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 $E$.  

2015年度:情報工学実験4:データマイニング班
## 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


- **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

- Input data, X
- 4 features or attributes
- Teach data
- supervisory signal
- output data, Y
- target
- 1 class in 3 classes

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

---

**Fisher’s Iris Data**

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
<tr>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
<tr>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
</tbody>
</table>
Example: boston house prices dataset

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

• 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

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

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.32E-03</td>
<td>1.80E+01</td>
<td>2.31E+00</td>
<td>4.98E+00</td>
<td>24.00</td>
</tr>
<tr>
<td>2.73E-02</td>
<td>0.00E+00</td>
<td>7.07E+00</td>
<td>9.14E+00</td>
<td>21.60</td>
</tr>
<tr>
<td>2.73E-02</td>
<td>0.00E+00</td>
<td>7.07E+00</td>
<td>4.03E+00</td>
<td>34.70</td>
</tr>
</tbody>
</table>
Example: *Iris* flower data set **WITHOUT classes**


- **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**

![Fisher's Iris Data Table]

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

Don’t use at learning
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 Learning: the problem setting


• 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][1]
  - Google: scikit-learn
    - Documentation
    - Quick start