Proof of Termination

w is updated once it sees a mistake

$$w^0 = 0 \rightarrow w^1 \rightarrow w^2 \rightarrow \ldots\ldots \rightarrow w^k \rightarrow w^{k+1} \rightarrow \ldots\ldots$$

$$w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)$$ (the relation of w^k and w^{k-1})

Proof that: The angle ρ_k between $\hat{w}$ and w_k is smaller as k increases

分子會越來越大

Analysis $\cos\rho_k$ (larger and larger?) $\cos\rho_k = \frac{\hat{w}}{\|\hat{w}\|} \cdot \frac{w^k}{\|w^k\|}$

$$\hat{w} \cdot w^k = \hat{w} \cdot \left(w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\right)$$

$$= \hat{w} \cdot w^{k-1} + \underbrace{\hat{w} \cdot \phi(x^n, \hat{y}^n) - \hat{w} \cdot \phi(x^n, \tilde{y}^n)}_{\geq \delta} \geq \hat{w} \cdot w^{k-1} + \delta$$

$\geq \delta$ (Separable)

Proof of Termination

w is updated once it sees a mistake

$$w^0 = 0 \rightarrow w^1 \rightarrow w^2 \rightarrow \ldots\ldots \rightarrow w^k \rightarrow w^{k+1} \rightarrow \ldots\ldots$$

$$w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)$$ (the relation of w^k and w^{k-1})

Proof that: The angle ρ_k between $\hat{w}$ and w_k is smaller as k increases

Analysis $\cos\rho_k$ (larger and larger?)

$$\cos\rho_k = \frac{\boxed{\hat{w}}}{\|\hat{w}\|} \cdot \frac{\boxed{w^k}}{\boxed{\|w^k\|}}$$

$$\hat{w} \cdot w^k \geq \hat{w} \cdot w^{k-1} + \delta$$

$$\hat{w} \cdot w^1 \geq \underbrace{\hat{w} \cdot w^0}_{=0} + \delta \qquad \hat{w} \cdot w^2 \geq \underbrace{\hat{w} \cdot w^1}_{\geq\delta} + \delta \qquad \ldots\ldots$$

$$\hat{w} \cdot w^1 \geq \delta \qquad \hat{w} \cdot w^2 \geq 2\delta \qquad \ldots\ldots$$

upper bound

$$\hat{w} \cdot w^k \geq k\delta$$

斜直線

(so what)

Proof of Termination

mistake: 前幾頁是說y^跟w取內積要是max，而一開始求argmax的時候有說明y~是找出投影上w(k−1)是最大的值

$$\cos \rho_k = \frac{\hat{w}}{\|\hat{w}\|} \cdot \frac{w^k}{\boxed{\|w^k\|}}$$

$$w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)$$

$$\|w^k\|^2 = \left\|w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\right\|^2$$

想像成(a+b)平方

$$= \|w^{k-1}\|^2 + \|\underbrace{\phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)}_{>0}\|^2 + \underbrace{2w^{k-1} \cdot (\phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n))}_{?\ <0 \text{ (mistake)}}$$

Assume the distance between any two feature vector is smaller than R

假設所有feature分佈之間的距離小於R

$$\leq \|w^{k-1}\| + R^2$$

$$\|w^1\|^2 \leq \|w^0\|^2 + R^2 = R^2$$

$$\|w^2\|^2 \leq \|w^1\|^2 + R^2 \leq 2R^2$$

$$\cdots$$

$$\|w^k\|^2 \leq kR^2$$

Proof of Termination

$$\cos\rho_k = \frac{\hat{w}}{\|\hat{w}\|}\cdot\frac{w^k}{\|w^k\|} \qquad \hat{w}\cdot w^k \ge k\delta \qquad \|w^k\|^2 \le k\mathrm{R}^2$$

