Topics in Data Engineering

Session 5

Masaomi Kimura

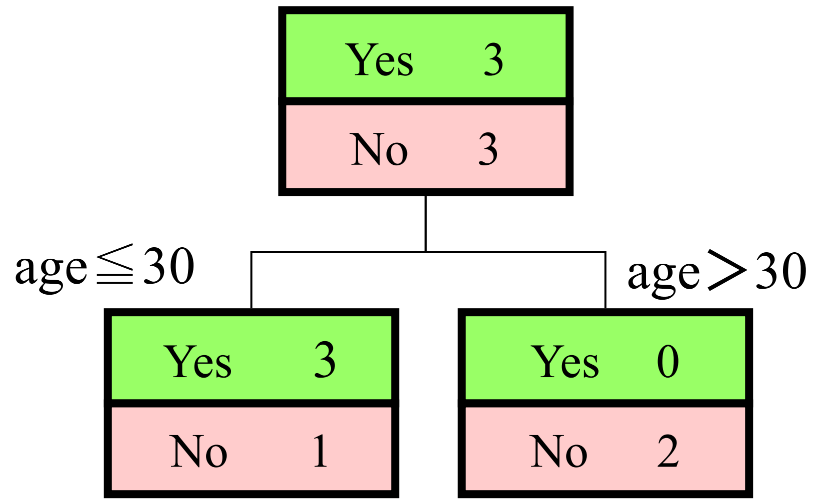
Today's topic

- Decision tree
- The foundation of Artificial Neural Network

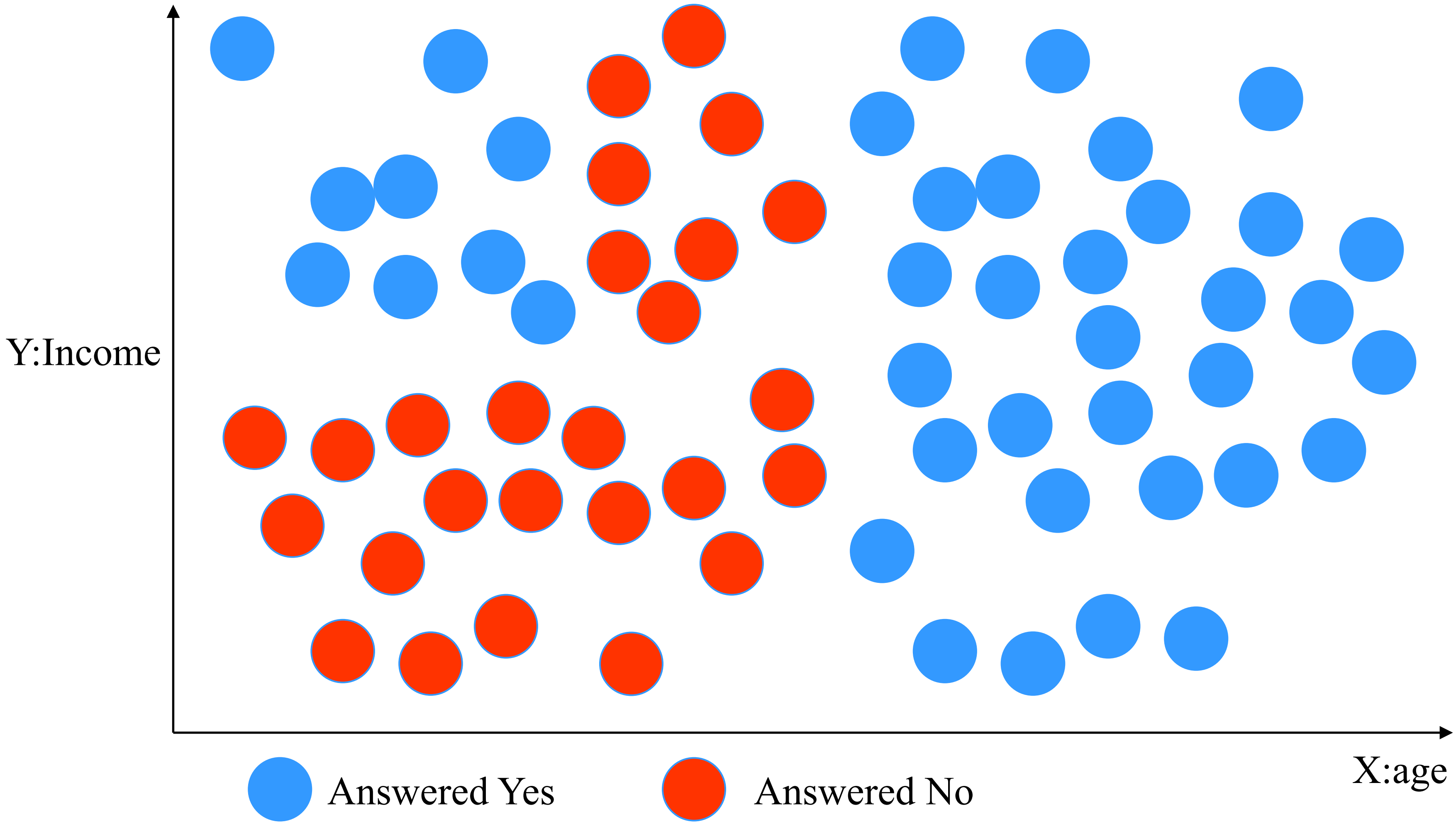
Decision trees

