CSCI 5521: Introduction to Machine Learning (Spring 2020)

Local Models

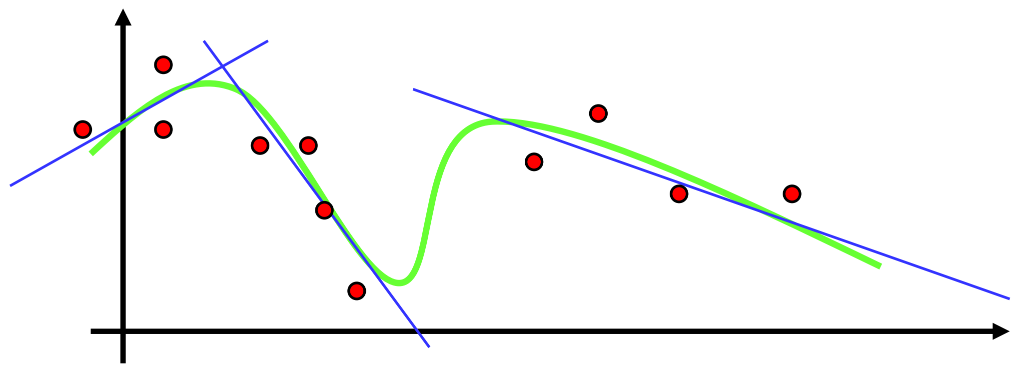
Rui Kuang

Department of Computer Science and Engineering

University of Minnesota

Introduction

- Divide the input space into local regions and learn simple (constant/linear) models in each patch

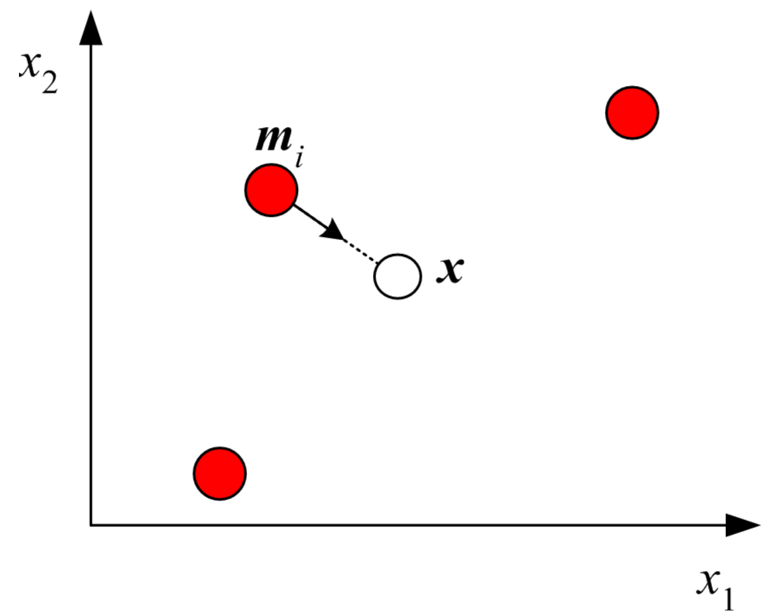


- Unsupervised: Competitive, online clustering
- Supervised: Radial-basis functions, mixture of experts

K-means Revisit

$$E\left(\{\mathbf{m}_i\}_{i=1}^{k} \middle| \mathcal{X}\right) = \sum_t \sum_i b_i^t \left\|\mathbf{x}^t - \mathbf{m}_i\right\|^2$$

$$b_i^t = \begin{cases} 1 & \text{if } \left\|\mathbf{x}^t - \mathbf{m}_i\right\| = \min_l \left\|\mathbf{x}^t - \mathbf{m}_l\right\| \\ 0 & \text{otherwise} \end{cases}$$



$$\text{Batch } k\text{-means: } \mathbf{m}_i = \frac{\sum_t b_i^t \mathbf{x}^t}{\sum_t b_i^t}$$

$$\text{Online } k\text{-means:}$$

$$\Delta m_{ij} = -\eta \frac{\partial E^t}{\partial m_{ij}} = \eta b_i^t \left(x_j^t - m_{ij}\right)$$

