An Introduction to Pattern Recognition

Topics of this lecture

- Concept and concept learning.
- Pattern classification and recognition.
- Feature vector representation of patterns.
- Nearest neighbor based learning.
- Discriminant function and decision boundary.
- Multi-class pattern recognition.
- General formulation of machine learning.
- The k-means algorithm.

Concept learning

- There are two types of knowledge: declarative (宣言的）and procedural（手続き的）knowledge.
- Declarative knowledge can be represented by concepts and relations between concepts (say, using a graphic model like semantic network).
- Procedural knowledge is basically a “transform” from one group of concepts to another group of concepts (e.g. a function in C-language).
- Learning various concepts based on observations or experiences is the first step to build an AI system.

Definition of a concept

- Concept is a sub-set of the universe of discourse.
- X: Universe of discourse
- A: concept defined on X

$$A = \{x | \mu_A(x) = True \bigwedge x \in X\}$$

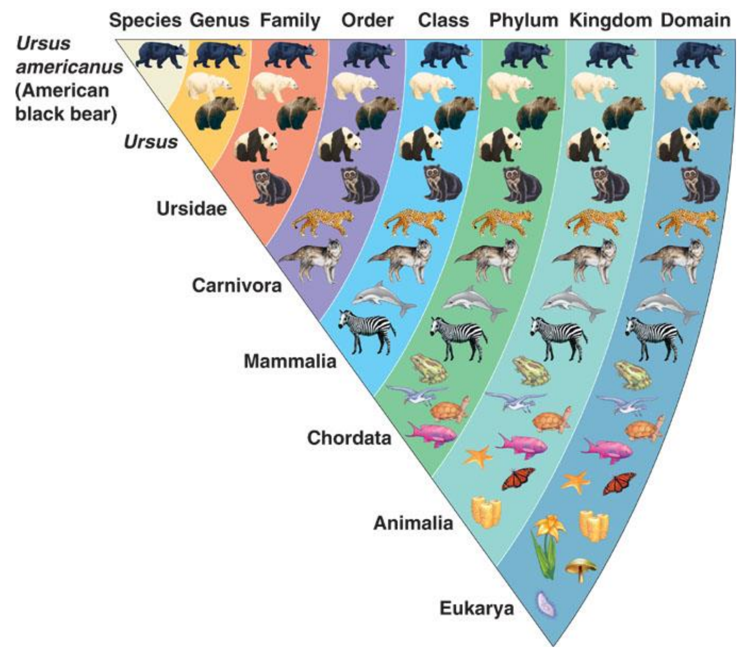
- In general $\mu_A(x)$ is a logic formula representing the ***membership function*** of A.
- For a fuzzy concept, the range of $\mu_A(x)$ is [0,1].

Pattern classification / recognition

- Pattern classification is the process for partitioning a given domain into various meaningful concepts.
- Pattern recognition is the process to determine to which concept an observed datum belongs.
- Example:
  - Domain: Chinese characters (Kanji)
  - Concepts: Nouns, verbs, adjectives, …
  - Given observation: 城 → noun; 走 → verb
- A concept is also called a class, a category, a group, a cluster, etc.

Why science is translated to "科学" (Kagaku or Kexue)?

- 「**科学**」is an interesting and "scientific" translation of the word "science".
- It means "study on classification or categorization" (分類、分科の学問).
- Based on the classification results, we can understand the world in a more organized way.



https://liorpachter.wordpress.com/2015/10/27/straining-metagenomics/

Vector representation of patterns

- To classify or recognize objects in a computer, it is necessary to represent them numerically.
- We usually transform an object into an n-dimensional vector, which is a point in the $n-D$ Euclidean space, as follows:

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n)^t$$

- Each element of the vector is called a **feature**, and the vector itself is called a **feature vector**. The set of all feature vectors is called the **feature space**.

Some terminologies for learning

- Learning of a concept is the process for determining the membership function of the concept.
- Training set is a set of data used for learning. Each datum is called a training datum or training pattern.
- Usually, each training pattern x has a label, which tells the name of the concept x belongs to. The label is also called ***teacher signal***.

Some terminologies for learning

- In many applications, we consider two-class problems.
  - Face or non-face;
  - Human or non-human;
  - Normal or abnormal.
- For two-class problems, the label takes only two values {-1, 1} (or {false, true}, or {0,1}).
- A pattern is often called ***positive*** (or ***negative***) if its label is 1 (or -1).