- The method to find conditions that purify target data after division

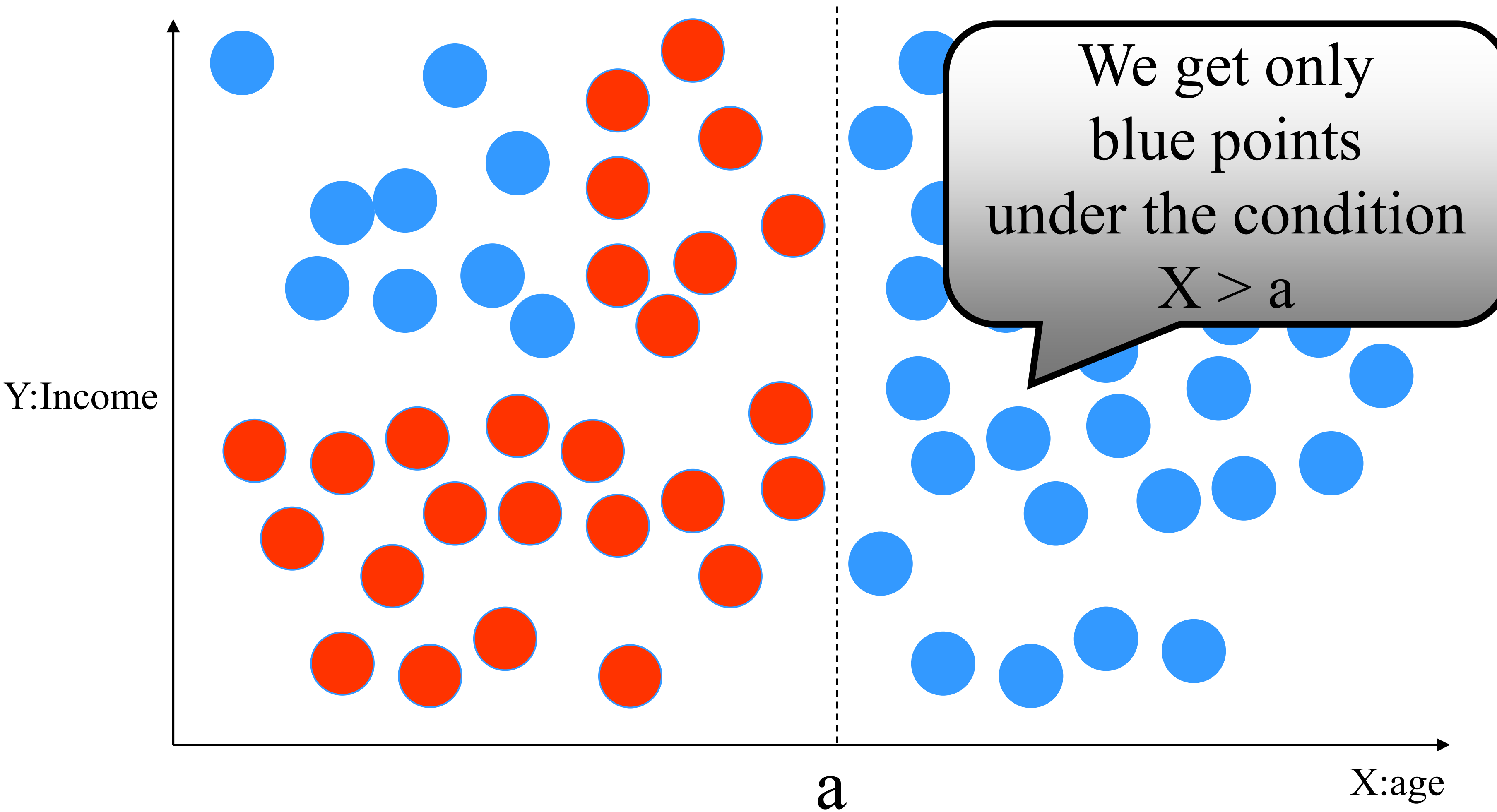
Want to buy?	ages
Yes	21
Yes	25
Yes	30
No	29
No	50
No	60



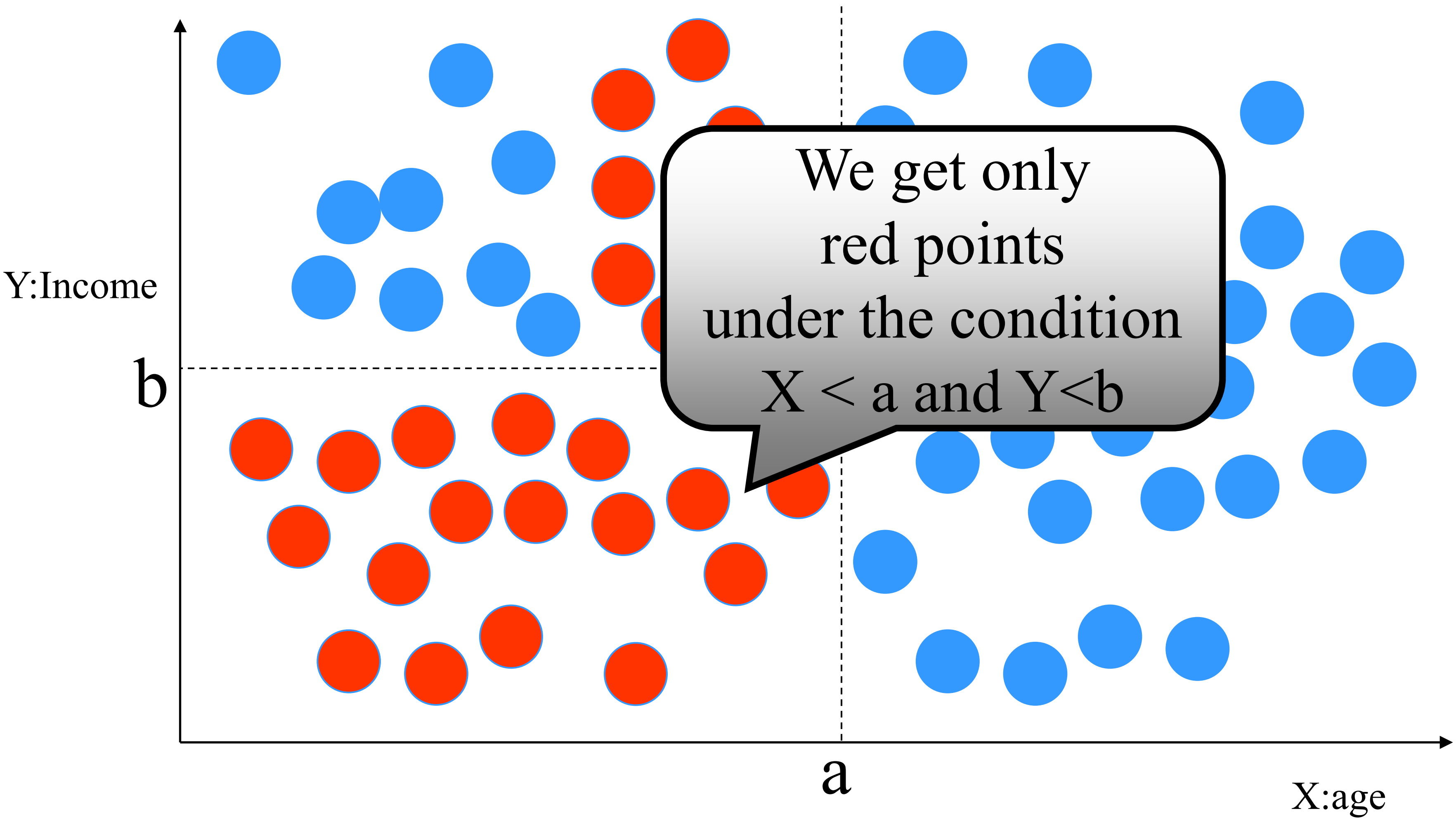
A key concept of decision tree:
Find borders to separate reds and blues