Online K-means

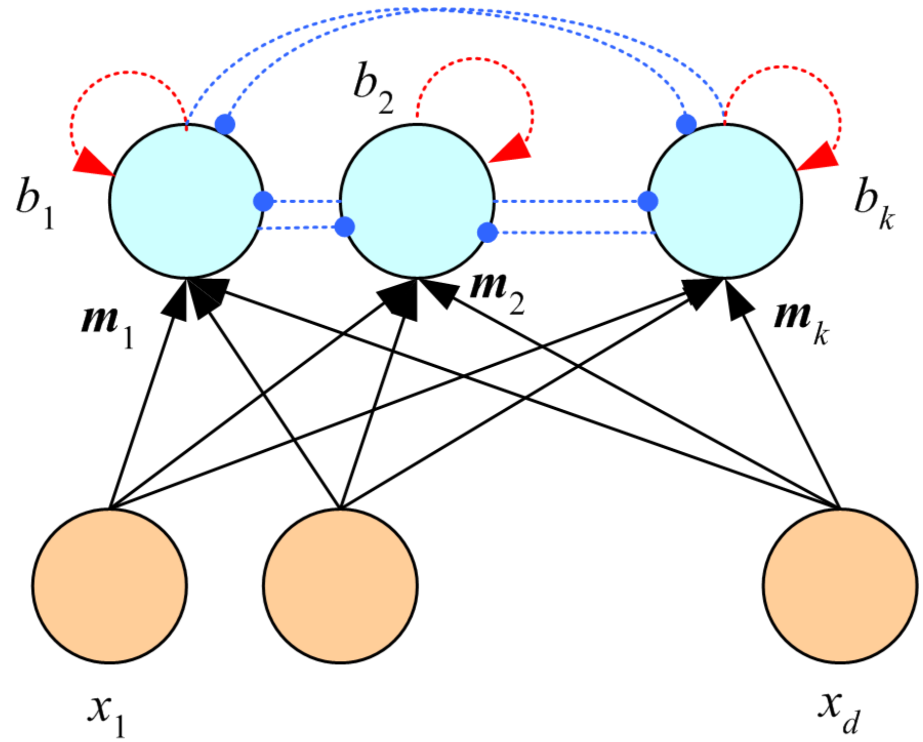
$$E^t = \sum_i b_i^t \left\| \mathbf{x}^t - \mathbf{m}_i \right\|^2$$

$$\Delta m_{ij} = -\eta \frac{\partial E^t}{\partial m_{ij}} = \eta b_i^t \left(x_j^t - m_{ij} \right)$$

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$
Repeat
    For all $\boldsymbol{x}^t \in \mathcal{X}$ in random order
        $i \leftarrow \arg\min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\|$
        $\boldsymbol{m}_i \leftarrow \boldsymbol{m}_i + \eta(\boldsymbol{x}^t - \boldsymbol{m}_i)$
Until $\boldsymbol{m}_i$ converge