Some terminologies for learning

- For any pattern, we can define its neighborhood using the Euclidian distance defined by

$$d(\boldsymbol{x}, \boldsymbol{q}) = \|\boldsymbol{x} - \boldsymbol{q}\| = \sqrt{\sum_{j=1}^{n} (x_j - q_j)^2}$$

- Patten $\boldsymbol{q}$ is said to be close to $\boldsymbol{x}$ if the distance is small.
- For any given pattern $\boldsymbol{x}$, its ε-neighbor, denoted by $N_\varepsilon(\boldsymbol{x})$, is a set of patterns in which any $\boldsymbol{p} \in N_{\boldsymbol{\varepsilon}}(\boldsymbol{x})$ satisfies the condition $d(\boldsymbol{x}, \boldsymbol{p}) \leq \varepsilon$.

Learning based on the neighborhood

- The simplest method for pattern classification is NNC, short for **nearest neighbor classifier**.
- To design an NNC, we just collect a set Ω of labeled training data, and use Ω directly for recognition.
- For any given pattern $\boldsymbol{x}$, $Label(\boldsymbol{x}) = Label(\boldsymbol{p})$ if

$$\boldsymbol{p} = arg \min_{\boldsymbol{q} \in \Omega} ||\boldsymbol{x} - \boldsymbol{q}||$$

- In this case, the whole training set Ω is an NNC.
- In general, NNC is defined by a set P of prototypes that can be a sub-set of Ω, or a set of templates found from Ω.

Learning based on the neighborhood

- Using NNC, we can define the membership function of a concept A as follows:

$$\mu_A(\mathbf{x}) = [\exists \mathbf{p} \in P^+][\forall \mathbf{q} \in P^-]\, \|\mathbf{x} - \mathbf{p}\| \leq \|\mathbf{x} - \mathbf{q}\|$$

- Where P^+ and P^- are the set of positive prototypes and set of negative prototypes, respectively.
- Physical meaning: For any given pattern x, if there is a positive prototype p, and the distance between x and p is smaller than that between x and any of the negative prototype, x belongs to A.

Properties of the NNC

- If the set P of prototypes contains enough number of observations, the error of the NNC is smaller than 2E, where E is the error of the “optimal” classifier (i.e. maximum posterior probability classifier).
- However, if the size of P is too big, it is very time consuming to make a decision for any given pattern x.
- In other word, NNC is easy to obtain, but difficult to use.

A method for reducing the cost

- One method for reducing the computational cost is to use a representative for each class.
- For 2-class problem, representatives can be given by

$$\mathbf{r}^{+} = \frac{1}{|\Omega^{+}|}\sum_{\mathbf{p}\in\Omega^{+}}\mathbf{p},\quad \mathbf{r}^{-} = \frac{1}{|\Omega^{-}|}\sum_{\mathbf{q}\in\Omega^{-}}\mathbf{q},$$

  where Ω^+ and Ω^- are, respectively, the set of positive training data and set of negative training data.
- Use the representatives, recognition is conducted by

$$\text{Label}(\mathbf{x}) = \begin{cases} +1 & \text{if } \left\|\mathbf{x}-\mathbf{r}^{+}\right\| < \left\|\mathbf{x}-\mathbf{r}^{-}\right\| \\ -1 & \text{if } \left\|\mathbf{x}-\mathbf{r}^{-}\right\| < \left\|\mathbf{x}-\mathbf{r}^{+}\right\| \end{cases}$$

From NNC to discriminant functions

- If the distance is defined as the Euclidean distance, pattern recognition can also be conducted as follows:

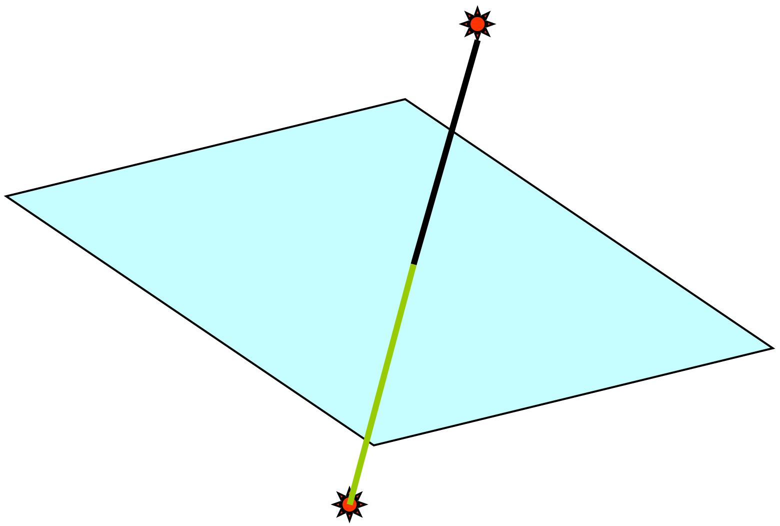
$$\mathrm{Label}(\mathbf{x}) = \begin{cases} +1 & \text{if } g^{+}(\mathbf{x}) > g^{-}(\mathbf{x}) \\ -1 & \text{if } g^{+}(\mathbf{x}) < g^{-}(\mathbf{x}) \end{cases}$$

