Distance-based neural networks

Topics of today

- Pattern recognition again
- Feature extraction/selection
- Self-organizing neural network
- Winner-take-all learning strategy
- Learning vector quantization
- R4-rule

Pattern recognition again

- Pattern classification is the process for partitioning a domain into various meaningful concepts; and pattern recognition is the process for determining the label of an observed pattern.
- 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.

Examples of patterns recognition

- Characters (kanji, kana, digits, etc.)
- Biometrics (finger prints, vein, iris, etc.)
- Signals (speech, sound, speaker, etc.)
- Images (face, expression, human, etc.)
- Videos (gaits, activity, gesture, etc.)
- Objects (human, car, obstacle, etc.)
- Anomaly (product, text, sequence, etc.)
- Time series (risk, chance, etc.)

Vector representation of patterns

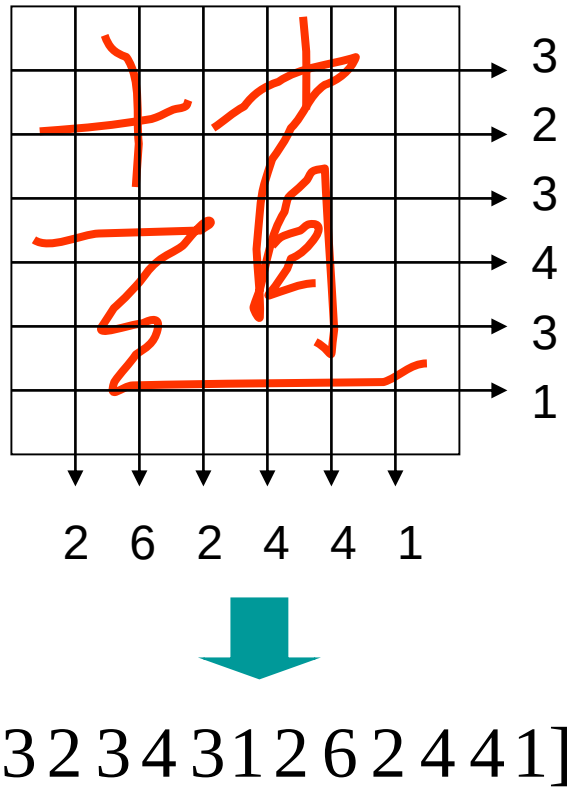
- To classify or recognize objects in a computer, it is necessary to represent them numerically.
- We usually transform an object into an n-dimensional vector, which is a point in the n-D Euclidean space, as follows:

$$\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^t$$

- 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**.