Network Interpretation

Winner-take-all network

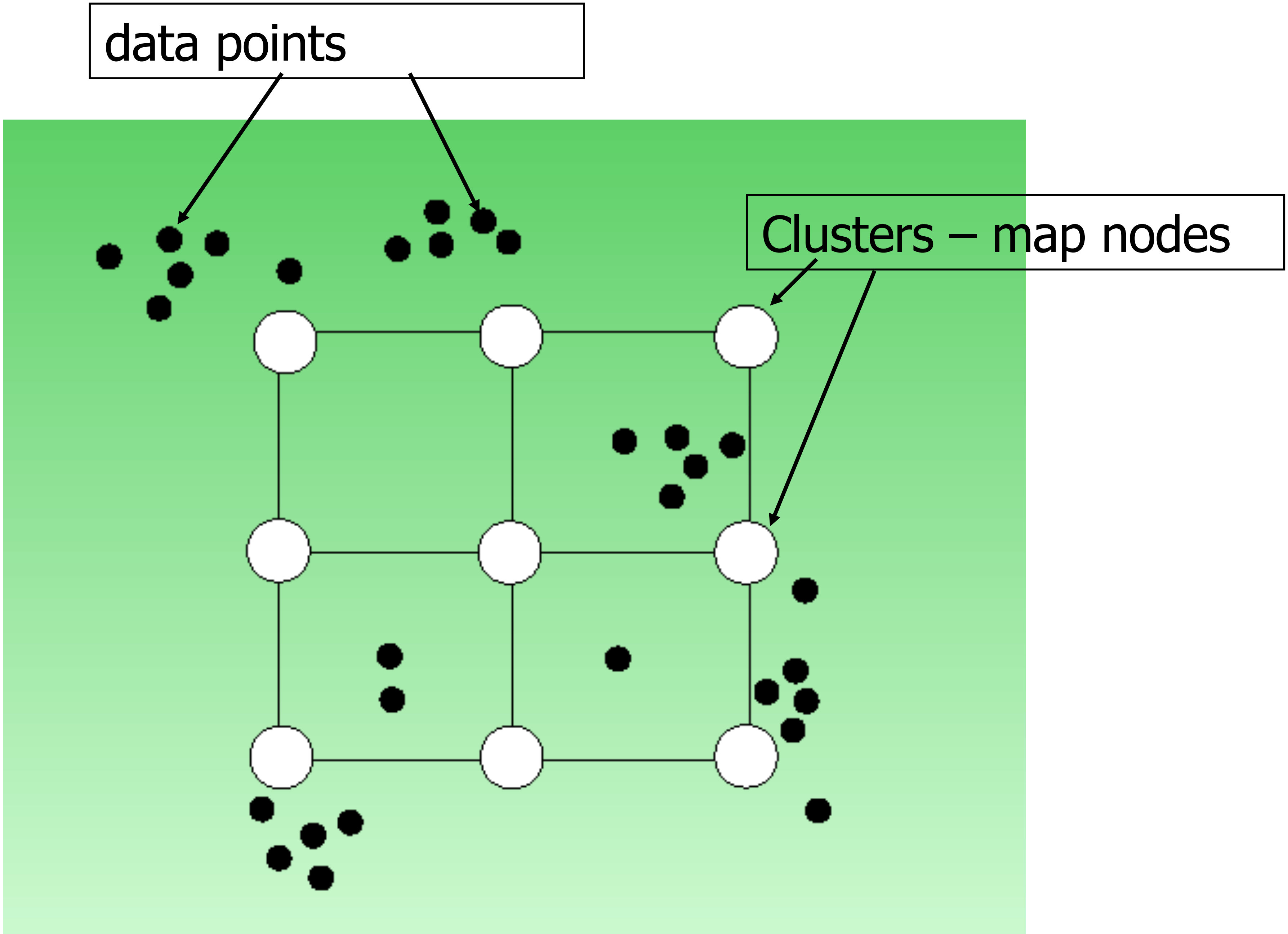


Renormalizing:

$$|\mathbf{m}_i| = 1, \forall i$$

Weight decay term:

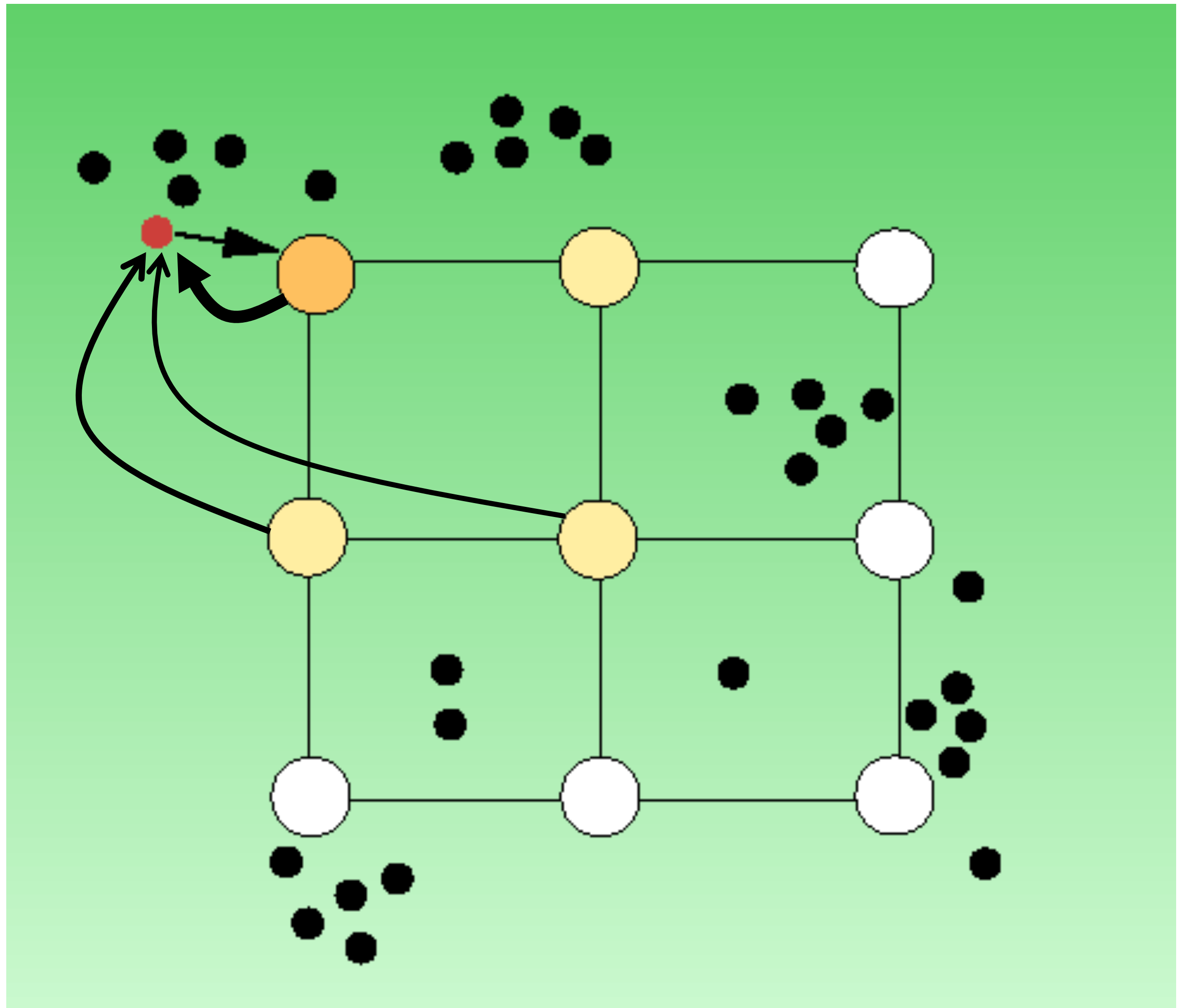
$$\Delta m_{ij} = \eta b_i^t \left(x_j^t - m_{ij} \right) = \eta b_i^t x_j^t - \eta b_i^t m_{ij}$$