- Here, $g^+(x)$ and $g^-(x)$ are called discriminant functions defined by

$$g^{+}(\mathbf{x}) = \sum_{j=1}^{n} x_j r_j^{+} - \frac{1}{2}\sum_{j=1}^{n} (r_j^{+})^2, \quad g^{-}(\mathbf{x}) = \sum_{j=1}^{n} x_j r_j^{-} - \frac{1}{2}\sum_{j=1}^{n} (r_j^{-})^2$$

- Since both functions are linear, they are also called **linear discriminant functions**.

Decision boundary



- For a 2-class problem, we need only one discriminant function defined by

$$g(\mathbf{x}) = g^{+}(\mathbf{x}) - g^{-}(\mathbf{x}) = \sum_{j=1}^{n} w_j x_j - \theta$$

$$H: \sum_{i=1}^{n} w_i x_i - \theta = 0$$

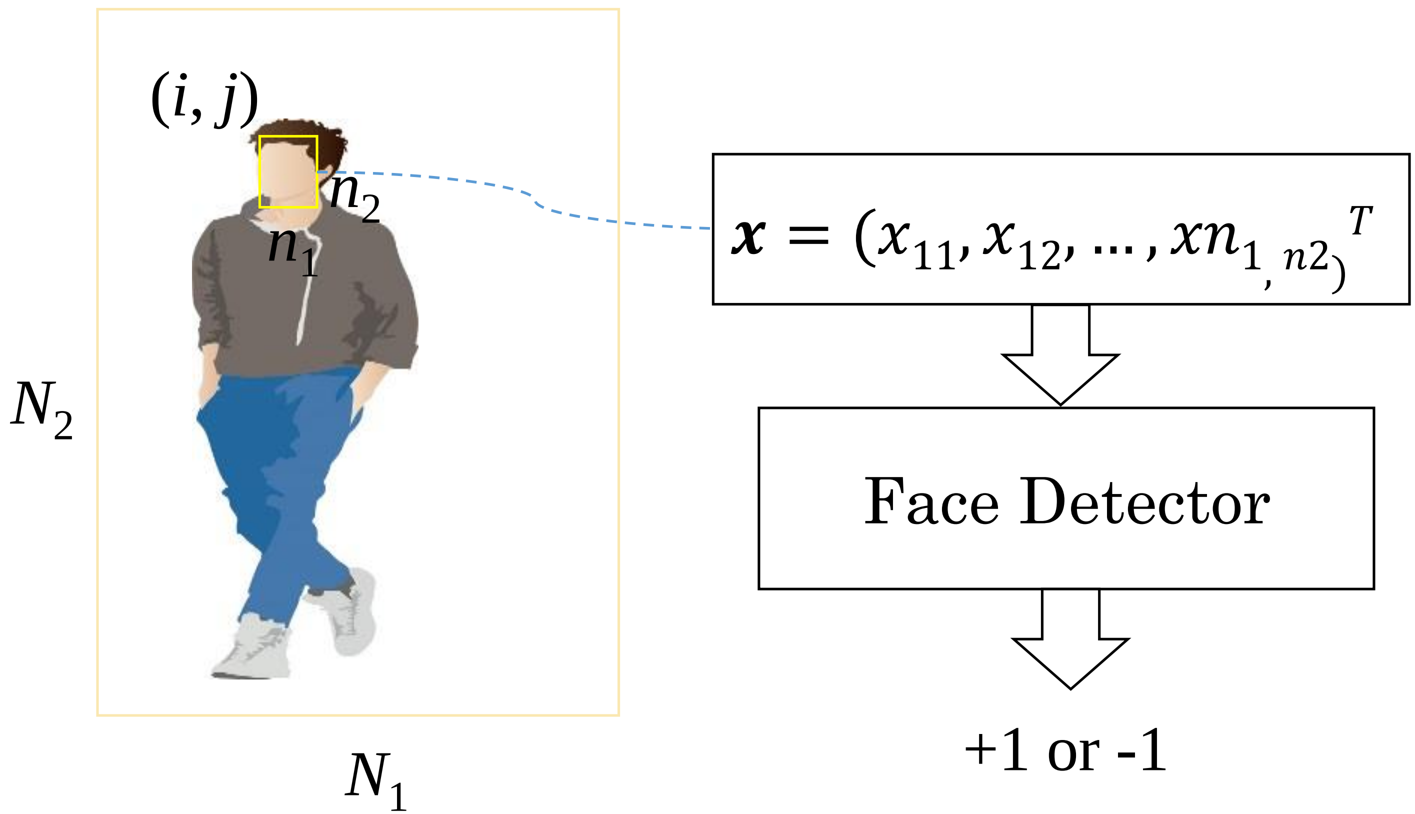