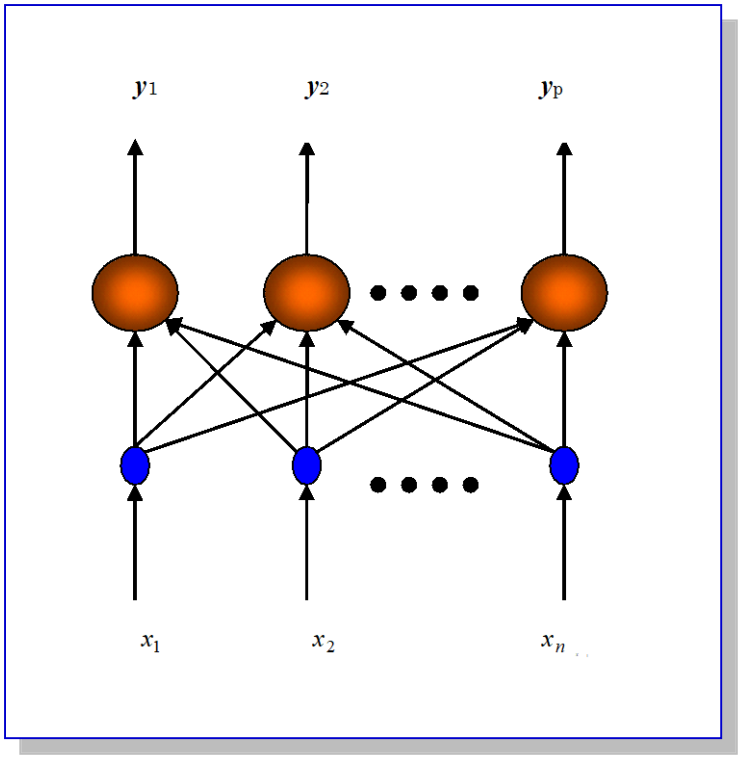
Feature extraction and selection

- The right figure shows an example of handwritten character recognition.
- Instead of using the whole picture, we can extract some features for recognition.
- The method given here is just an illustration.
- Usually, the number of extracted features can be large, and we should select some of the most useful ones.
- Feature extraction / selection is the first step for pattern recognition.

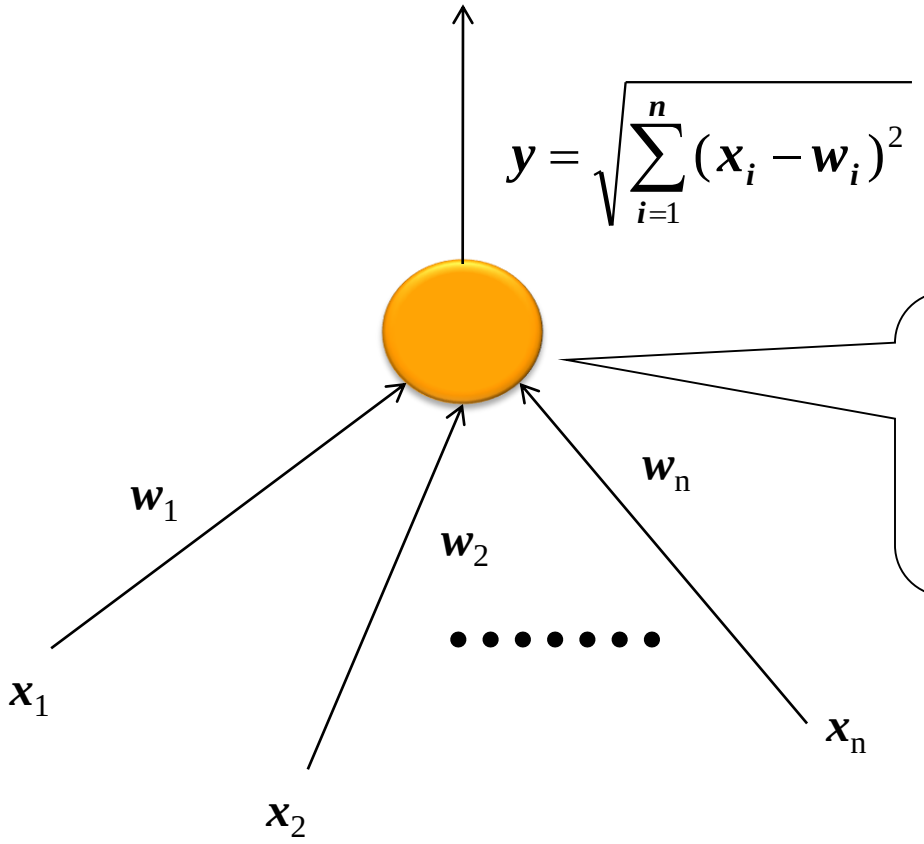


Distance-based learning

- The right figure shows a neural network implementation of the nearest neighbor classifier (NNC).
- There are $\boldsymbol{p}$ neurons, and each of them is a prototype.
- This network is called Kohonen neural network (KNN), and is useful both for self-organization (WTA) and for supervised learning (LVQ).



A neuron in a KNN



The output of the neuron provides a measure of similarity between the input pattern x and the memorized one w.