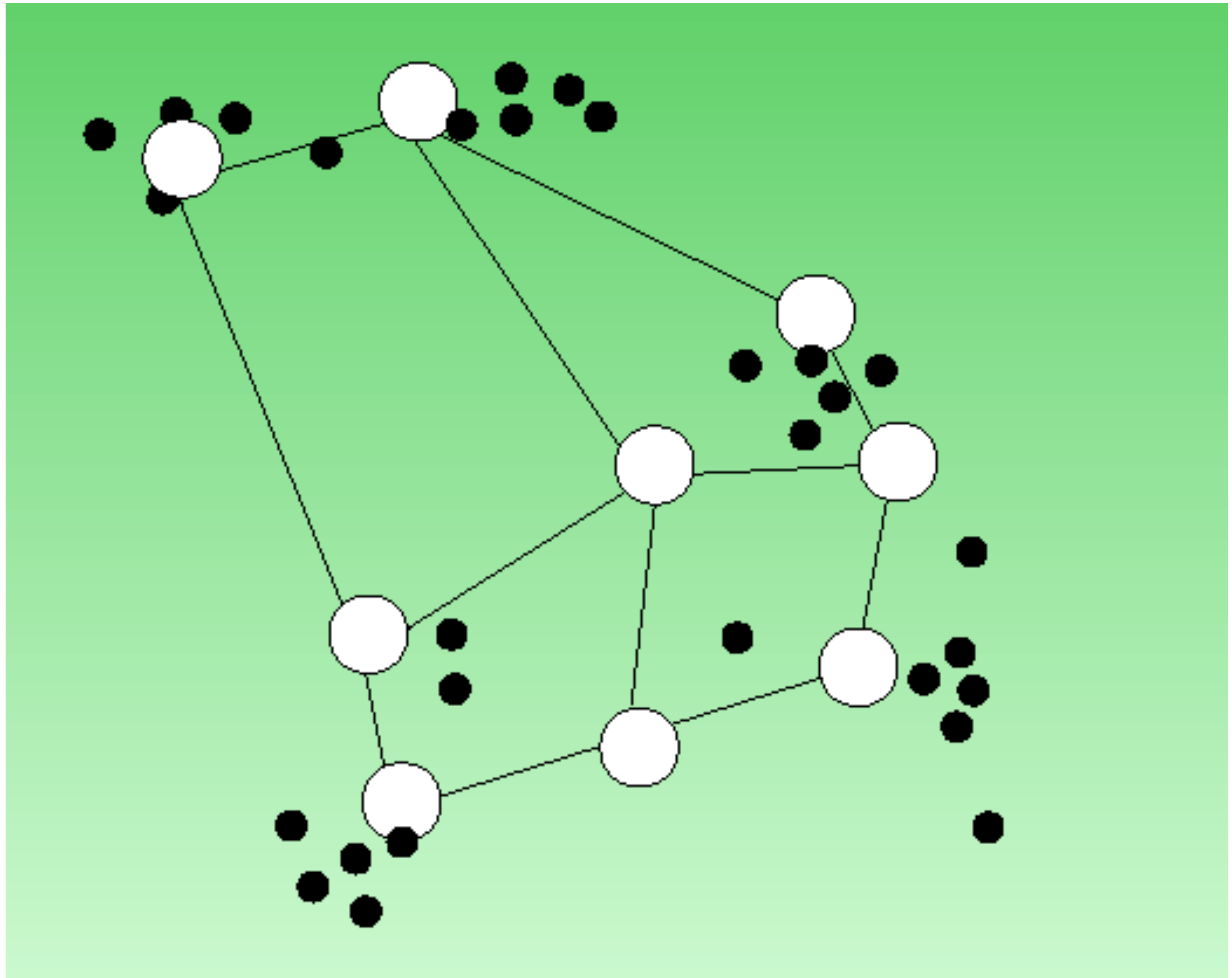
www.tau.ac.il/lifesci/bioinfo/teaching/2005-2006/**clustering**-Jan06.**ppt**