$$\ge \frac{k\delta}{\sqrt{kR^2}} = \boxed{\sqrt{k}\,\frac{\delta}{R}}$$

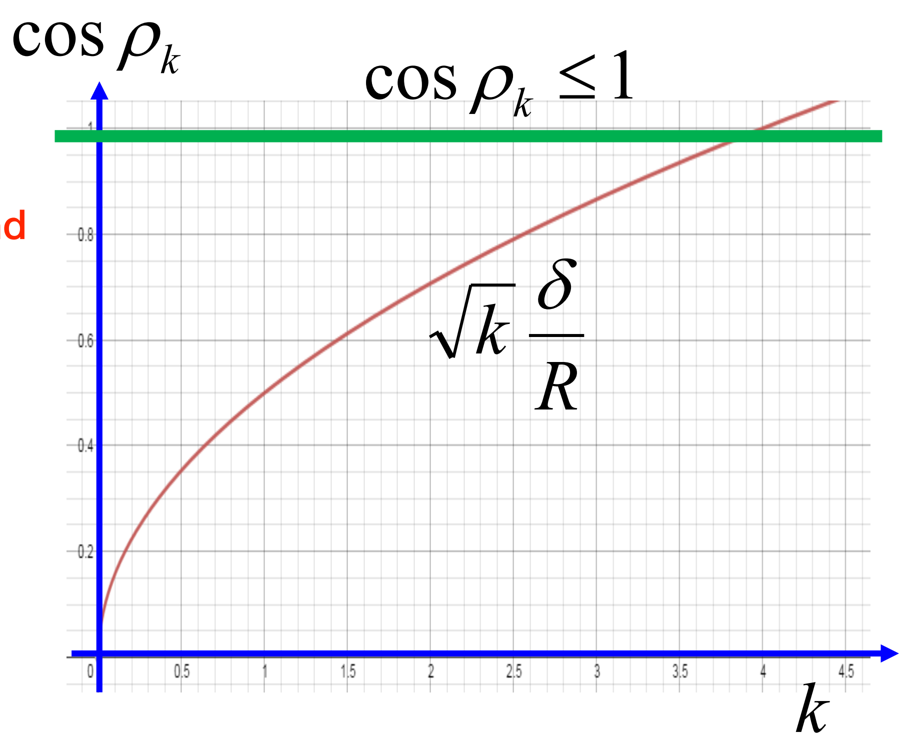
cos的lower bound

$$\sqrt{k}\,\frac{\delta}{R} \le 1$$

$$k \le \left(\frac{R}{\delta}\right)^2$$

k的最大值

（最多update這麼多次這個演算法就會結束）



Proof of Termination

feature間距離的最大值

$$k \le \left(\frac{R}{\delta}\right)^2$$

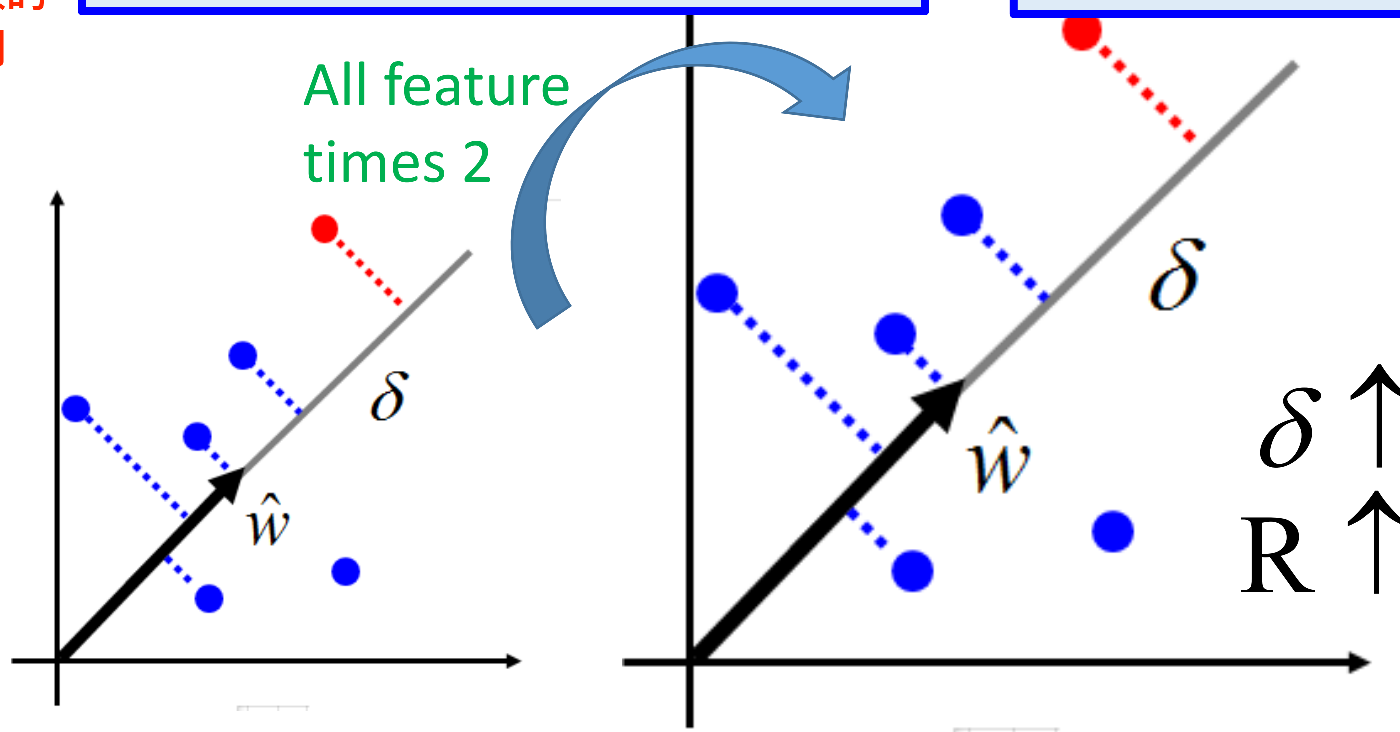
The largest distances between features → Normalization