KNN is smarter

- Using the KNN, it is not necessary to remember all training patterns.
- Instead, a small number of representatives can be obtained through learning.
- The representatives are not the training data, but abstracted patterns (e.g. smile face, angry face, etc.).
- There are several ways for obtaining the neurons.
- Here, we first introduce a method called winner-take-all (WTA) that is useful for self-organization.

Winner-take-all Learning
勝者独占学習

Step 1: Initialize the **weights** of all neurons at random.

Step 2: For any $\mathbf{x}$ in the training set, find the most similar neuron, let it be $\mathbf{w}_m$. This neuron is called the winner.

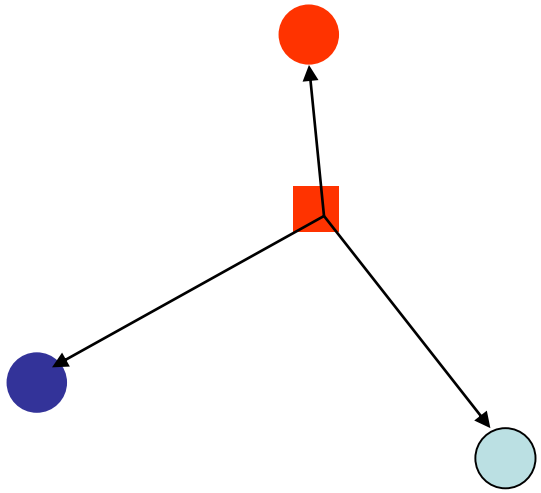
Step 3: Update the weights of the winner as follows:

$$\mathbf{w}_m^{k+1} = \mathbf{w}_m^k + \alpha(\mathbf{x} - \mathbf{w}_m^k)$$

- Where α is a positive number in (0,1). Usually, α is called the learning rate.

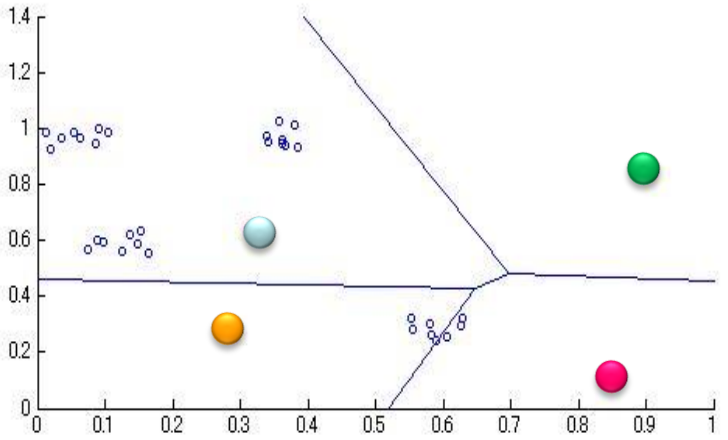
Physical meaning of the WTA

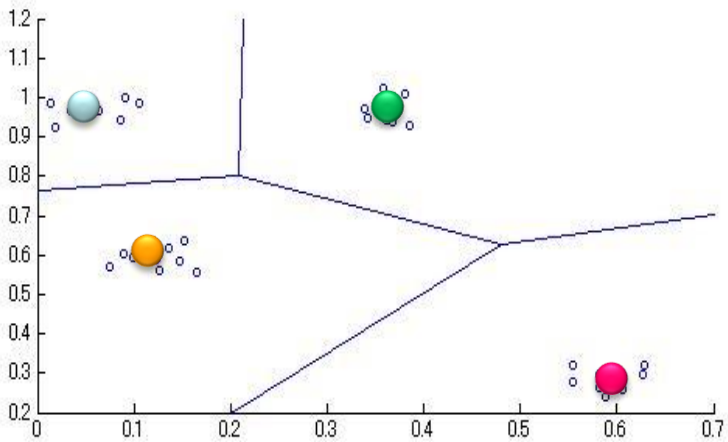
- Using WTA, we move the winner towards the input pattern, so that it can win again if a similar pattern is provided.
- This way, a neuron can become an expert for recognizing a group of “similar patterns”, and the weight of the neuron can be used as an “icon” of this group of patterns.