SOM - Scheme

- Randomly choose a data point.
- Find its closest map node
- Move this map node towards the data point
- Move the neighbor map nodes towards this point, but to lesser extent
- Iterate over data points



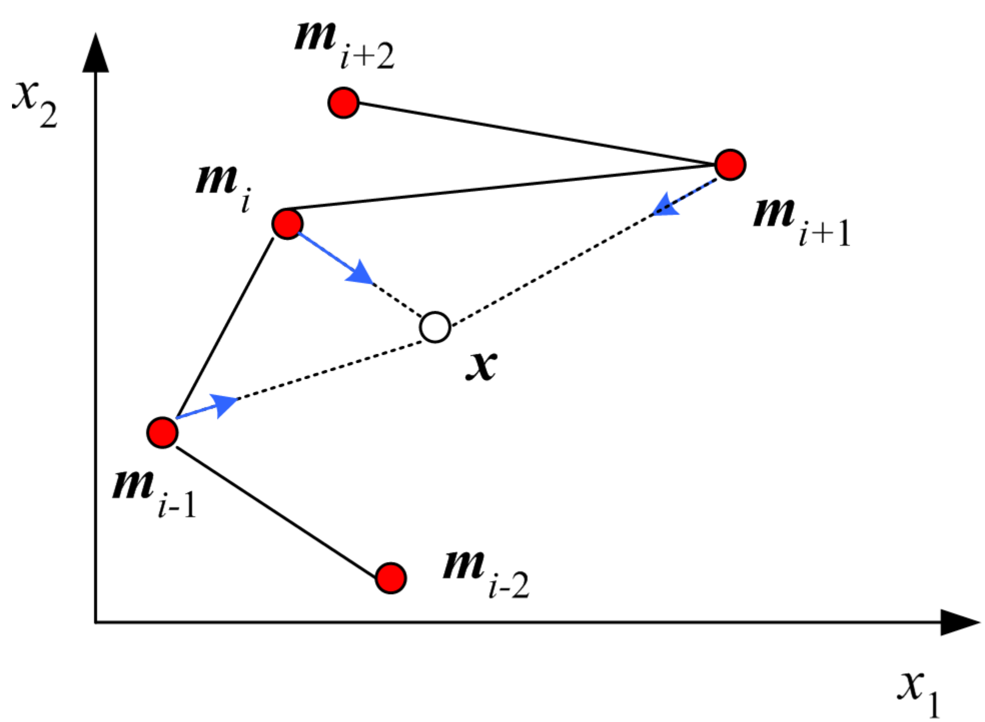
- The extent of node displacements is relaxed with the iteration number
- After thousands of iterations:
- Assign each data point to the map node (cluster) it is most similar to



www.tau.ac.il/lifesci/bioinfo/teaching/2005-2006/**clustering**-Jan06.**ppt**

Self-Organizing Maps

- Units have a neighborhood defined; $\boldsymbol{m}_i$ is "between" $\boldsymbol{m}_{i-1}$ and $\boldsymbol{m}_{i+1}$, and are all updated together



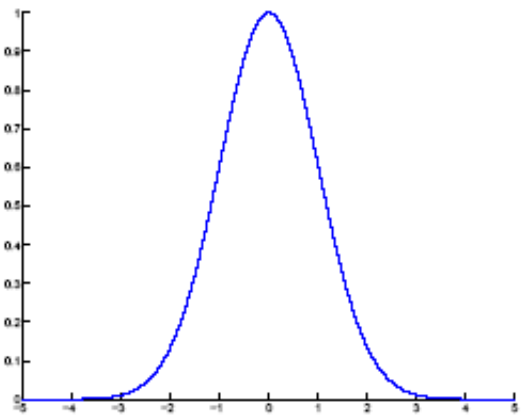
$$\Delta \mathbf{m}_l = \eta e(l,i)\left(\mathbf{x}^t - \mathbf{m}_l\right)$$

$$e(l,i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(l-i)^2}{2\sigma^2}\right]$$