Margin: Is it easy to separable red points from the blue ones → Larger margin, less update

delta:
正確的example跟錯誤的example分佈的有多開

● $\phi\left(x^r, \hat{y}^r\right)$

● $\phi\left(x^r, y\right)$



Structured Linear Model: Reduce 3 Problems to 2

如果function是linear的話可以
用structure perceptron來解

$F(x,y)=w\cdot\phi(x,y)$

前提是要先能夠解出arg max

Problem 1: Evaluation

- How to define F(x,y)

Problem 2: Inference

- How to find the y with the largest F(x,y)

Problem 3: Training

- How to learn F(x,y)

Problem A: Feature

- How to define $\phi(x,y)$

Problem B: Inference

- How to find the y with the largest $w\cdot\phi(x,y)$

Graphical Model

A language which describes the evaluation function

Structured Learning

We also know how to involve hidden information.

Problem 1: Evaluation 假設為linear

- What does F(x,y) look like? $F(x,y) = w \cdot \phi(x,y)$

Problem 2: Inference

- How to solve the "arg max" problem

$$y = \arg\max_{y \in Y} F(x,y)$$

Problem 3: Training

- Given training data, how to find F(x,y) Structured SVM, etc.

solve: structure perceptron/structure SVM

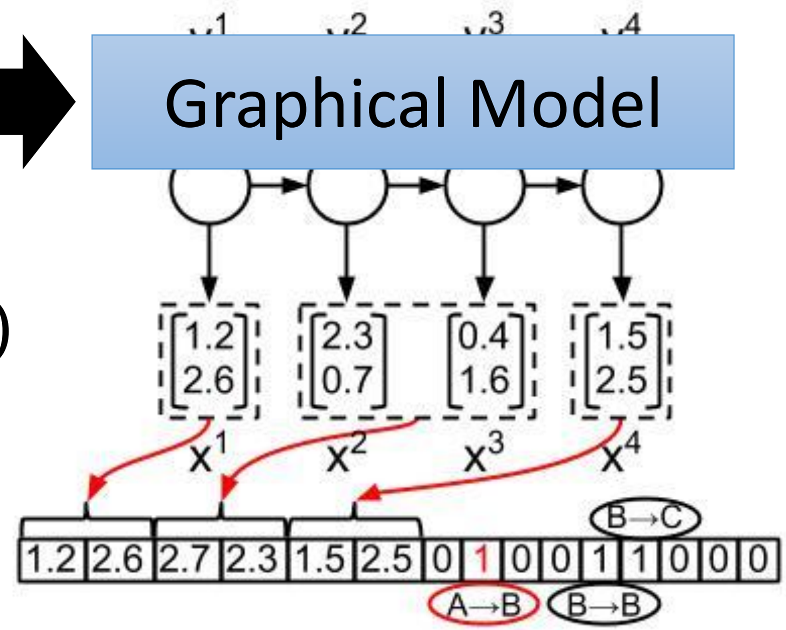
Difficulties

怎麼設計evaluation function？

Difficulty 1. Evaluation

Graphical Model

$$F(x,y) = w \cdot \phi(x,y)$$



$\phi(x,y)$

Hard to figure out? Hard to interpret the meaning?

怎麼解inference

Difficulty 2. Inference

Gibbs Sampling

We can use Viterbi algorithm to deal with sequence labeling. How about other cases?