Example

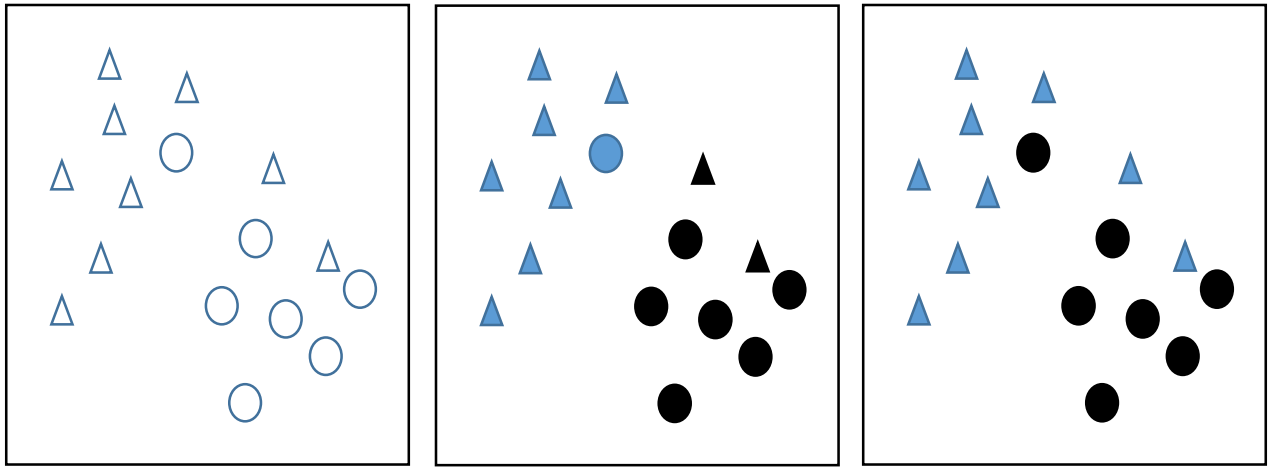
- The right figure shows patterns generated according to 4 different distributions.
- At the beginning, the neurons cannot represent the patterns well.
- After WTA-based learning however, the neurons become the centers of the patterns.