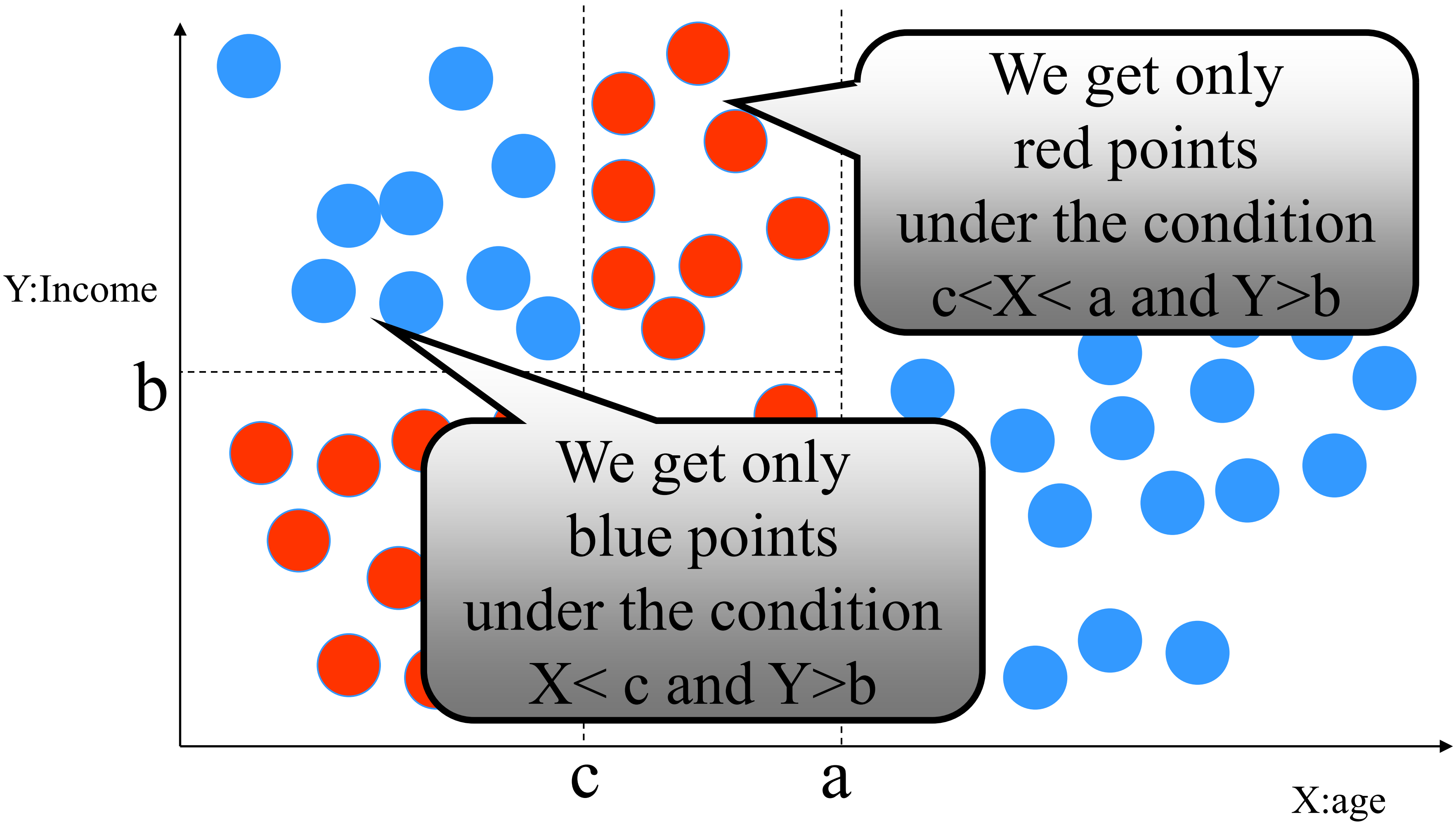
A key concept of decision tree: Find borders to separate reds and blues