Graphical Model

$$F(x, y) \Longleftrightarrow \text{Graph}$$

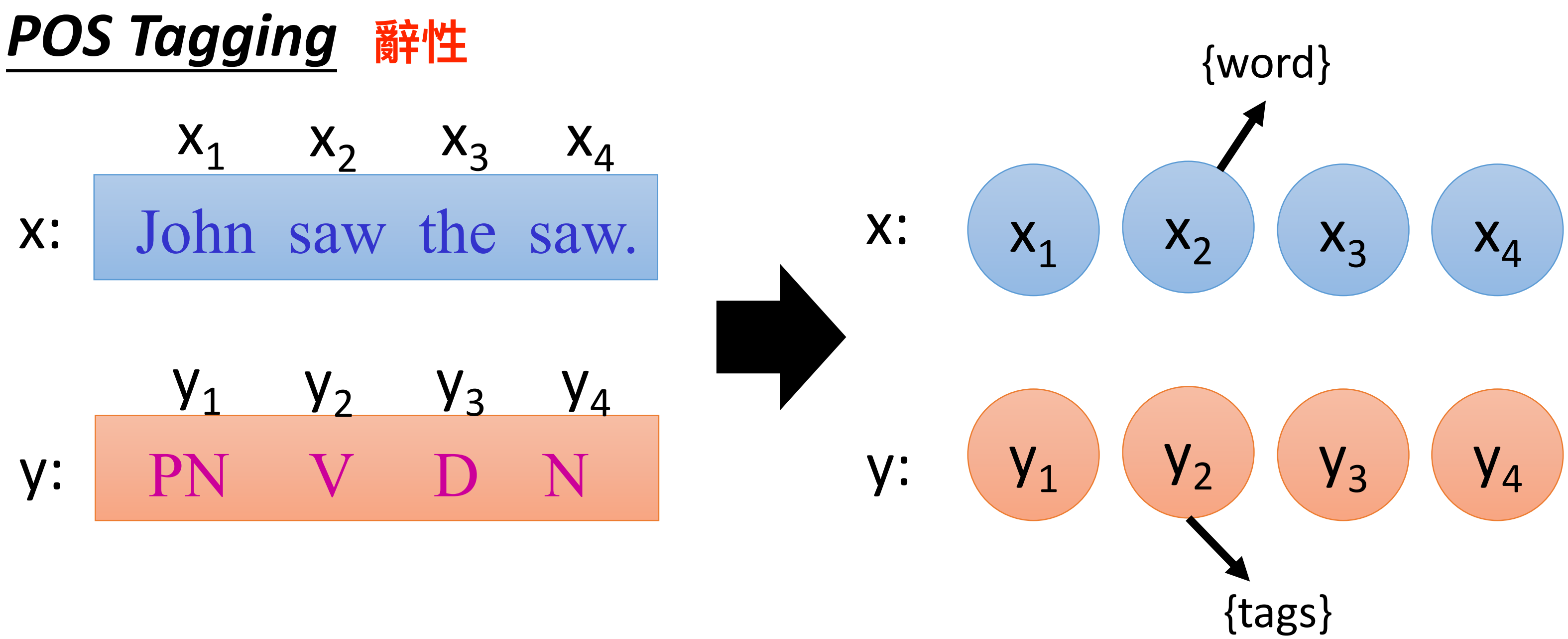
- Define and describe your evaluation function F(x,y) by a graph
- There are three kinds of graphical model.
  - *Factor graph*, *Markov Random Field* (MRF) and *Bayesian Network* (BN)
  - Only *factor graph* and *MRF* will be briefly mentioned today.

Decompose F(x,y)

- $F(x, y)$ is originally a ***global*** function
  - Define over the whole x and y x,y是一個有結構的物件
- Based on graphical model, $F(x, y)$ is the composition of some ***local*** functions
  - x and y are decomposed into smaller components 拆成很多local function的和，且每個local function 代表x,y的一部分components(features)
  - Each local function defines on only a few related components in x and y
  - Which components are related → defined by Graphical model

Decomposable x and y

- x and y are decomposed into smaller components



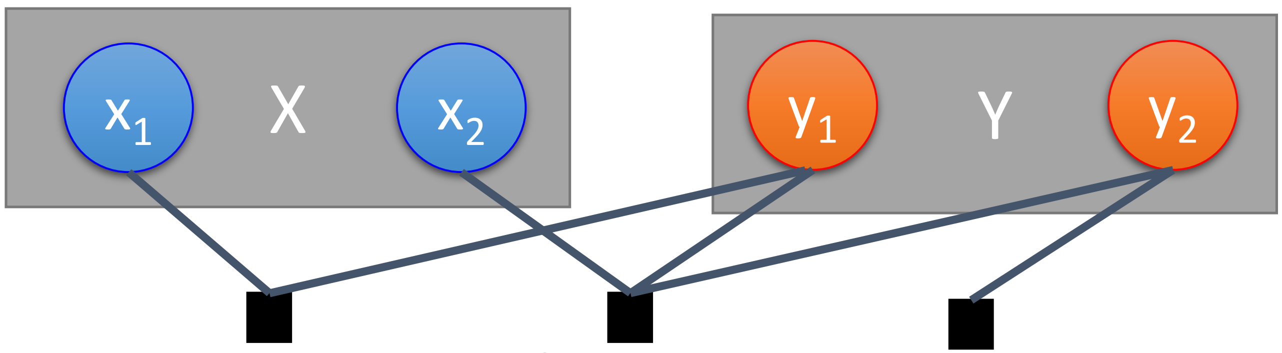
Factor Graph

假設x & y的關係是由一些factor所組成

每一個factor都對應到一個function

Each factor influences some components.