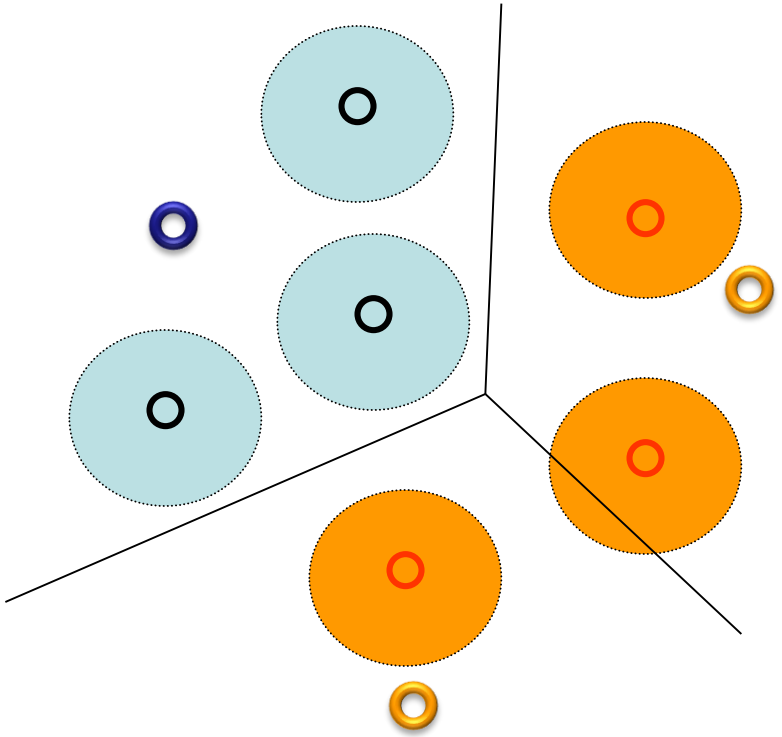
Distance-based supervised learning

- In WTA-based learning, it is not necessary to have teacher signals. Thus, it is an un-supervised learning.
- We can also conduct supervised learning based on the distance.
- One of the distance-based supervised learning methods is Learning Vector Quantization (LVQ). Supervised learning can be more accurate (see the figure given below).



Supervised learning can be more efficient

- Using self-organized learning, we can get cluster centers.
- But if we know the class labels of the cluster centers, we can reduce the number of representatives for each class.



Learning vector quantization

Step 1: Initialize the weights of all neurons at random.

Step 2: Take a pattern **x** from the training set, and find the winner **r**.

Step 3: If **r** has the same label as **x**, update it using Eq. (6.18); otherwise, update using Eq. (6.19)

$$\mathbf{r} = \mathbf{r} + \alpha(\mathbf{x} - \mathbf{r}) \qquad (6.18)$$

$$\mathbf{r} = \mathbf{r} - \alpha(\mathbf{x} - \mathbf{r}) \qquad (6.19)$$

Step 4: Terminate if the given condition is satisfied. Otherwise, return to Step 2.