A key concept of decision tree:
Find borders to separate reds and blues

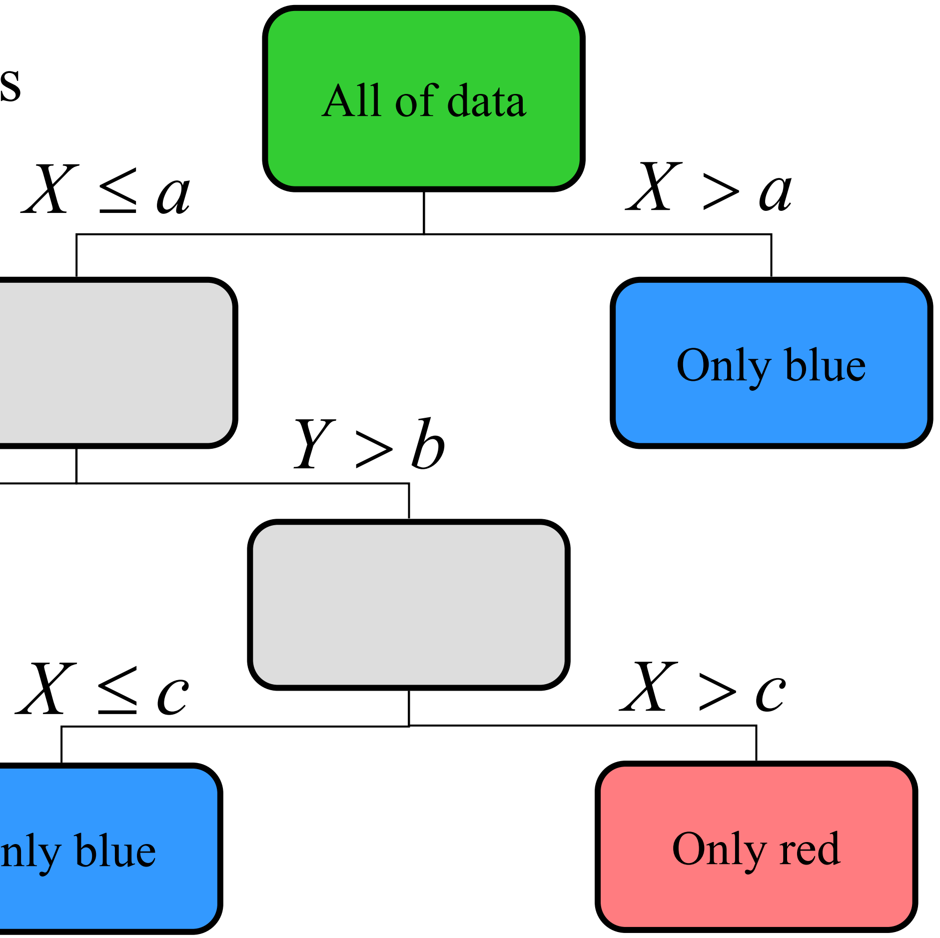


A key concept of decision tree:
Find borders to separate reds and blues



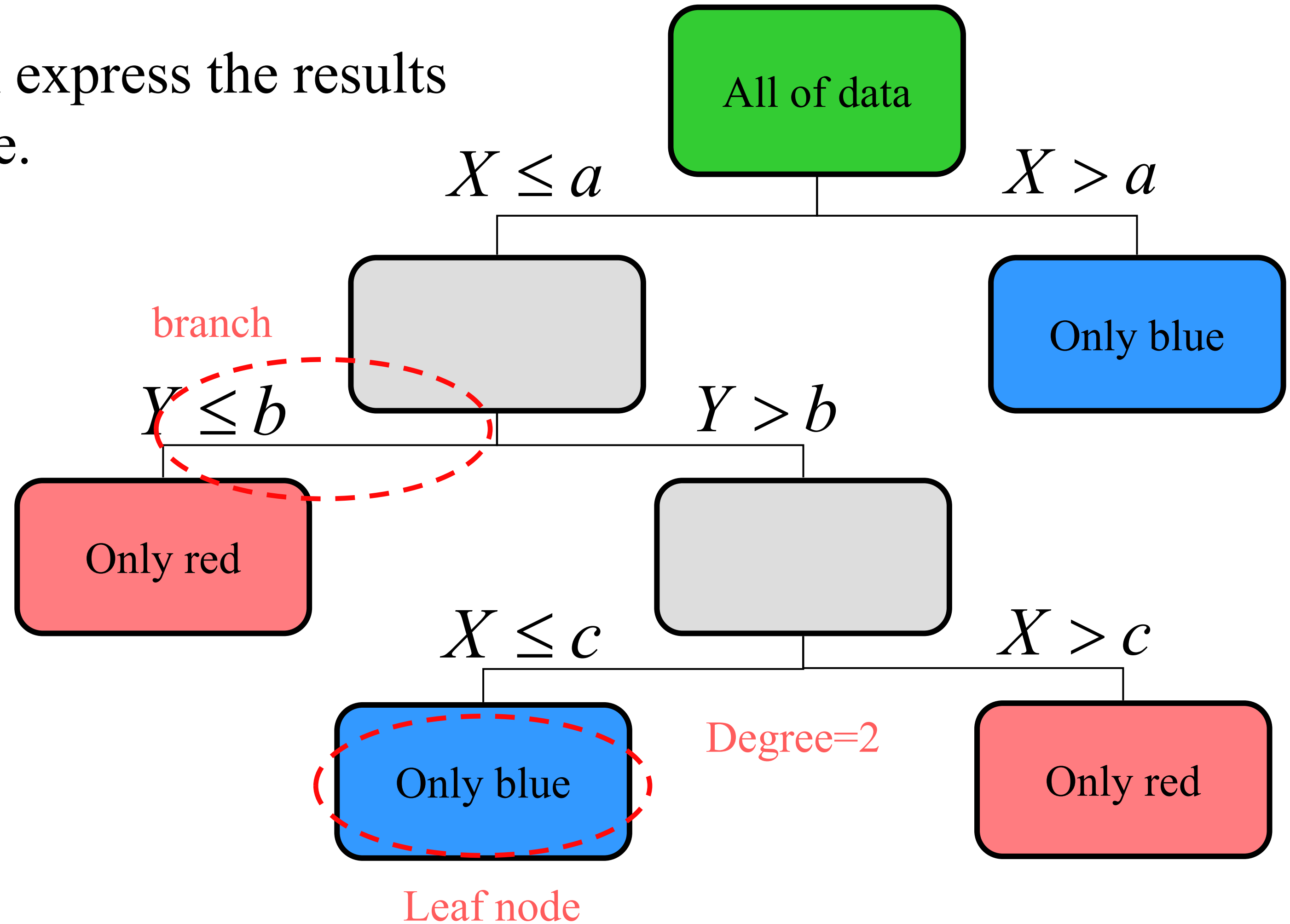
In summary,…

We can express the results in a tree.



In summary,…

We can express the results in a tree.



In reality,…

- It is usually difficult to distinguish ‘blues’ and ‘reds’ in reality,
  - especially, if the borders are not parallel to axes
- However! The idea to find borders is useful to get conditions to separate a target.

⇒

We need a method to find a border that gives biased separation into homogenous parts
(red part/blue part in this case)

Homogeneity indices

- Gini index

$$GINI = 1 - \sum_{i=1}^{C} p_i^2$$

- Entolopy

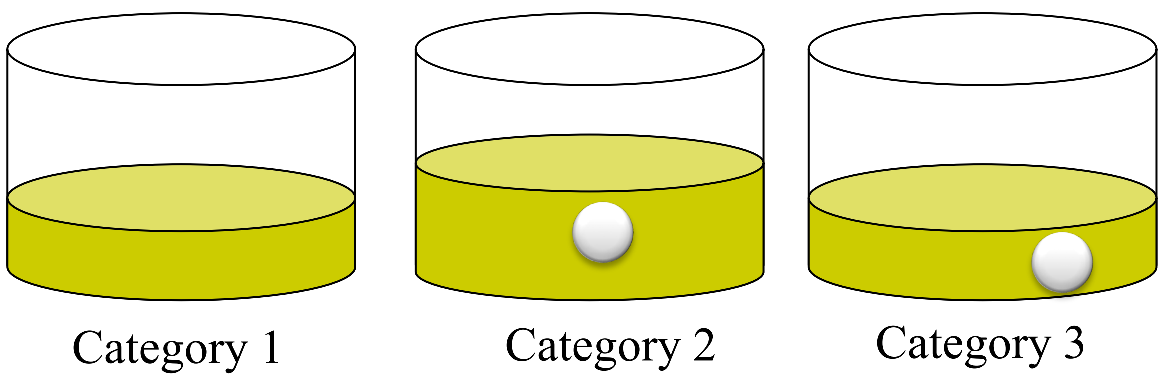
$$S = -\sum_{i=1}^{C} p_i \log p_i$$

C: the number of category
p_i: relative frequencies in the i^{th} category

Gini index(1)

$$GINI = 1 - \sum_{i=1}^{C} p_i^2 = \sum_{i=1}^{C} \sum_{j \neq i} p_i p_j$$

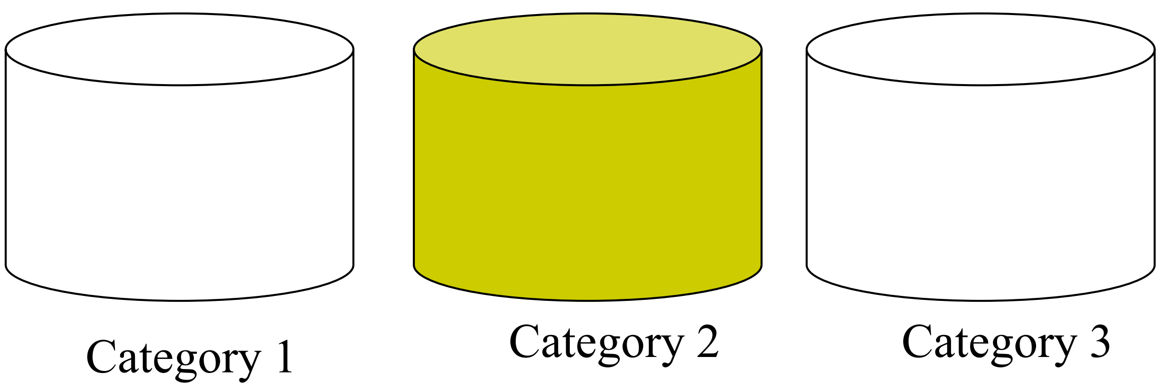
If we interpret p_i as a probability, GINI index is a probability that two data belong to different categories.



