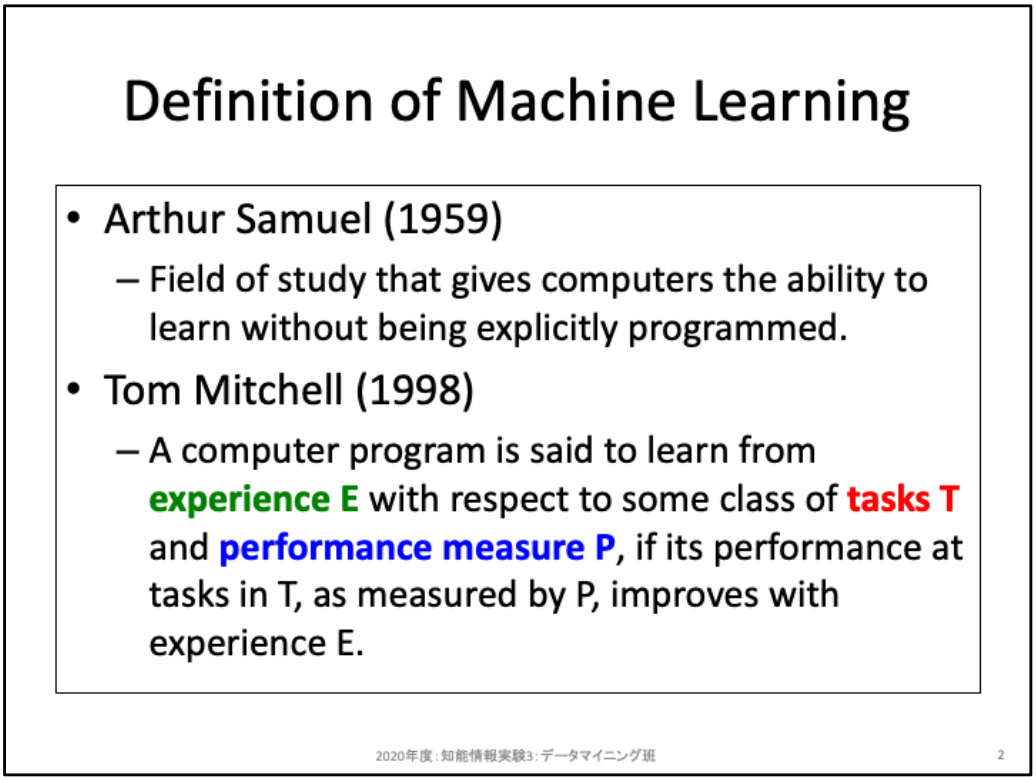
知能情報実験3：データマイニング班

(week 2) 機械学習概観

1. 機械学習の定義
2. 専門用語
3. 問題設定例（分類, 回帰, クラスタリング）
4. 検討課題
5. 問題設定サマリ
6. 機械学習の種別
7. クイックスタート（scikit-learn）