Each factor corresponds to a local function.



learn from training data!!

factor a $f_a(x_1, y_1)$

factor b $f_b(x_2, y_1, y_2)$

factor c $f_d(y_2)$

Larger value means more compatible.

$$F(x,y) = f_a(x_1,y_1) + f_b(x_2,y_1,y_2) + f_c(y_2)$$

evaluation function即為所有factor所代表的function組合而成

You only have to define the factors.

因此其實我們只需要定義factor即可，因為只需要定義某幾個component之間的關係是比較容易的

The local functions of the factors are learned from data.

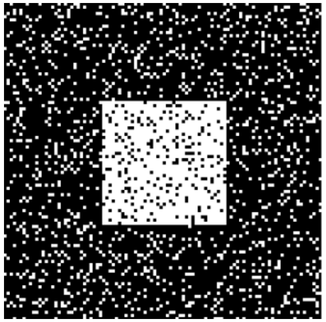
Factor Graph - Example

- ***Image De-noising***

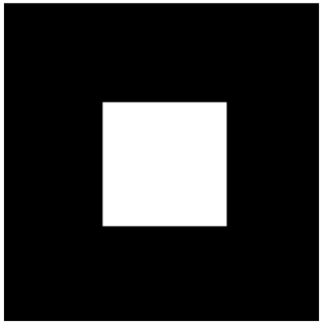
把image拆成每個pixel代表一個component

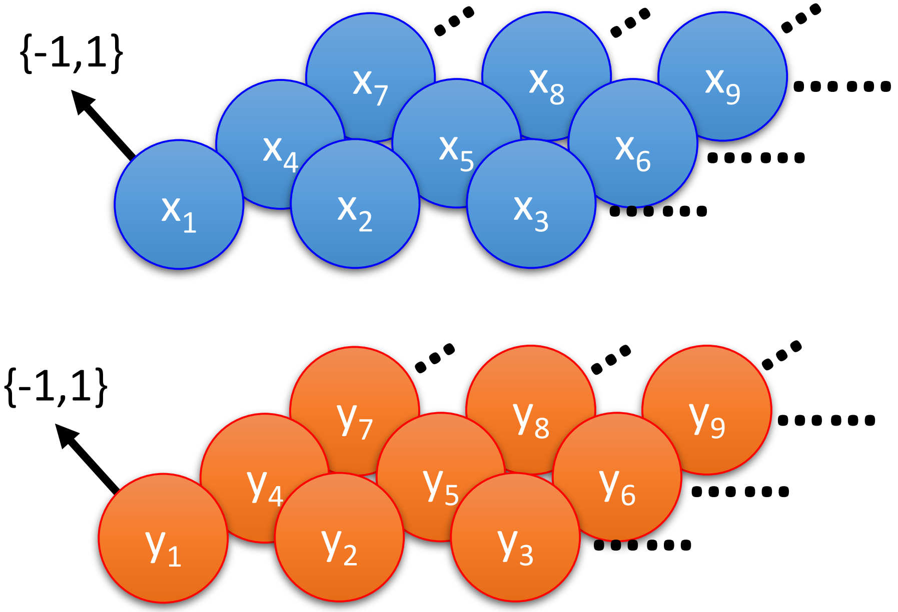
Each pixel is one component

Noisy image
x



Clean image
y





http://cs.stanford.edu/people/karpathy/visml/ising_example.html

Factor Graph - Example

Factor:

Noisy and clean images are related

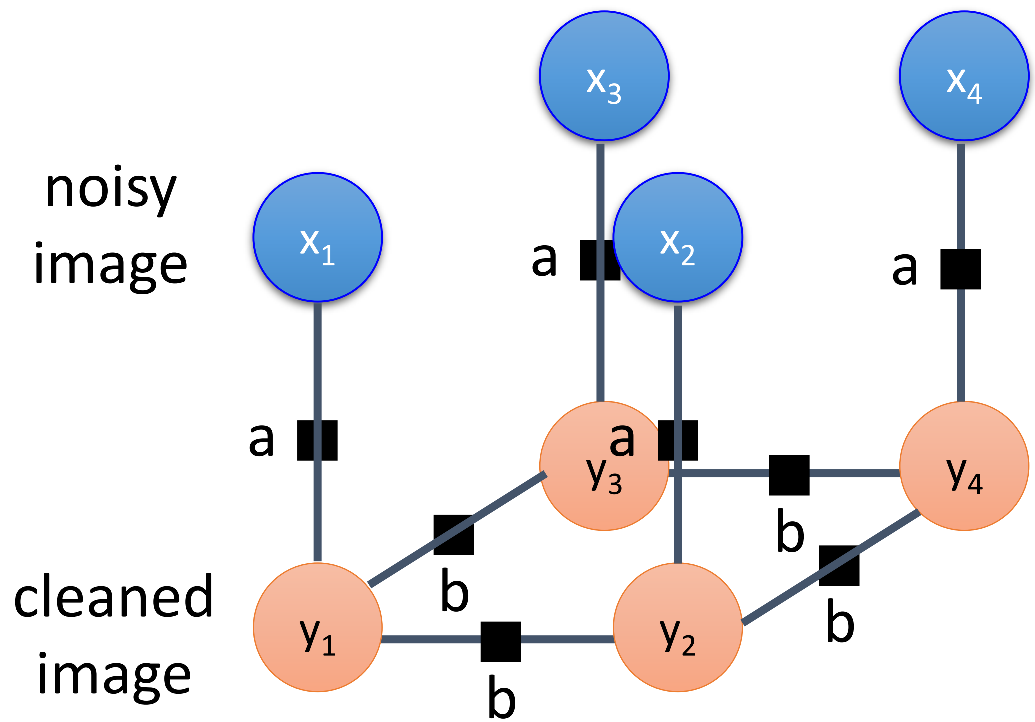
同一位置的pixel之對應

➢ **a**: the values of x_i and y_i

The colors in the clean image is smooth.

假設clean image相鄰pixel是平滑的

➢ **b**: the values of the neighboring y_i



factor/function

$$f_a(x_i, y_i) = \begin{cases} 1 & x_i = y_i \\ -1 & x_i \neq y_i \end{cases}$$

factor/function

$$f_b(y_i, y_j) = \begin{cases} 2 & y_i = y_j \\ -2 & y_i \neq y_j \end{cases}$$

The weights can be learned from data.

Factor Graph - Example

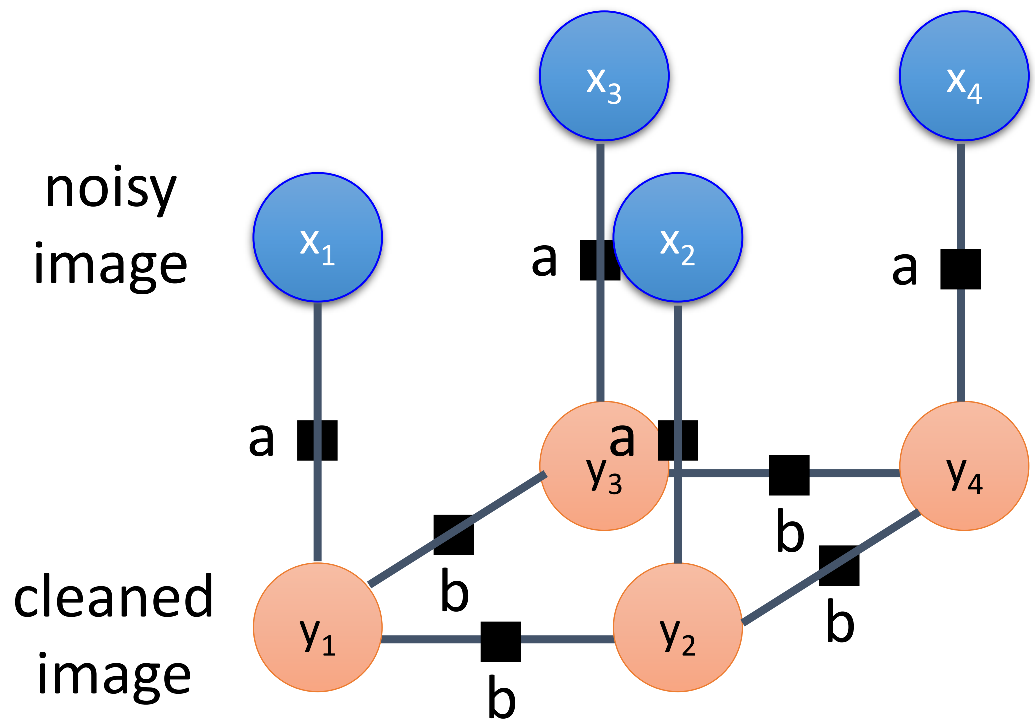
Factor:

Noisy and clean images are related

- **a**: the values of x_i and y_i

The colors in the clean image is smooth.