Gini index(2)

$$GINI = 1 - \sum_{i=1}^{C} p_i^2 = 1 - (0^2 + 1^2 + 0^2) = 0$$

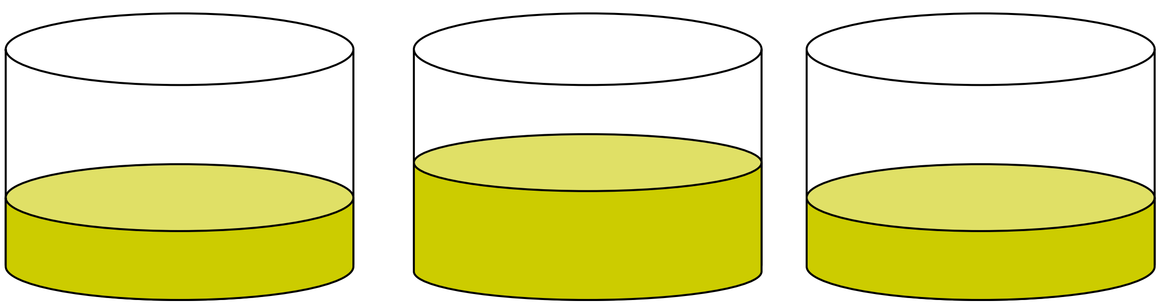
$$p_1 = 0, p_2 = 1, p_3 = 0$$



Gini index(3)

$$GINI = 1 - \sum_{i=1}^{C} p_i^2 = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{4}\right)^2 \right) = \frac{5}{8} > 0$$

$$p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$$



Category 1 Category 2 Category 3

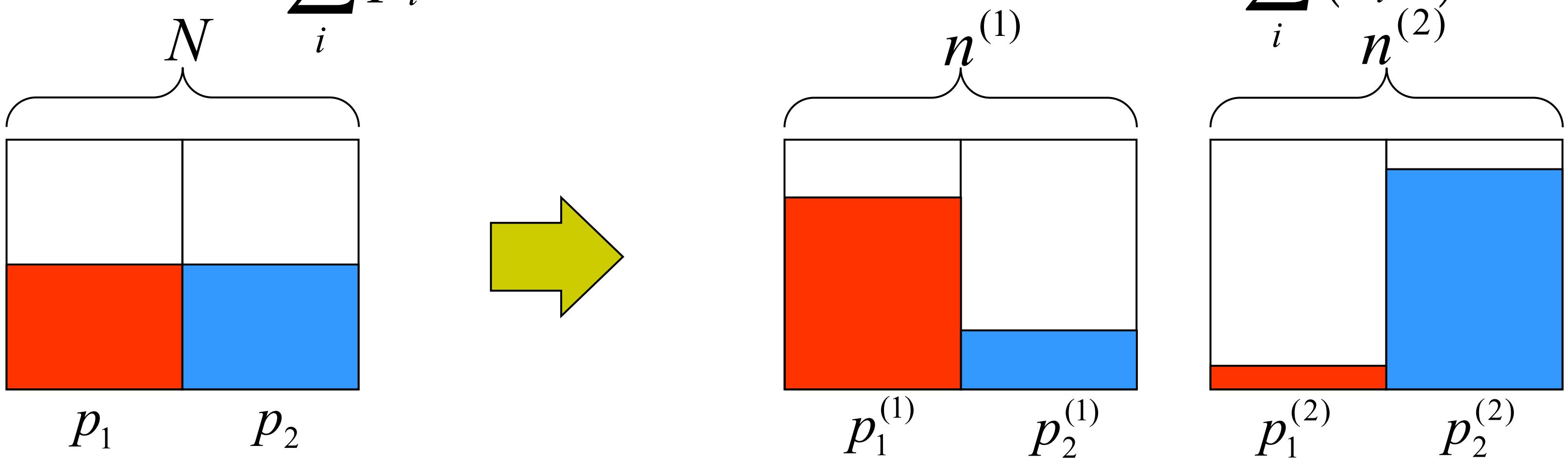
Find a condition based on Gini index

- Choose a condition that gives the largest difference of original Gini from averaged Gini after separation

$$\Delta GINI = GINI - \sum_j \frac{n^{(j)}}{N} GINI^{(j)}$$

$$GINI = 1 - \sum_i p_i^2$$

$$GINI^{(j)} = 1 - \sum_i \left(p_i^{(j)}\right)^2$$



Find a condition based on Gini index

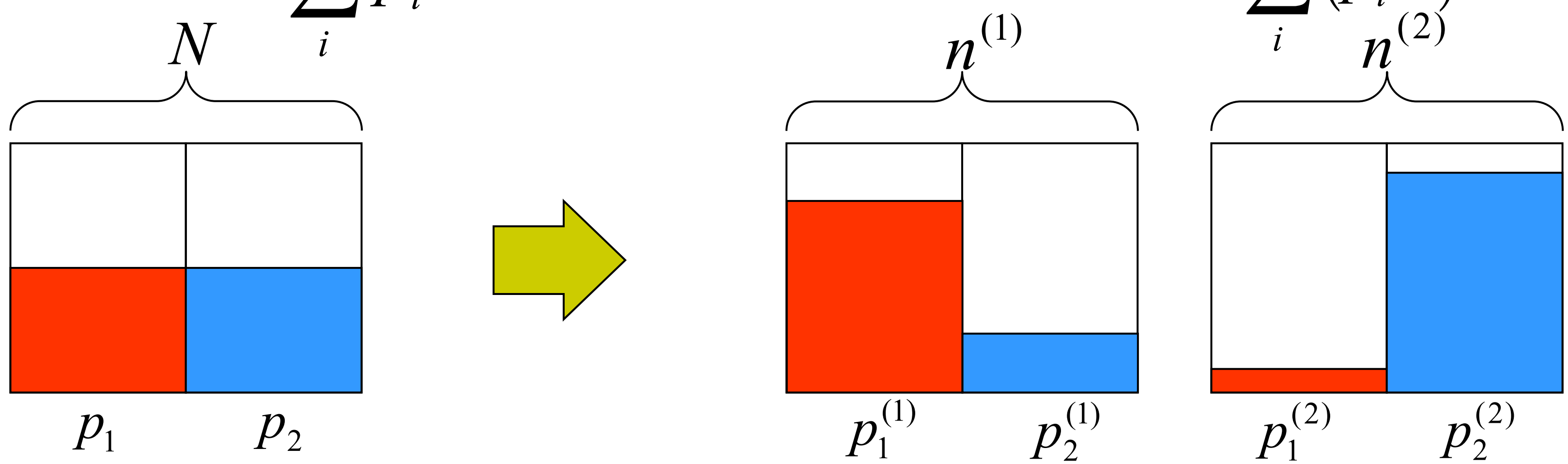
- Choose a condition that gives the largest difference of original Gini from averaged Gini after separation