$$w_i = r_i^{+} - r_i^{-};$$

$$\theta = \frac{1}{2} \sum_{i=1}^{n} [(r_i^{+})^2 - (r_i^{-})^2]$$

- This function is actually a hyper-plane. Patterns on this plan cannot be classified.
- Thus, this hyper-plane forms the **decision boundary**.

Example 6.1 pp. 115-116

Illustration of face detection



Multi-class classification

- To solve a multi-class problem, we can use Eq. (6.2) and Eq. (6.3) given in the textbook to realize an NNC.
- We can also use the following rule:

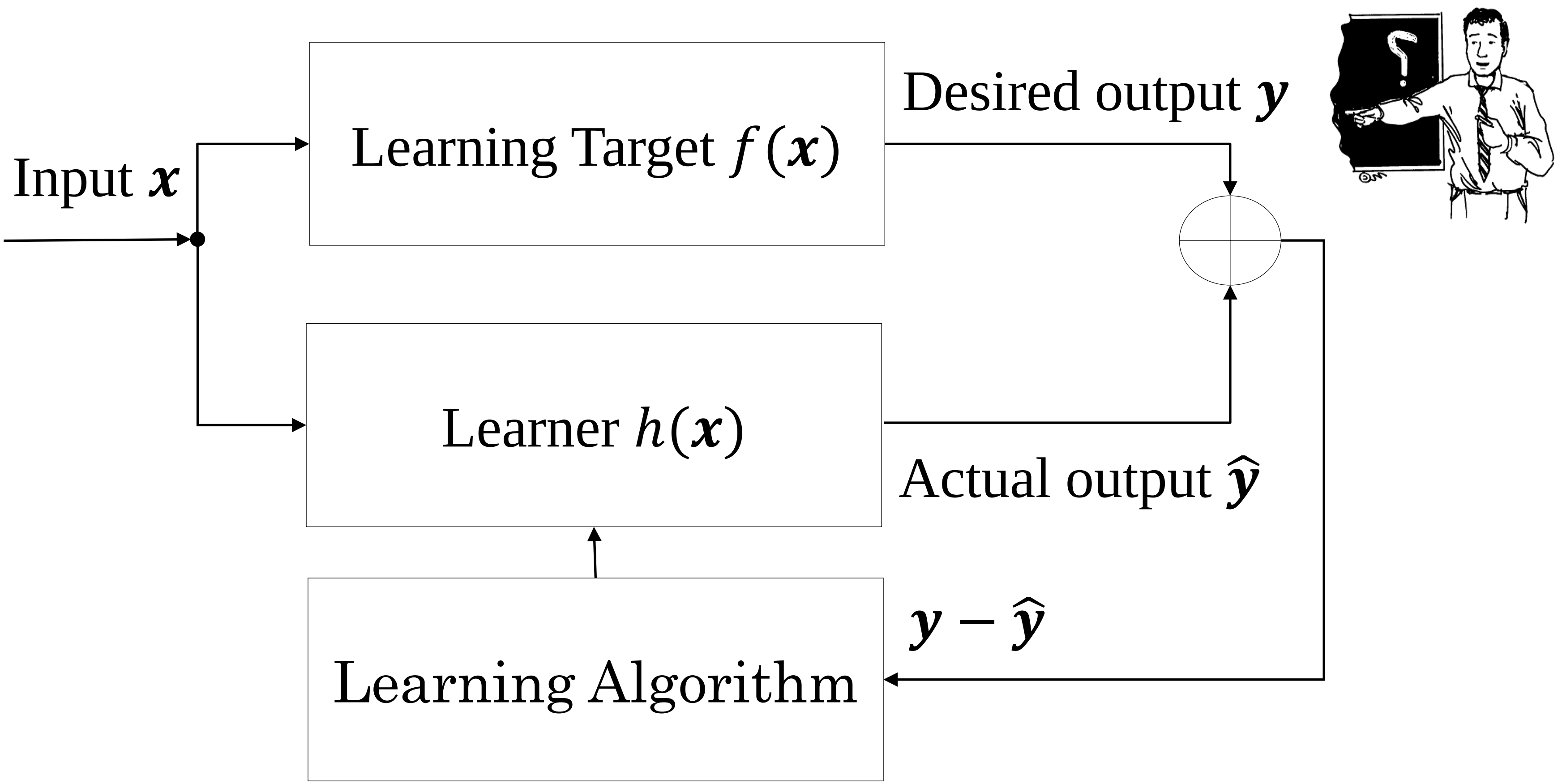
Given $\mathbf{x}$, Label($\mathbf{x}$)=k if
k=arg max $g_i(\mathbf{x})$, for i=1,2,…,N_c