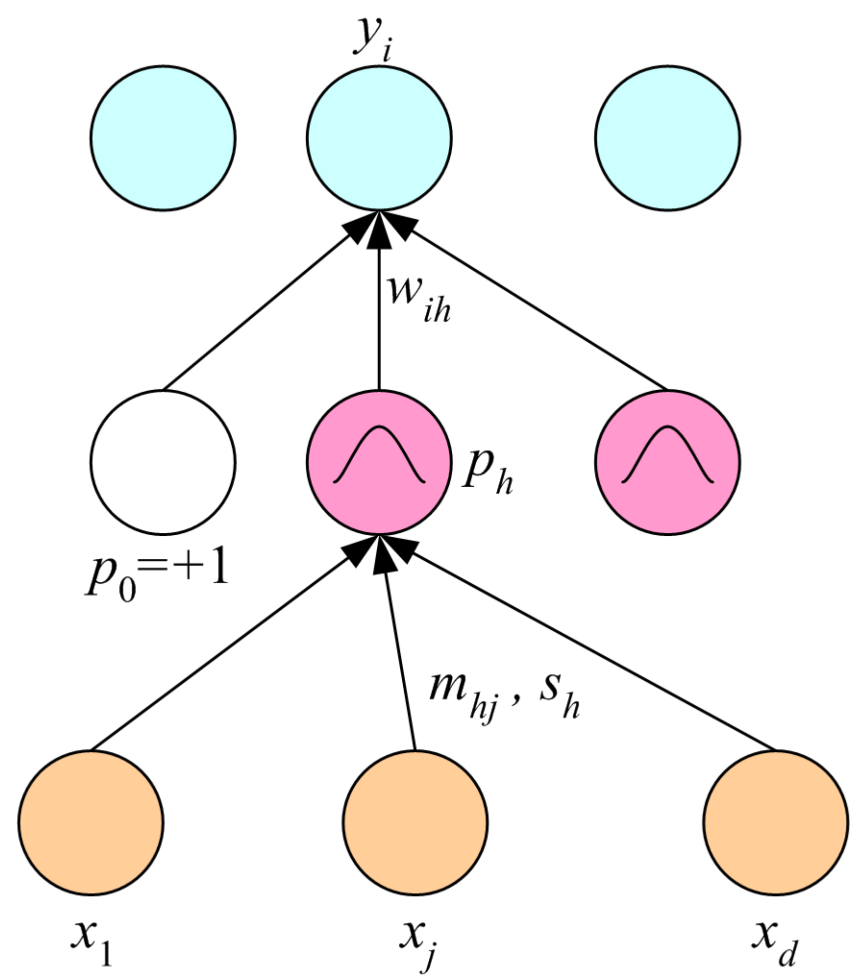
Radial-Basis Functions

- Locally-tuned units:

$$p_h^t = \exp\left[-\frac{\left\|\mathbf{x}^t - \mathbf{m}_h\right\|^2}{2s_h^2}\right]$$



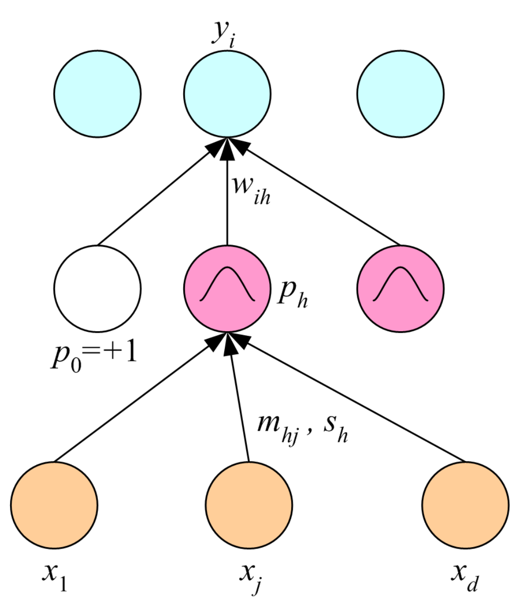
$$y^t = \sum_{h=1}^{H} w_h p_h^t + w_0$$

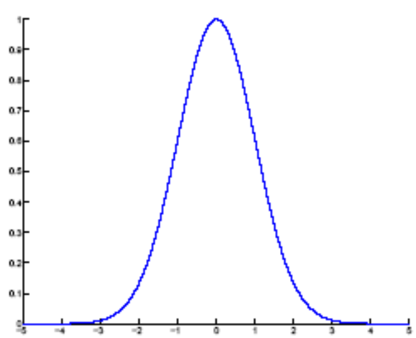


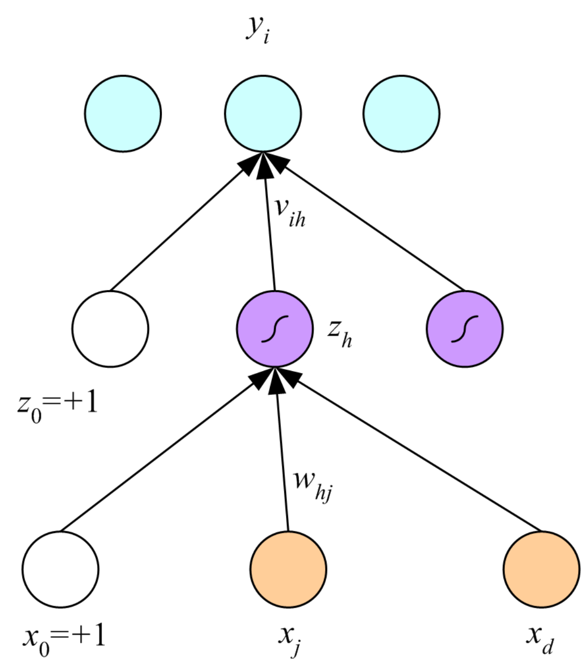
E. Alpaydin, Introduction to Machine Learning