Less, if homogenous

constant

$$\Delta GINI = GINI - \boxed{\sum_j \frac{n^{(j)}}{N} GINI^{(j)}}$$

$$GINI = 1 - \sum_i p_i^2 \qquad GINI^{(j)} = 1 - \sum_i \left(p_i^{(j)}\right)^2$$



Entropy (1)

$$S = -\sum_{i=1}^{C} p_i \log_2 p_i$$

Entropy: the average amount of information

C: the number of category

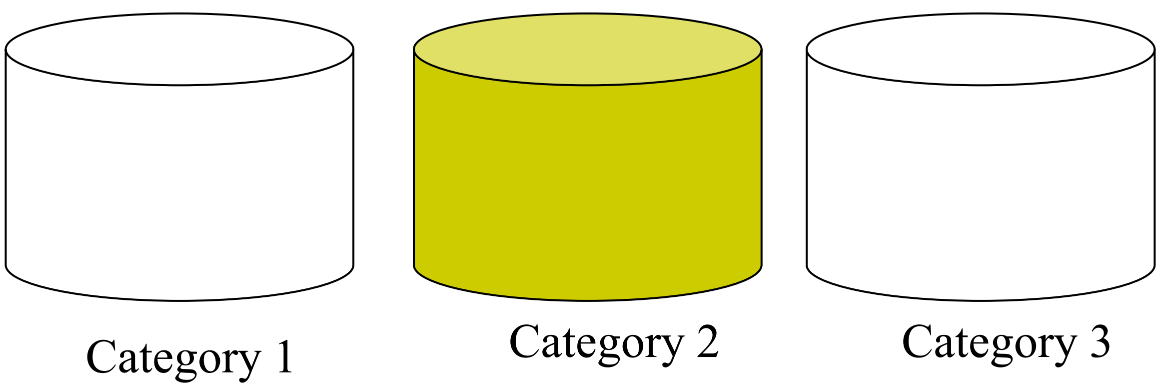
p_i: relative frequencies in the i^{th} category

Entropy (2)

$$S = -\sum_{i=1}^{C} p_i \log_2 p_i = -0\log_2 0 - 1\log_2 1 - 0\log_2 0 = 0$$

0log0 is defined to be 0, because x logx →0 (x→0)

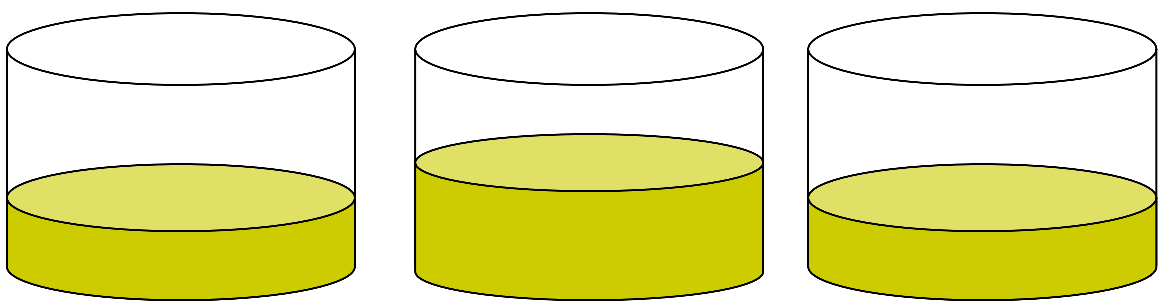
$$p_1 = 0, p_2 = 1, p_3 = 0$$



Entropy (3)

$$S = -\sum_{i=1}^{C} p_i \log_2 p_i = \frac{1}{4}\log_2 4 + \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 = \frac{3}{2} > 0$$

$$p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$$



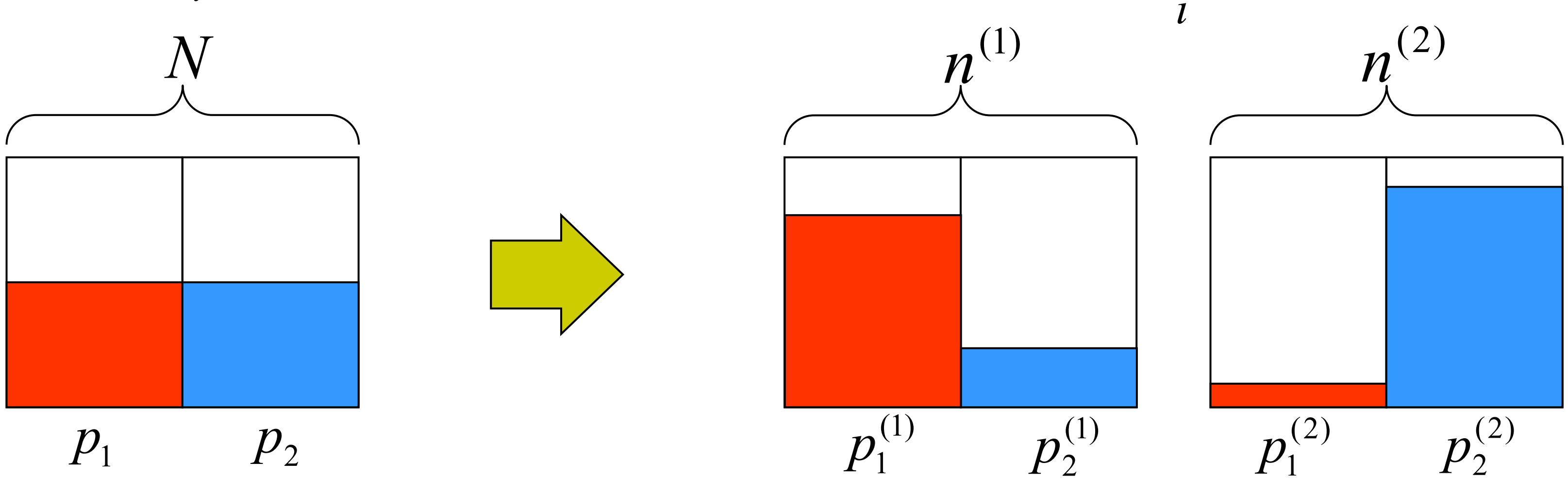
Category 1 Category 2 Category 3

Find a condition based on entropy

- Choose a condition that gives the largest difference of original entropy from averaged entropy after separation

$$\Delta S = S - \sum_j \frac{n^{(j)}}{N} S^{(j)}$$

$$S = -\sum_i p_i \log p_i \qquad S^{(j)} = -\sum_i p_i^{(j)} \log p_i^{(j)}$$



CART(C&RT)

- A binary decision tree
  - At each junction, there are two branches.
  - A tree tends to be grown to leaves (downwards).
  - Because of a binary tree structure, if natural degree (=the number of branches) is more than two, the tree might have unnatural branches.
- Suitable for numerical data
- It uses Gini index

C5.0

- A decision tree suitable for categorical data
  - The degree of each node can be more than two.
  - A tree tends to be grown to width direction.
  - Because of the high degree at a node, the number of data corresponding to the node can decrease quickly.
    - Therefore, reliability of training/prediction can be low.
- It uses entropy.