- Here, g_i(x) is the discriminant function of the i-th class defined by

$$g_i(\mathbf{x}) = \sum_{j=1}^{n} x_j r_j^i - \frac{1}{2}\sum_{j=1}^{n}(r_j^i)^2, \quad i = 1,2,\ldots, N_c$$

- And r^i is the representative of the i-th class.

Formulation of machine learning



Formulation of machine learning

- Concepts to learn: $X_1, X_2, \ldots, X_{Nc}$

$$X_i = \{\mathbf{x} \in X \mid f(\mathbf{x}) = \mathbf{y}_i, \mathbf{y}_i \in Y\}$$

- $Y=\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{Nc}\}$ is the label set.
- A training datum is usually given as a pair (x,y), where x is the observation and y is the label given by a "teacher".
- **Supervised learning**: If y is available for all training data.
- **Un-supervised learning**: If y is not available.

Learning is the process to find a "learner" or learning model h(x) to approximate the target function f(x).

Formulation of machine learning

- The learner h(x) is usually determined by a set of parameters $w=\{w_1,w_2,\ldots,w_m\}$. That is, h(x) can be represented by h(x,w).
- In this case, finding the best h(x,w) is to find the best w. This kind of learning is called **parametric learning**.
- For a given w, h(x,w) is a **hypothesis**. The set H of all possible h(x,w) is called the **hypothesis space**.
- Parametric learning is an **optimization problem** for finding the best hypothesis from the hypothesis space H.

$$L = \sum_{\forall \mathbf{x} \in \Omega} \left\| f(\mathbf{x}) - h(\mathbf{x},\mathbf{w}) \right\|^2 + \lambda \frac{1}{p} \left\| \mathbf{w} \right\|_p^p = \sum_{\forall \mathbf{x} \in \Omega} \left\| f(\mathbf{x}) - h(\mathbf{x},\mathbf{w}) \right\|^2 + \lambda \frac{1}{p} \sum_i \left| w_i \right|^p$$

