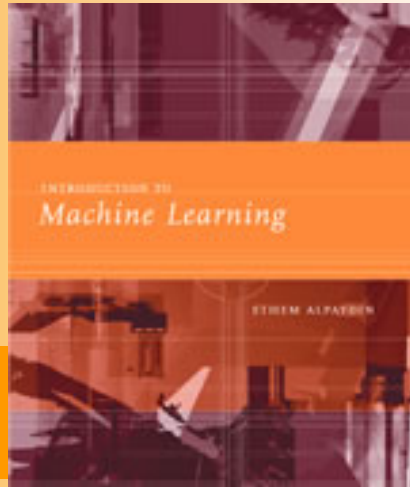




Lecture Slides for

INTRODUCTION TO

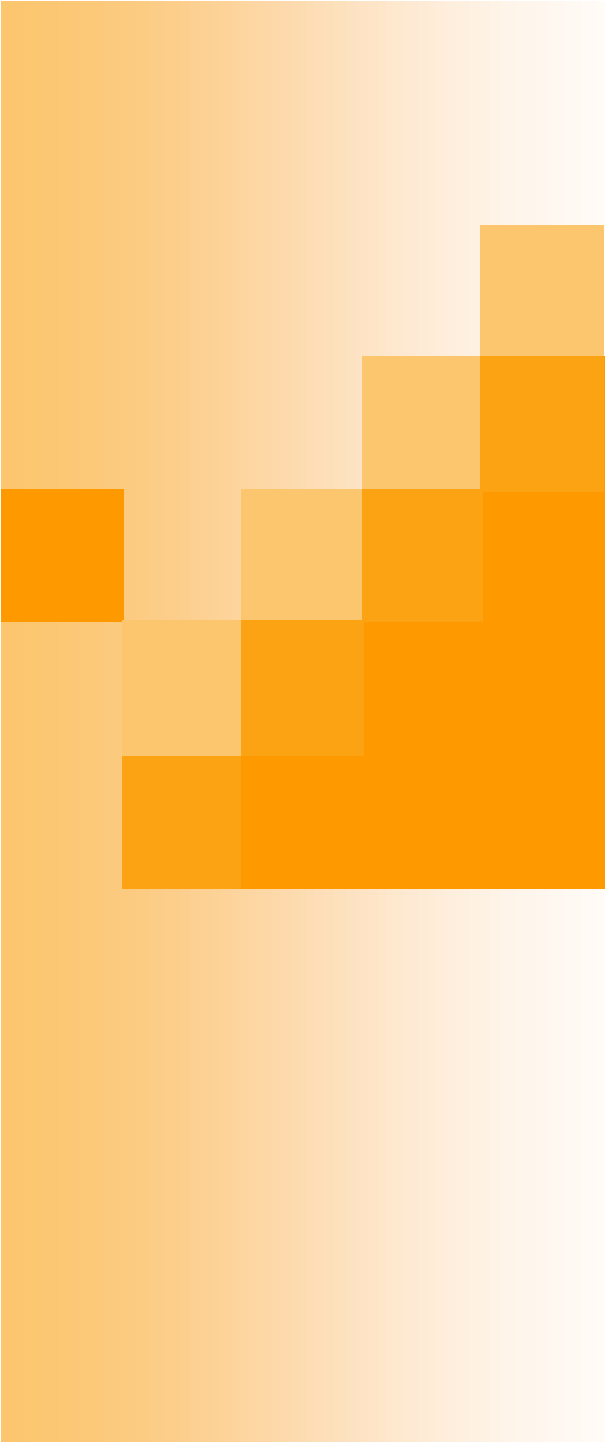
Machine Learning

ETHEM ALPAYDIN

© The MIT Press, 2004

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml>

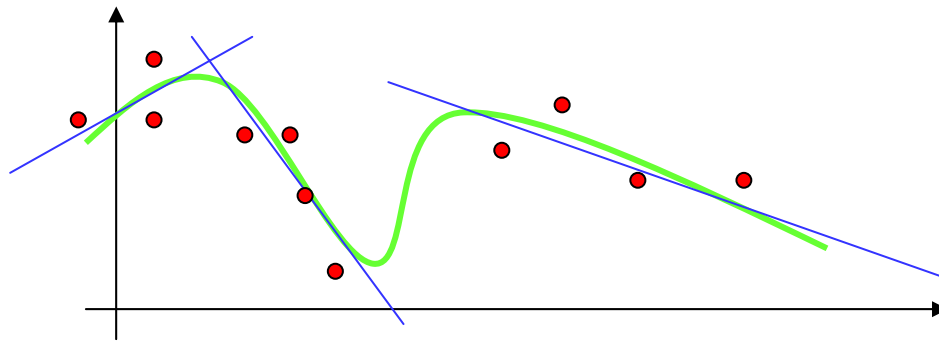
A decorative graphic on the left side of the slide consisting of a grid of orange squares of varying shades, arranged in a stepped pattern that tapers towards the top.

CHAPTER 12:

Local Models

Introduction

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