実験ページ: http://ie.u-ryukyu.ac.jp/~tnal/2020/info3/dm/



「機械学習」の定義は歴史を追う毎に変化している。

Arthur Samuel先生（1959）によると、明確なプログラムなしに学ぶ能力を有するコンピュータに関する学問分野のことを指していた。

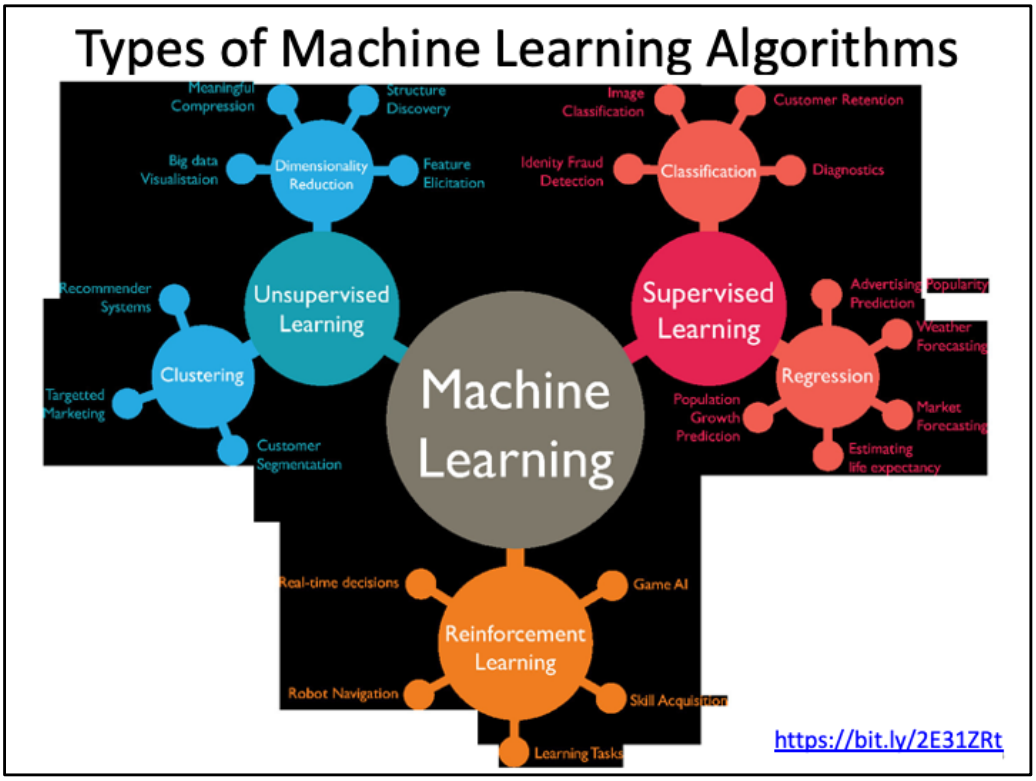
これに対しTom Mitchell先生（1998）では、ある程度の条件が追加された定義を提唱している。すなわち、パフォーマンスを計測するPを有するタスクTの元で経験Eを通し、そこからそのタスクTにおけるパフォーマンスPを改善するように学習することを機械学習と定義した。

Terminology

- ML types
  - supervised, unsupervised, semi-supervised
  - (reinforcement learning, genetic algorithm,,,)
- Task types
  - classification, regression, clustering
- sample
- features, attributes
  - numerical value
  - categorical value
  - true or false
- supervisory signal, teacher, class, label, target variable

- input, output
- Input types
  - training data / training set
  - test (for evaluation)
  - validation (for hyper params)
- model
- parameters
  - hyper parameters
  - weights, parameters
- learn, fit
- predict, estimate
- evaluation
  - open or close test
  - cross validation