- **b**: the values of the neighboring y_i



Realize $F(x,y)$ easily from the factor graph

$$F(x,y) = \sum_{i=1}^{4} f_a(x_i, y_i)$$

global evaluation function

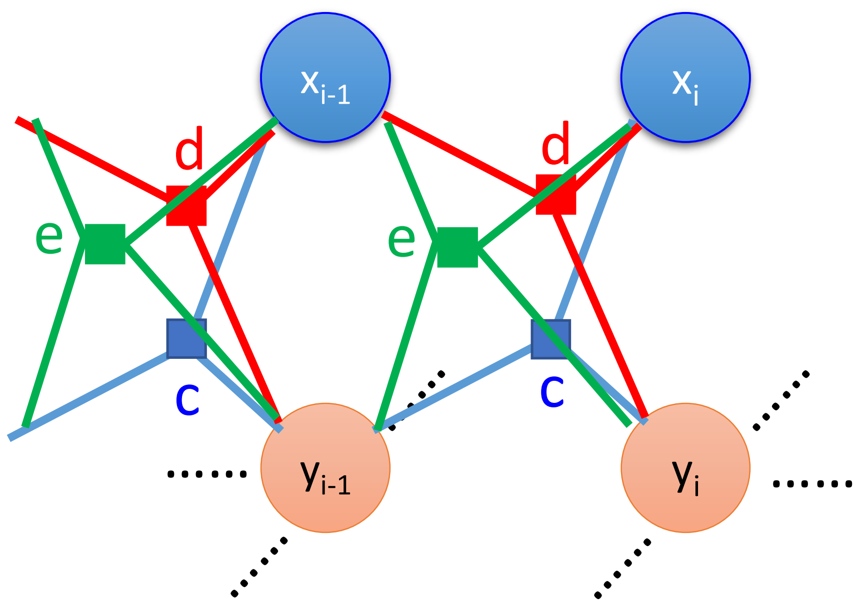
$$+f_b(x_1,y_2) + f_b(x_1,y_3)$$
$$+f_b(x_2,y_4) + f_b(x_3,y_4)$$

Factor Graph - Example

factor可以隨便亂定義，如下～

Factor:

- **c**: the values of x_i and the values of the neighboring y_i
- **d**: the values of the neighboring x_i and the values of y_i



$$f_c(x_i, y_i, y_{i-1})$$

$$f_d(x_i, x_{i-1}, y_i)$$

$$f_e(x_i, x_{i-1}, y_i, y_{i-1})$$

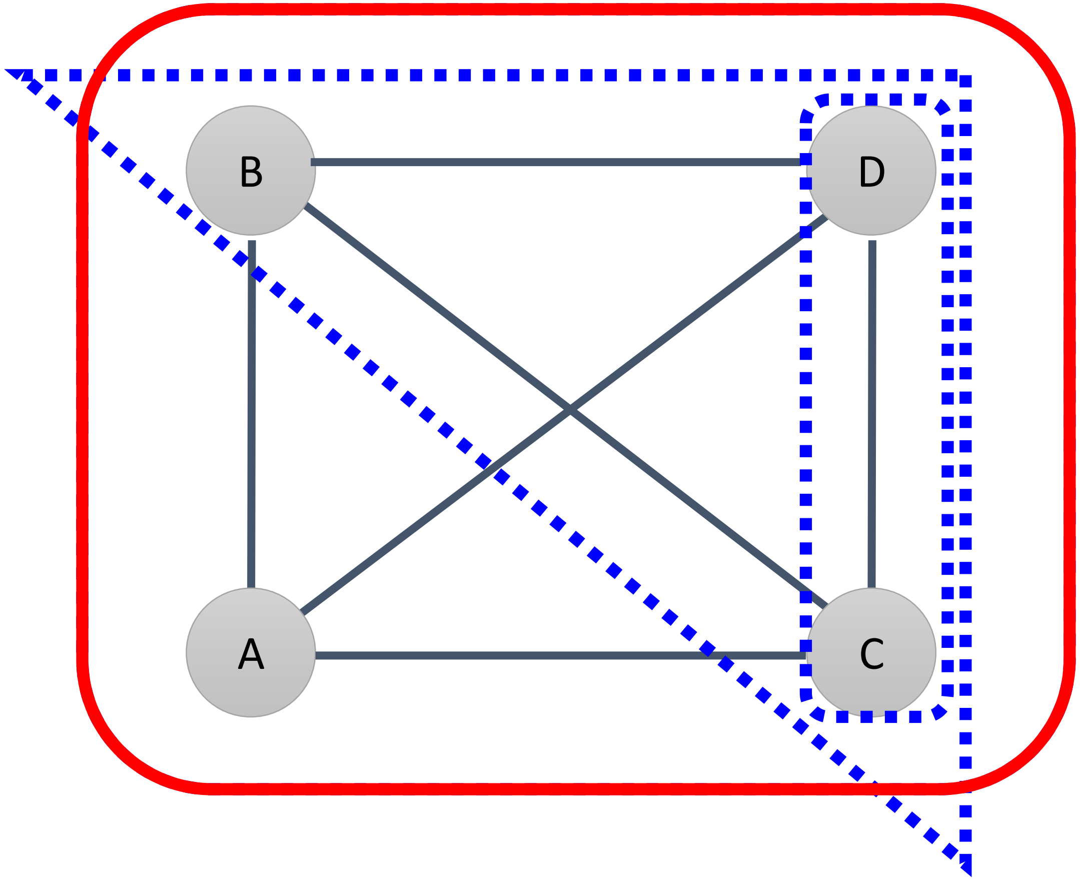
Markov Random Field (MRF)