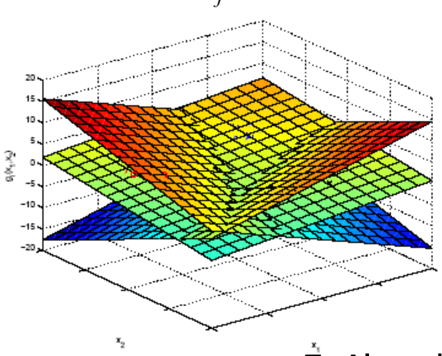
Radial-Basis vs Linear Functions

- What does the hidden layer do?



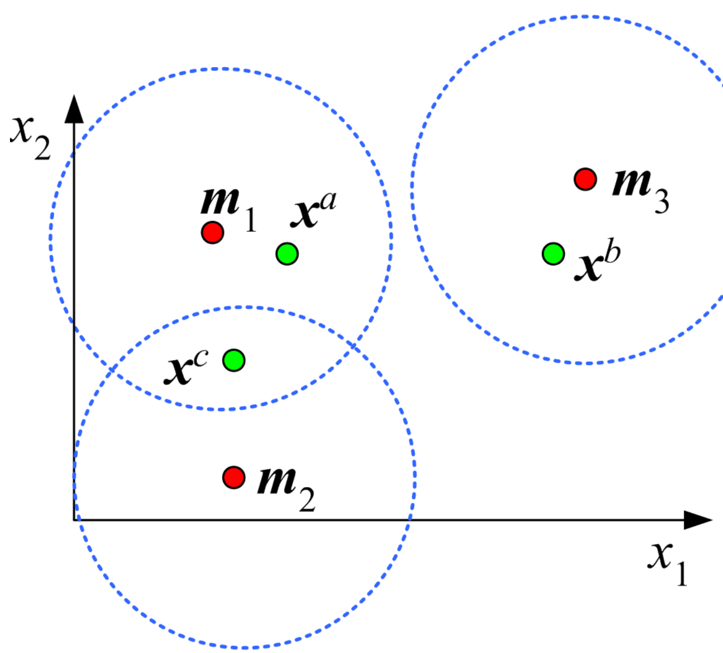






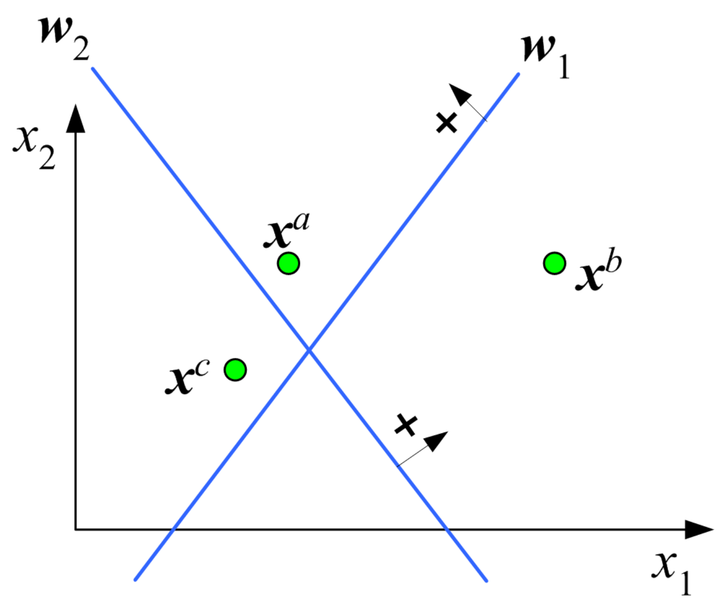
E. Alpaydin, Introduction to Machine Learning

Local vs Distributed Representation



Local representation in the space of (p_1, p_2, p_3)

$\boldsymbol{x}^a$: (1.0, 0.0, 0.0)
$\boldsymbol{x}^b$: (0.0, 0.0, 1.0)
$\boldsymbol{x}^c$: (1.0, 1.0, 0.0)



Distributed representation in the space of (h_1, h_2)

$\boldsymbol{x}^a$: (1.0, 1.0)
$\boldsymbol{x}^b$: (0.0, 1.0)
$\boldsymbol{x}^c$: (1.0, 0.0)