Example 6.6 p. 131

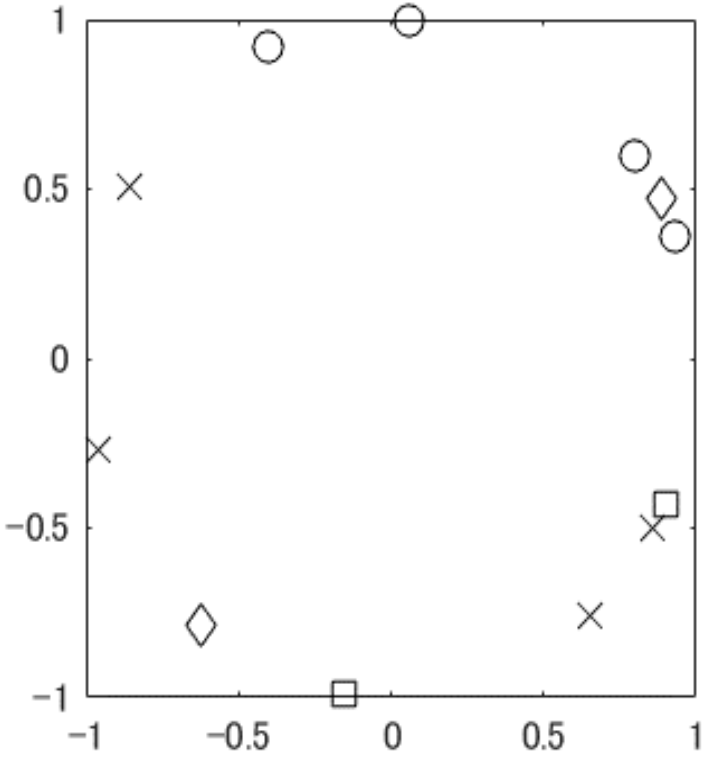
Training data

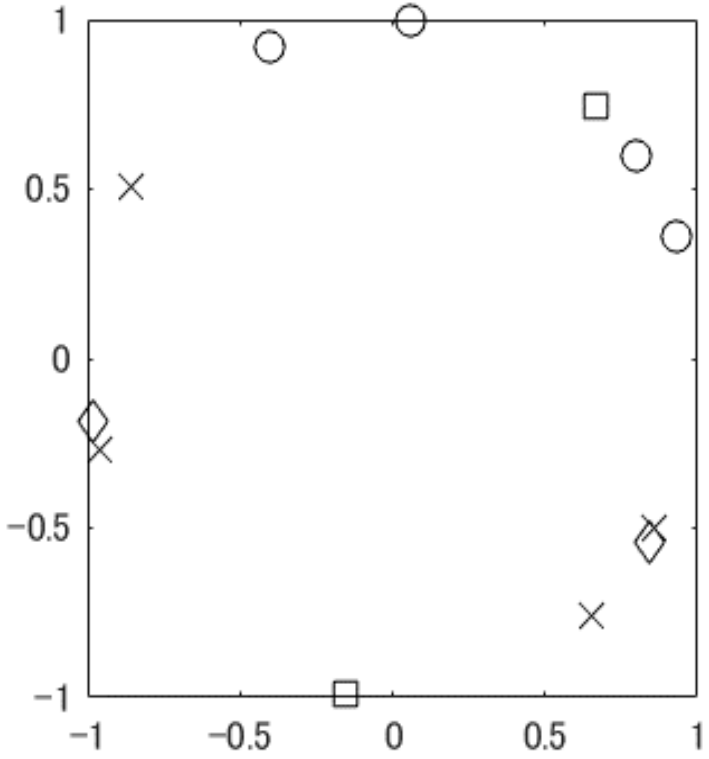
Class 0	Class 1
-0.96, -0.27	-0.400000,0.92
-0.86, 0.51	0.800000, 0.60
0.86, -0.50	0.060000, 1.00
0.65, -0.76	0.930000, 0.36

Training results

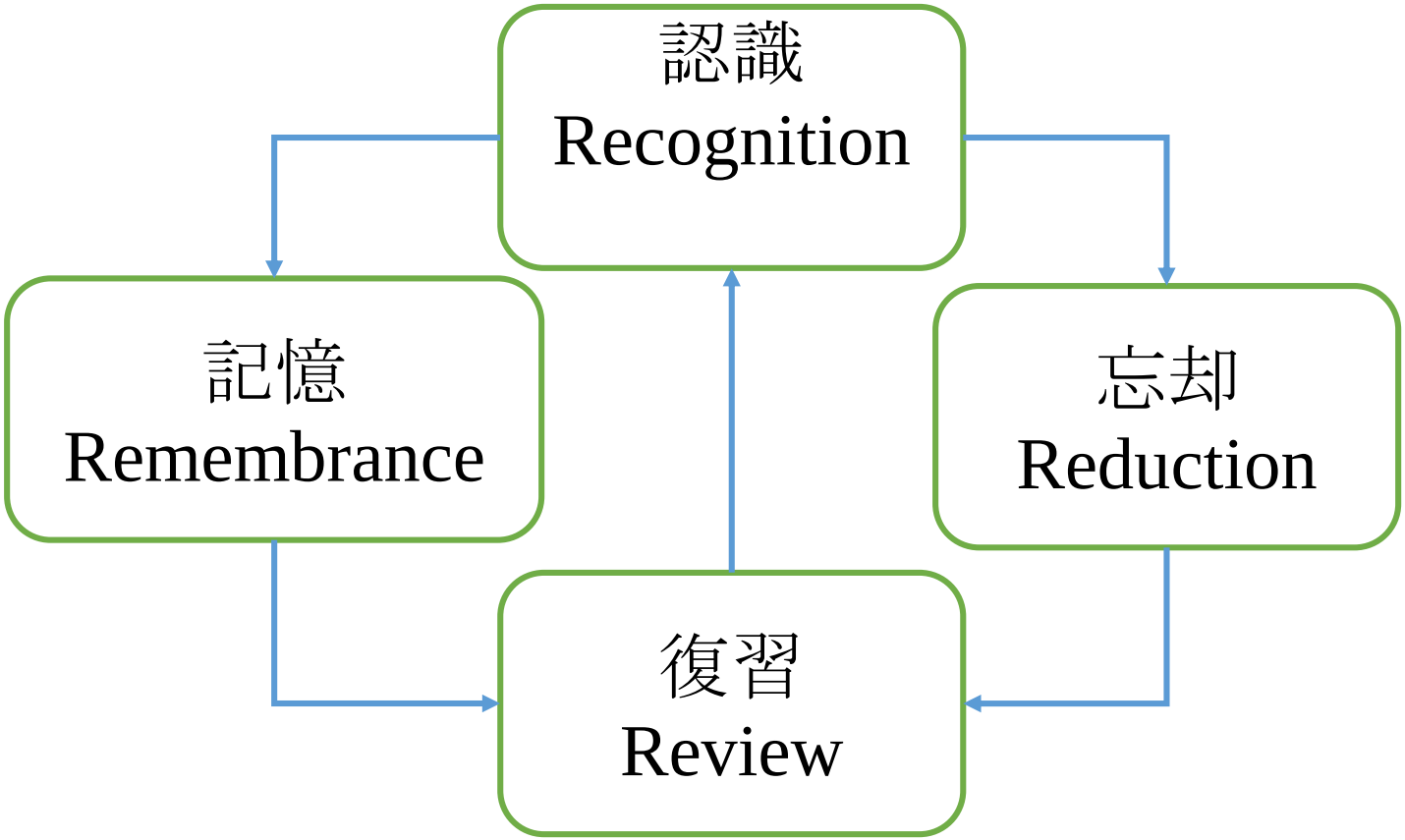
Initial neurons	Neurons after learning
0.882983 0.469404 0;	0.842607 -0.538529 0;
-0.619131 -0.785288 0;	-0.983217 -0.182438 0;
0.901795 -0.432164 1;	0.665301 0.746575 1;
-0.151238 -0.988497 1;	-0.151238 -0.988497 1;