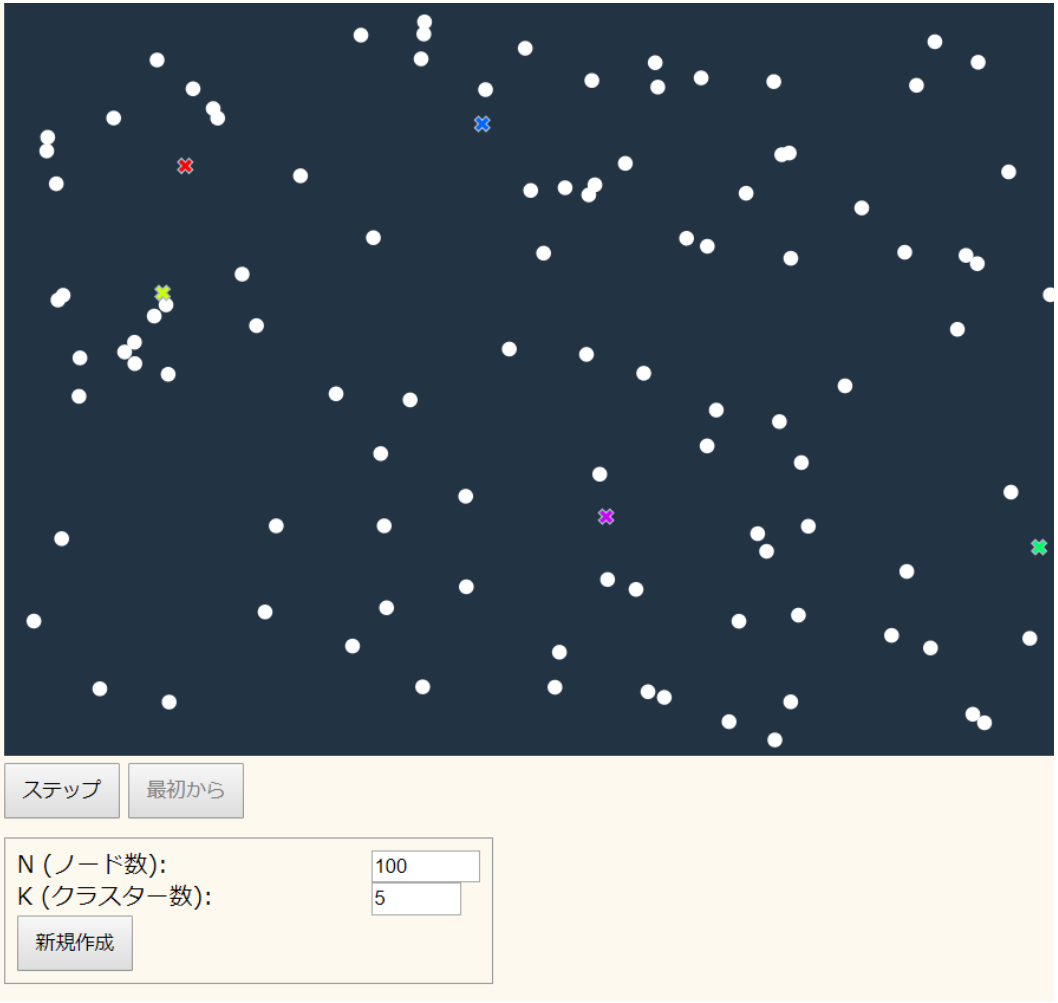
- Here, L is called the lost function, and the second term is the **regularization factor**.

K-means: An un-supervised algorithm for finding the representatives

Consider the problem to classify the domain D into K clusters based on un-labeled data.

- Step 1: Define a representative for each cluster at random (or select a representation from each class at random).
- Step 2: For each training data x, find the nearest representative. If the nearest representative is $\mathbf{r}_i$, label(x)=i.
- Step 3: For each cluster, re-define the representative by using the average of all data assigned to this cluster.
- Step 4: If the new representatives are ***almost the same*** as the old ones, Stop; otherwise, return to Step 2.

Demo of the k-means algorithm



How to use the results of K-means?

- Since K-means is an un-supervised learning algorithm, the results are representatives of K clusters.
- For any given new pattern x, we can “recognize” x by finding the nearest representative, and then assign the “index” of this representative to x.
- If we can obtain the class labels of the representatives later, the result of recognition can be the class label, rather than the index of the cluster.
- The class label of a cluster can be determined via majority voting. That is, if most data contained in a cluster belong to the i-th class, the cluster label is i.

Homework of lecture 10 (1)
(submit to the TA during the exercise class)

- Read Example 6.1 in pp. 115-116 carefully, and try to solve Problem 6.2.
- Purpose of this homework
  - Understand the meaning of two-class problem.
  - Understand the meaning of NNC.
  - Understand the basic process for face expression recognition.

Homework of lecture 10 (2)

- Complete the program for implementing the k-means algorithm.
- Test your program using the “Iris dataset”.
  - There are three types of “Iris” flowers (あやめの花) in the dataset.
  - http://archive.ics.uci.edu/ml/datasets/Iris
- Test your program using 3-fold cross-validation (3分割交差検証).

1	2	3

  - That is, divide the whole dataset into 3 parts as above, and use each part for testing, and the other data for training. This way, we get 3 results. The “average” of the results is used for evaluation.

Quizzes of today

1. What is the purpose of pattern classification?
2. What is the purpose of pattern recognition?
3. For a two-class problem, we usually call the two classes positive and ________________________________.
4. In an NNC, recognition is conducted by finding the_________________of the given pattern x.
5. To reduce the computational cost of an NNC, we can use __________ of each class.
6. If the desired class label is available for each training datum, learning is called ____________ learning.
7. If the learner or learning model is determined by a set of parameters, learning is called__________learning.