Regression

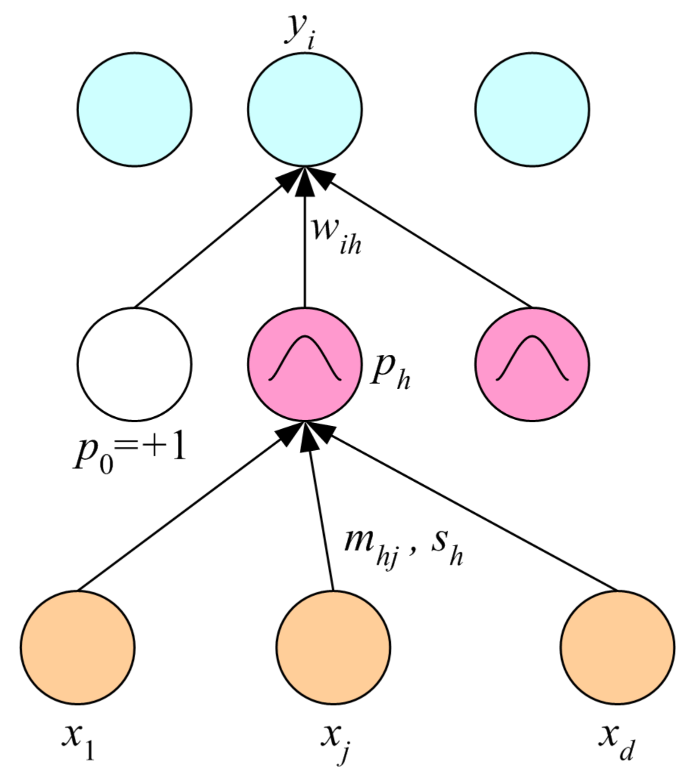
$$E\left(\left\{\mathbf{m}_h, s_h, w_{ih}\right\}_{i,h} \mid \mathcal{X}\right) = \frac{1}{2}\sum_t \sum_i \left(r_i^t - y_i^t\right)^2$$

$$y_i^t = \sum_{h=1}^{H} w_{ih} p_h^t + w_{i0}$$

$$\Delta w_{ih} = \eta \sum_t \left(r_i^t - y_i^t\right) p_h^t$$

$$\Delta m_{hj} = \eta \sum_t \left[\sum_i \left(r_i^t - y_i^t\right) w_{ih}\right] p_h^t \frac{\left(x_j^t - m_{hj}\right)}{s_h^2}$$

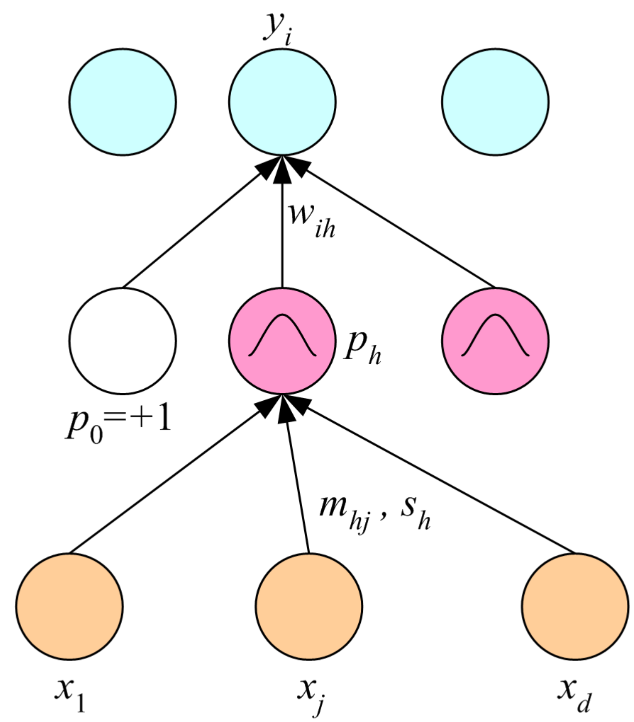
$$\Delta s_h = \eta \sum_t \left[\sum_i \left(r_i^t - y_i^t\right) w_{ih}\right] p_h^t \frac{\left\|\mathbf{x}^t - \mathbf{m}_h\right\|^2}{s_h^3}$$



Classification

$$E\left(\left\{\mathbf{m}_h, s_h, w_{ih}\right\}_{i,h} \mid \mathcal{X}\right) = -\sum_t \sum_i r_i^t \log y_i^t$$

$$y_i^t = \frac{\exp\left[\sum_h w_{ih} p_h^t + w_{i0}\right]}{\sum_k \exp\left[\sum_h w_{kh} p_h^t + w_{k0}\right]}$$



The updates are the same as the regression problem.