Example 6.6





R^4-rule: A method for determining the network size



Zhao and Higuchi, IEEE Trans. on Neural Networks, 1996

1st R: Recognition

- **Recognition** is an operation for:
  - Evaluating the current system; and
  - Evaluating the fitness of each neuron.
- The first result is used to select the next operation; and the second is used to delete useless neurons.

Determine the winner:

- For any given pattern x, a neuron y is the winner if:
  - y has the same class label as x;
  - y is closer to x than any neuron with a different class label;
  - y has the highest fitness among all neurons satisfying the above two conditions.
- Increase the fitness of the winner y.

2nd R: Remembrance

- **Remembrance** is an operation for remembering un-recognizable patterns (e.g. new patterns or difficult patterns).
- Usually, only one of the un-recognizable patterns is remembered a time.
  - Method for selection: random.
  - Method for remembrance: as is.

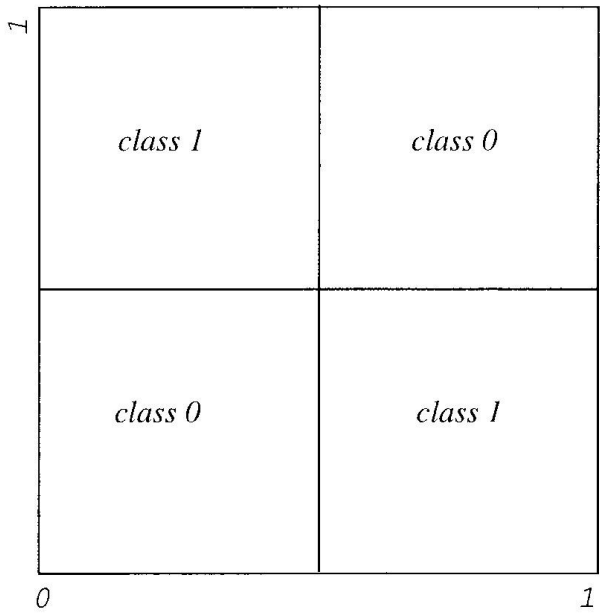
3rd R: Reduction

- **Reduction** is an operation for deleting useless neurons.
- If the current system is good enough, we can select one of the neurons with low fitness, and delete.