彼此之間有連接的

Clique: a set of components connecting to each other

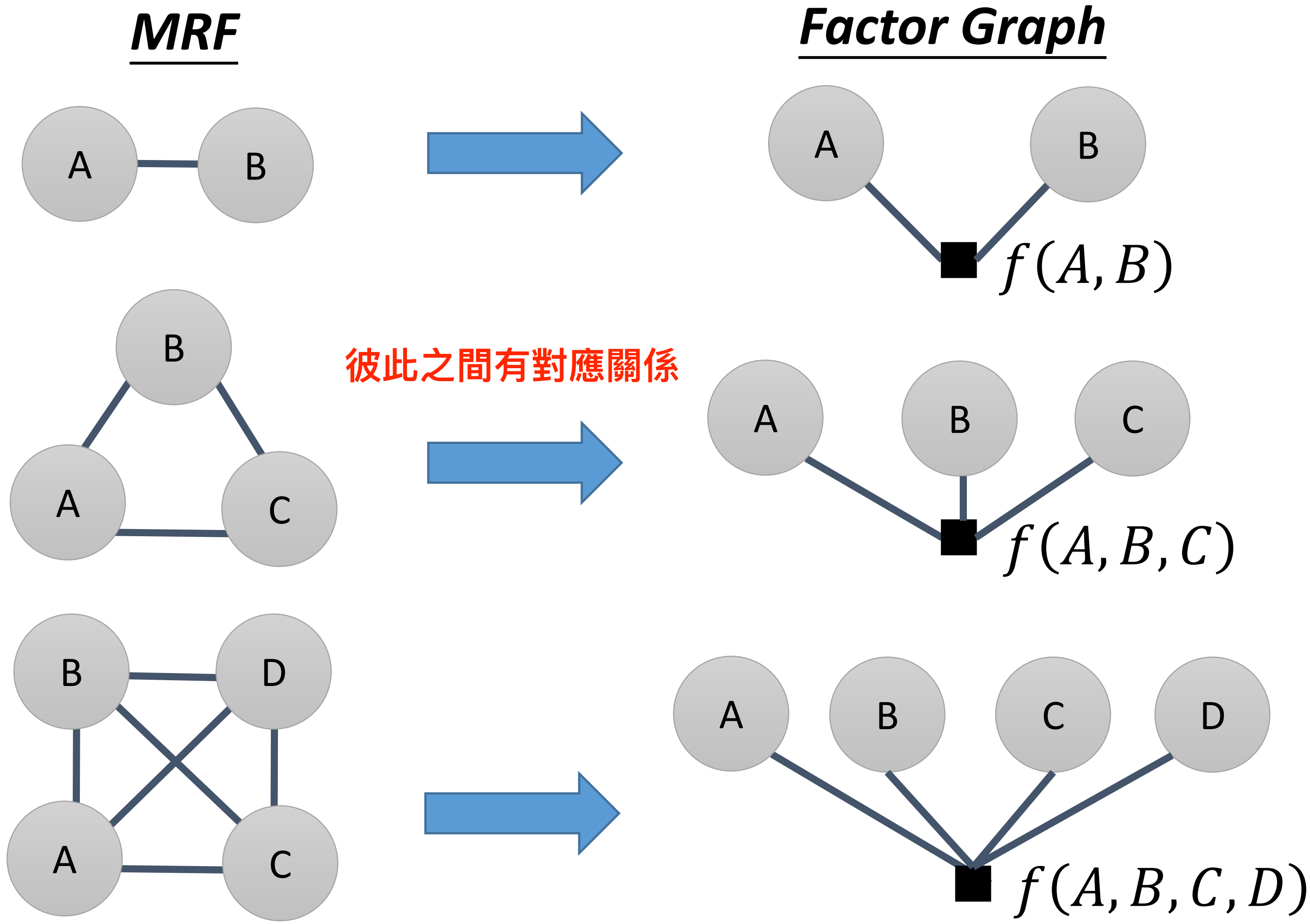
Maximum Clique: a clique that is not included by other cliques

最大的clique也不被其他clique包含



MRF

Each maximum clique on the graph corresponds to a factor



MRF

MRF → ***Factor Graph***