ここに示している用語は、これから少しずつ定義を学んでいこう。



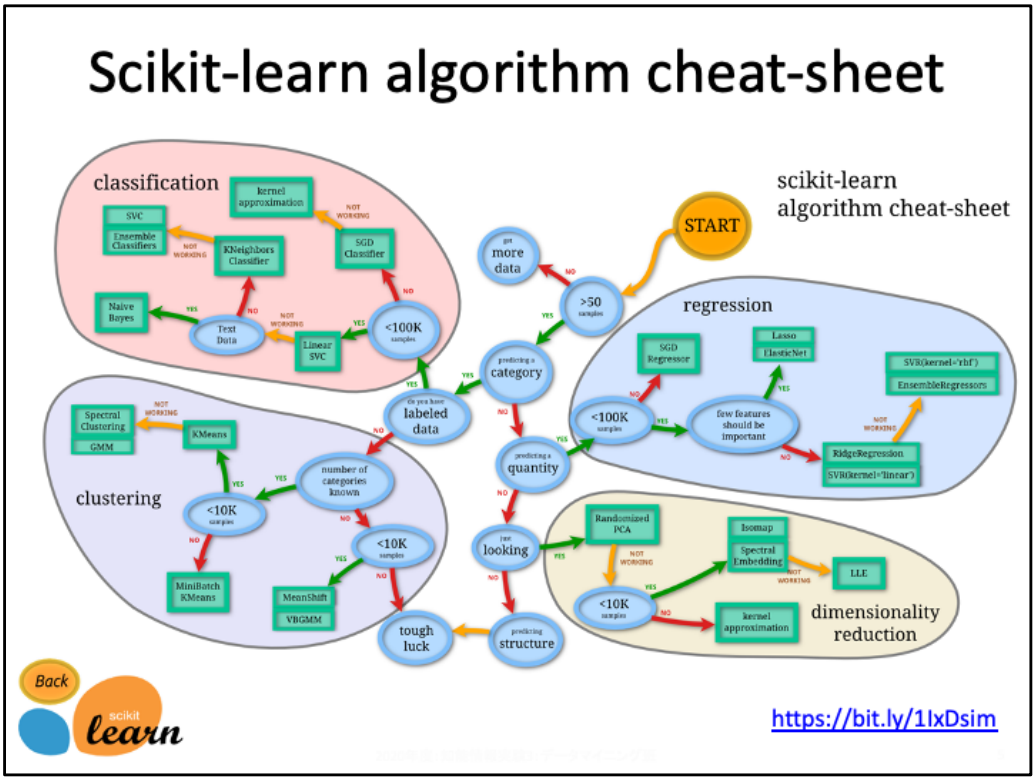
機械学習の体系図の例を示している。

大別すると教師あり学習（supervised learning）、教師なし学習（unsupervised learning）、強化学習（reinforcement learning）に分かれる。

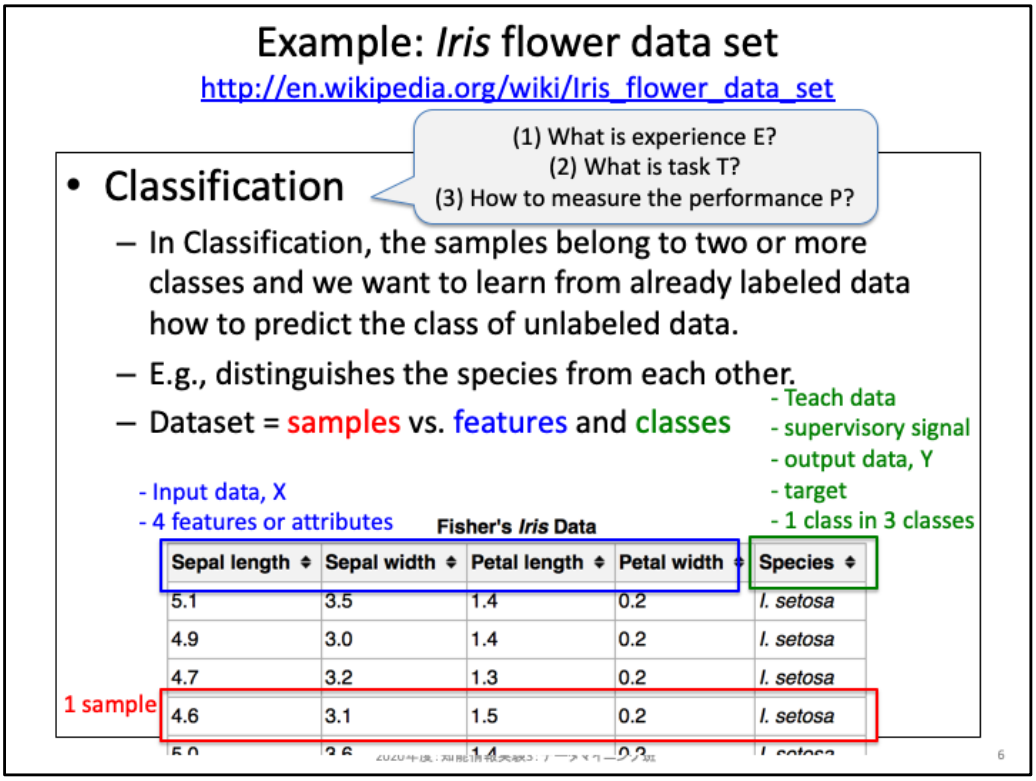
教師あり学習とは、例えば写真に写っているのが犬か猫かを判別（識別）したいときに事前に答えを犬・猫といったラベルとして用意しているような問題に適用する学習を指す。その次のClassificationは今示したような分類タスクのことを意味しており、更にその先の末端にあるものはより具体的な事例を紹介している。Regressionは回帰と呼ばれ、識別結果が犬・猫のようなカテゴリではなく、数値（連続値）を取るケースを回帰タスクと呼ぶ。

