情報工学実験4：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/2015/info4/dm/

Definition of Machine Learning

- Arthur Samuel (1959)
  - Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998)
  - A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E.

Terminology

- supervised, unsupervised learning
- classification, regression, clustering
- sample
- features, attributes
  - numerical value
  - categorical value
  - true or false
- supervisory signal, teacher, class, label, output data, target variable

- input, output
- training data / training set
- test data / test set
  - open test
  - close test
- model
- parameters
- learn, fit
- predict, estimate
- evaluation

Example: *Iris* flower data set

http://en.wikipedia.org/wiki/Iris_flower_data_set

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

- Classification
  - In Classification, the samples belong to two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data.
  - E.g., distinguishes the species from each other.
  - Dataset = samples vs. features and classes

- Teach data
- supervisory signal
- output data, Y
- target
- 1 class in 3 classes

- Input data, X
- 4 features or attributes

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*

1 sample

Example: boston house prices dataset

http://archive.ics.uci.edu/ml/datasets/Housing

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

- Regression
  - If the desired output consists of one or more continuous variables, then the task is called *regression*.
  - E.g., concerns housing values in suburbs of Boston.
  - Dataset = samples vs. features and continuous variables

13 features

Continuous variable

CRIM	ZN	INDUS	（中略）	LSTAT	MEDV
6.32E−03	1.80E+01	2.31E+00		4.98E+00	24.00
2.73E−02	0.00E+00	7.07E+00		9.14E+00	21.60
2.73E−02	0.00E+00	7.07E+00		4.03E+00	34.70

1 sample

Example: *Iris* flower data set WITHOUT classes

http://en.wikipedia.org/wiki/Iris_flower_data_set

(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

4 features

Don't use at learning

1 sample

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*

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?

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.

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