Pruning

- Too many leaf nodes usually cause over-fit to training data
  - because of small number of training data in the leaf node
- Prune branches based on a standard
  - Minimize <u>error rate + the number of leaf nodes</u>
  - Minimize error rate for test data other than training data

Advantage/disadvantage

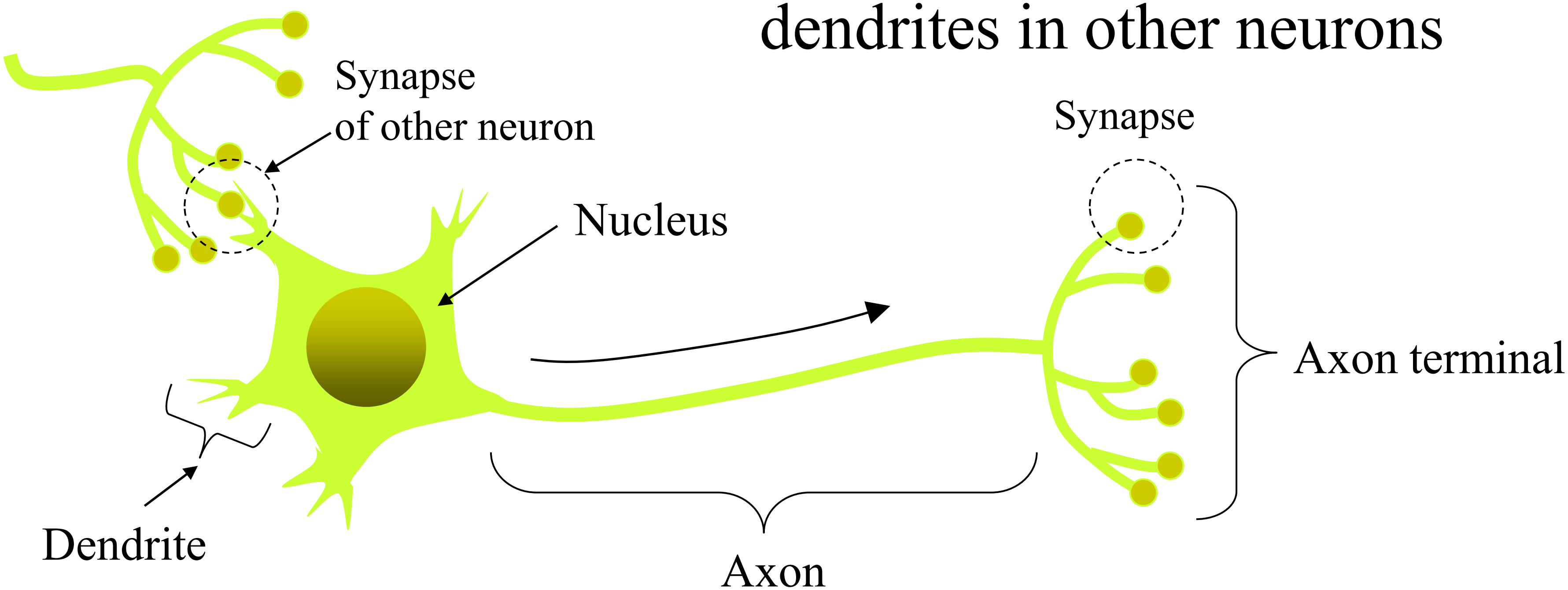
- Advantage
  - Easy to understand results and their reason
  - Applicable to both numerical variables and categorical variables
- Disadvantage
  - Has difficulty for data whose borders are not parallel to axes of explanatory variables
  - Gives low accuracy with too many leaf nodes

Perceptron

An introduction of artificial neural network

Neuron

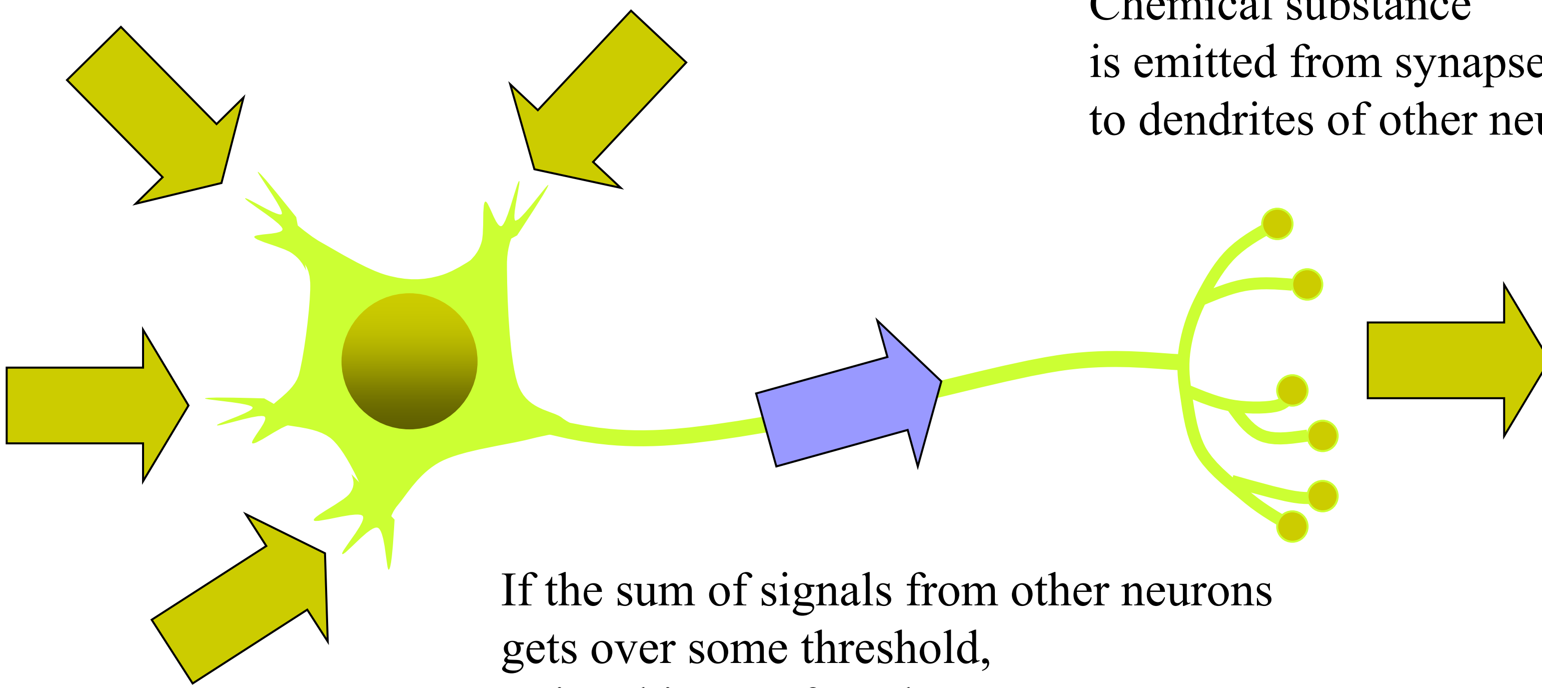
- cells in nervous system
  - contains dendrites, a nucleus, an axon
  - Synapses at an axon terminal send a signal to dendrites in other neurons