教師なし学習とは、分類タスクのように明瞭な答えを用意していない（できないケースも含む）場合に行うタスク群である。Dimensionality reductionは次元削減と呼ばれ、例えば100次元のデータは直接的に眺めることが困難である。これを2～3次元に圧縮したり、重要性の低い次元を削除することで次元を削減することを目的としたタスクである。Clusteringはクラスタリングと呼ばれ、とにかく似ているものをまとめたい、グルーピングしたい、というような目的で利用されることが多い。まとめたものを一つのカテゴリとして設定することで教師あり学習に利用するといったことも多く見られる。

強化学習とは、あるゴールに辿り着くための最良な方法を模索するために行うことが多い。例えば将棋で勝つための手順を獲得したり、サッカーでゴールに辿り着くための戦略を練ったり、といったことだ。教師あり・教師なし学習と明確に異なる点は、指標自体は明確に設定できるが、その指標上でどのぐらい良いのか悪いのかはシミュレーションを要する点にある。



機械学習には様々なモデルがあり、タスクや特徴ベクトル等様々な観点を考慮して選ぶことが望ましい。その背tなく基準の例がこれだ。必ずしも適切とは限らないが、一つの例として参考にしよう。



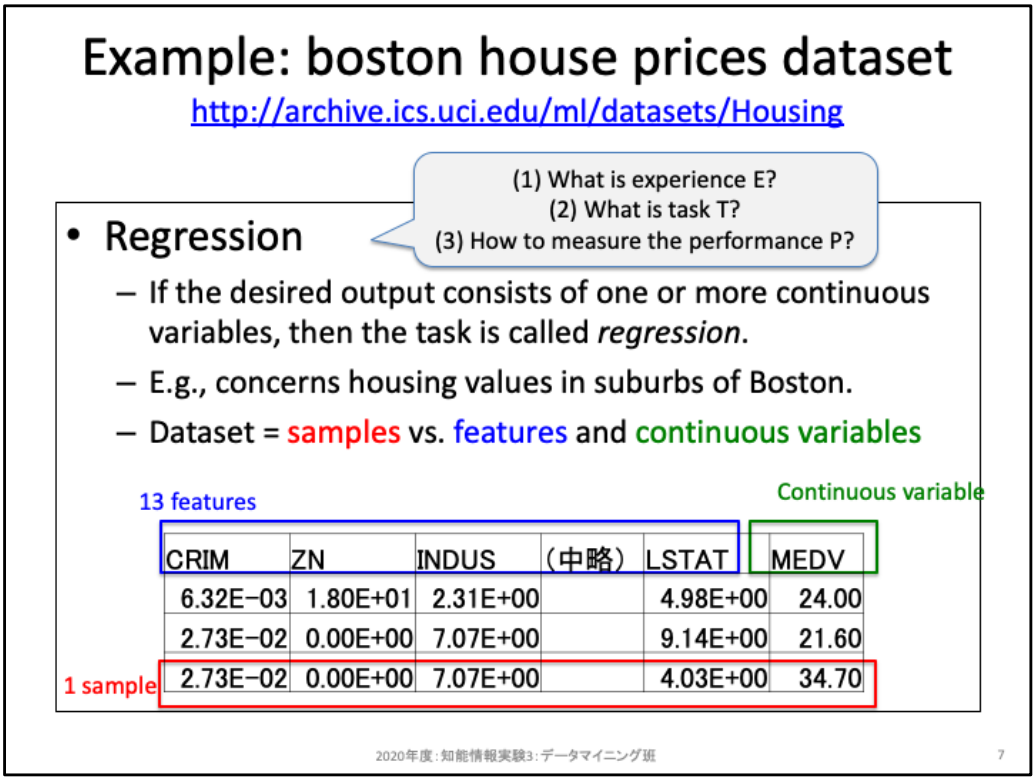
分類タスクについて定義と事例で確認してみよう。

分類タスクとは、各サンプルをクラスに分類することが目的である。クラスはラベルやカテゴリ、教師データなどとも呼ばれる。分類タスクの目的は、事前にラベルが付与されたデータセットからラベル付与に関する傾向を学習し、まだラベルが付与されていないデータセットに対してラベルを推測することである。

スライド下部の表は Iris flower dataset と呼ばれ、花の4箇所の長さ（sepal length～petal length）からその端の種別「I. setosa, I. Versicolour, I. Virginica」について収集されたデータセットである。長さはcm単位の連続値であり、種別はラベル（カテゴリ）である。