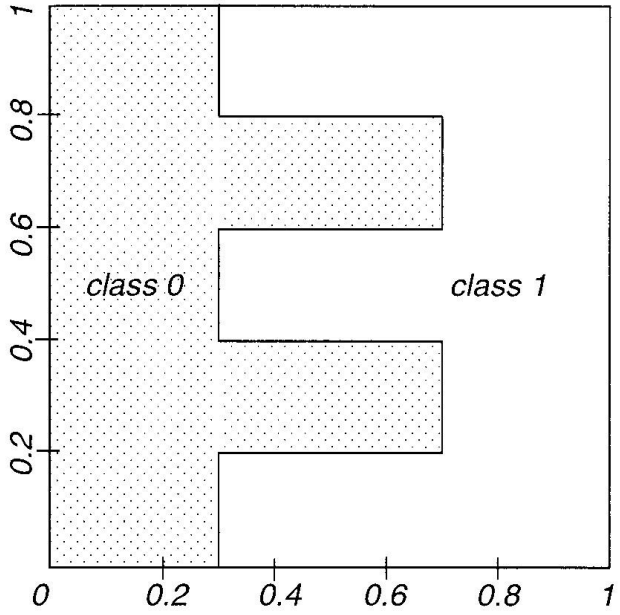
4th R: Review

- **Review** is an operation for updating the current system.
- We can use any LVQ algorithm to implement “review”.
- We recommend the DSM (decision surface mapping) algorithm because it is more efficient.

Examples



The generalized XOR problem



The straight line class boundaries problem

Results

	Generalized XOR		SLCB problem	
Methods	Number of neurons	Error rate (%)	Number of neurons	Error rate (%)
NNC	6400	0.91	6400	0.97
CNN	231	1.11	307	1.17
RNN	162	1.27	216	1.14
RCE	183	0.69	243	0.56
R^4-rule	**4**	**0.13**	**10**	**0.48**

Evaluation of algorithms

- To evaluate the performance of a learning algorithm, we usually divide the data into two parts: **training data and testing data**.
- Training data are used to train a learning model, and testing data is used to test.
- The performance of the model for the training data can be used to check if learning is actually conducted or not.
- The performance of the model for the testing data can be used to check the "**generalization ability**" of the model. This is usually considered more important in machine learning.

Evaluation of algorithms

- If we just test the model using one testing data set, the result might be biased.
- To void this problem we often use n-fold cross validation.
- In n-fold cross validation, we divide the data into n (equal) parts, and use each of them for testing and the remaining for training.
- We can thus train n models, and the average performance is often used to draw a conclusion.
- The standard deviation is often used to measure the confidence of the conclusion.

Homework for lecture 11 (1)

- Solve Problem 6.5 in p. 132 of the textbook, and submit the results during the exercise class.

Homework for lecture 11 (2)

- Based on the skeleton program,
  - Complete a program for WTA-based self-organized learning.
  - There is a data file generated from 5 different distributions. Find the representative patterns using this data file, and draw a figure.
  - You may use “make plot” to draw the figure.
  - Write your comments based on the figure in “summary.txt”.

Quizzes of today

- KNN is the short for _____ .
- WTA is the short for _____.
- LVQ is the short for_____.
- R4-rule has the following 4 operations:
  - R_____,
  - R_____ ,
  - R_____ , and
  - R_____.

- Try to explain the physical meaning of the WTA algorithm.
- Why KNN is smarter compared with the NNC using all training data?