Signal transfer

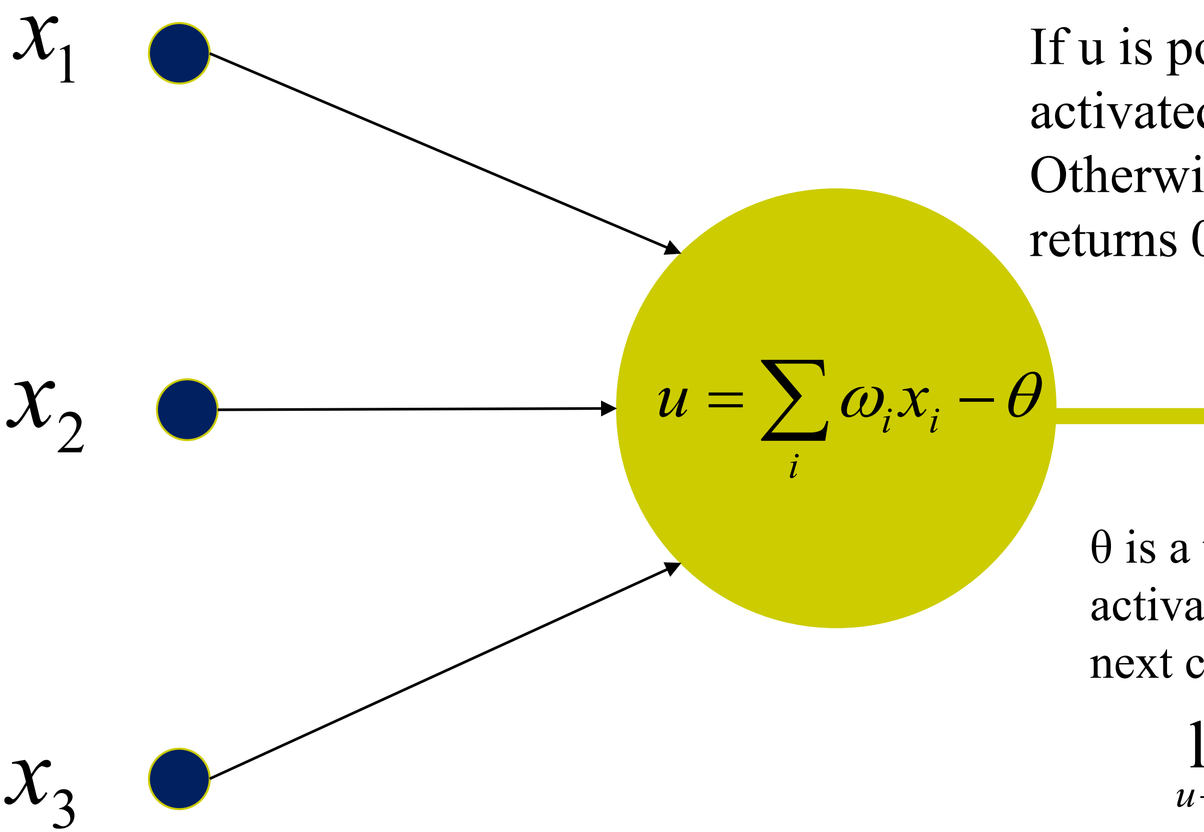
Signals from synapses of other neurons

Chemical substance
is emitted from synapses
to dendrites of other neurons



If the sum of signals from other neurons
gets over some threshold,
a signal is transferred to an axon.

Perceptron



If u is positive, the neuron is activated and returns 1 as output. Otherwise, it is not activated and returns 0.

$$z = f(u)$$

θ is a threshold, f(u) is an activation function satisfying next conditions

$$\lim_{u \to -\infty} f(u) = 0,$$

$$\lim_{u \to \infty} f(u) = 1.$$

Activation function *f(u)*

- Step function

$$f(u) = \begin{cases} 0 & (u < 0) \\ 1 & (u \geq 0) \end{cases}$$

- Sigmoid function

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

- Hyperbolic tangent function

$$\tanh(u) = \frac{e^{u} - e^{-u}}{e^{u} + e^{-u}}$$

Exercise

- Let’s find the behavior of a sigmoid function $\sigma(x)$.
  - If we let x get larger ($x \to \infty$), explain the behavior of $\sigma(x)$.
  - If we let x get smaller ($x \to -\infty$), explain the behavior of $\sigma(x)$.
  - Write a graph of $\sigma(x)$ in the range $-5 < x < 5$.
- Calculate the derivative$\sigma'(x)$
  - Find the range of$\sigma'(x)$

Training of perceptrons

- A training set of an input vector $\{x_i\}$ and its expected output value z_i^* is used to determine weights $\{\omega_k\}$.
- adjust the weights $\{\omega_k\}$ in order to minimize the mean squared error R based on Gradient descent (z_i is an output for the an input vector x_i, z_i^* is a correct output, ε is a constant.)

$$R = \frac{1}{2}\sum_i (z_i - z_i^*)^2$$

$$\delta\omega_i = -\varepsilon\frac{\partial R}{\partial\omega_i} = \varepsilon\sum_k (z_k^* - z_k)\frac{\partial z_k}{\partial\omega_i} = \varepsilon\sum_k (z_k^* - z_k)x_k^i\frac{\partial f(u_k)}{\partial u}$$

Exercise

- Let $w' = w - \epsilon \frac{\partial}{\partial w} R(w)$, where ϵ is a small and positive constant.
- Which is larger, $R(w)$ or $R(w')$?
  - Explain why?

Exercise

- Let $z(x_1, x_2) = \sigma(w_1 x_1 + w_2 x_2 - b)$.

- Discuss whether we can make a perceptron which satisfies the following table:

x_1	x_2	z
0.0	0.0	0.0
1.0	0.0	1.0
0.0	1.0	1.0
1.0	1.0	0.0