分類タスクにおいて、ラベルを学んだり推定したりする個々の事例を「サンプル」と呼ぶ。スライドにおいては「5.1, 3.5, 1.4, 0.2, I. setosa」という1行が1サンプルに相当する。各サンプルは特徴と教師データで構成されている。ここでは「Sepal length」や「Sepal width」が特徴であり、各サンプルは4つの特徴を有している。一般的には複数の特徴を有することから、これらの数字列を「特徴ベクトル」と呼ぶことが多い。最後の「Species」は、この分類タスクにおけるラベルであり、1つのサンプルが1つのラベルの場合を「シングルラベル」と呼び、複数のラベルを有する場合には「マルチラベル」と呼び分ける。

このように、分類タスクにおいては分類したい対象をサンプル集合として構成する。各サンプルは特徴ベクトルと教師信号（ラベル）で構成する。特徴ベクトルは原則として数値（実数）ベクトルである必要があり、数値でない場合には何らかの方法で変換することが一般的である。



回帰タスクの定義と例を眺めてみよう。

こちらも各サンプルが特徴ベクトルと教師信号で構成されている点は同様だ。ただし教師信号が連続値である点が異なる。

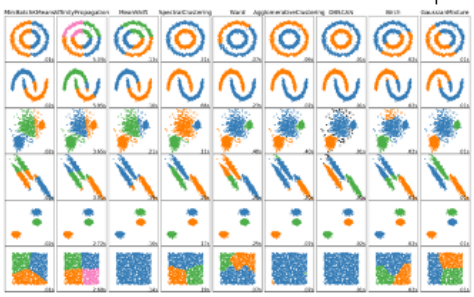
Boston house prices datasetは回帰タスクのデータセット例であり、予測したい値は連続値（価格）である。

Example: *Overview of clustering methods*

https://scikit-learn.org/stable/modules/clustering.html

(1) What is experience E?
(2) What is task T?
(3) How to measure the performance P?

- **Clustering**
  - Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).
  - Training data consists of a set of input vectors x **without any corresponding target values**.
  - Dataset = samples vs. features



2020年度：知能情報実験3：データマイニング班 8

クラスタリングの定義と例を眺めてみよう。

こちらは教師なし学習であり、各サンプルが特徴ベクトルを持つのみである。教師信号はない。右下の図はスライド上部のURL先から引用したものであり、各点がサンプルに相当する。各サンプルは2次元ベクトルであり、横方向に同一データセットを並べ、それを異なるクラスタリングアルゴリズムに適用した際の結果を色分けして出力している。この図からも分かるように、同じデータセットであったとしてもアルゴリズムが異なれば得られる結果も異なる（同じであることもある）。クラスタリングに限った話ではないが、学習アルゴリズムに適用する際には「サンプルをどのような特徴ベクトル（ベクトル空間）として表現するか」が重要である。

Exercises

- Make a group of 2~4 students.
  - Choose one kind of problem settings on machine learning.
  - Try to design an example under the problem setting.
    - Input? Features? Output?
    - What is experience E?
    - What is task T?
    - How to measure the performance P?

今回は分類タスクの事例を考えてみよう。

その際、各サンプルはどのように構成されるのか。モデルへの入出力はどのようになるだろうか。
機械学習の定義における、経験E、タスクT、パフォーマンスPはどのようになるだろうか。

Machine Leaning: the problem setting

http://scikit-learn.org/stable/tutorial/basic/tutorial.html

- In general, a learning problem considers a set of n samples of data and then tries to predict properties of unknown data. If each sample is more than a single number and, for instance, a multi-dimensional entry (aka multivariate data), is it said to have several attributes or features.

これまで眺めてきた定義や事例を整理するとこのようになる。（読んでみよう）

Types of Machine Learning

- Targets of this class
  - Supervised Learning
    - Classification
    - Regression
  - Unsupervised Learning
    - Clustering
  - (Semi-supervised Learning)

- Others
  - Principal component analysis
  - Reinforcement Learning
  - Artificial Neural Networks
  - Genetic Algorithm
  - Recommender System
  - Decision Trees
  - …

Quick Start

- http://scikit-learn.org/stable/tutorial/basic/tutorial.html
  - Google: scikit-learn
    - Documentation
    - Quick start

Scikit-learnのquick startをやってみよう。