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

Competitive Learning

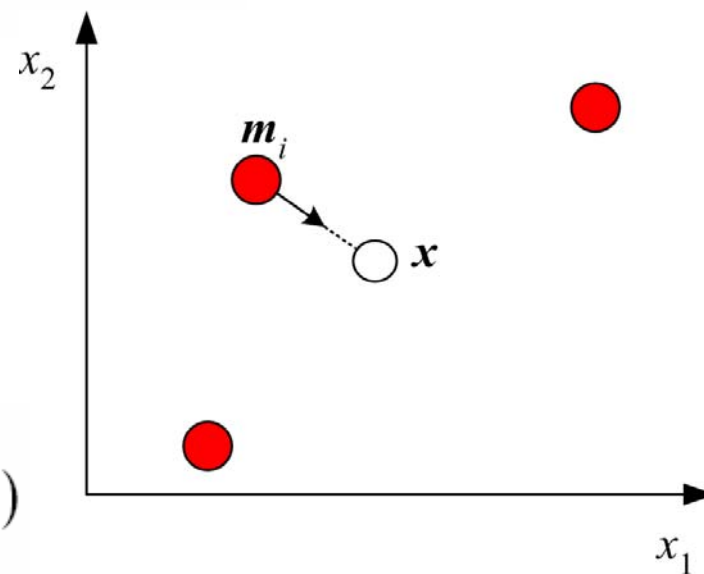
$$E(\{\mathbf{m}_i\}_{i=1}^k | \mathcal{X}) = \frac{1}{2} \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|^2$$

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

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

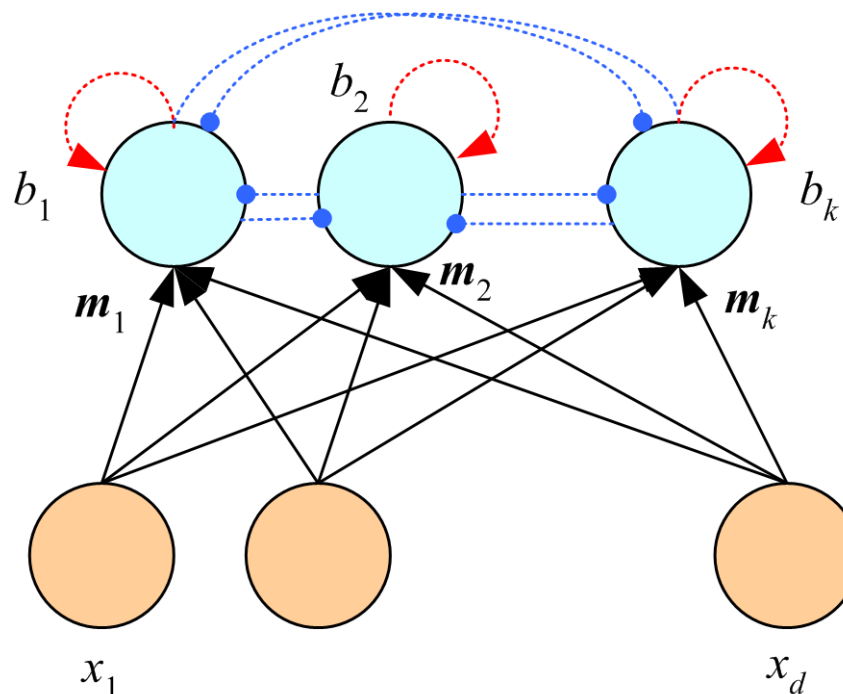
- Online k -means:

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



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

*Winner-take-all
network*

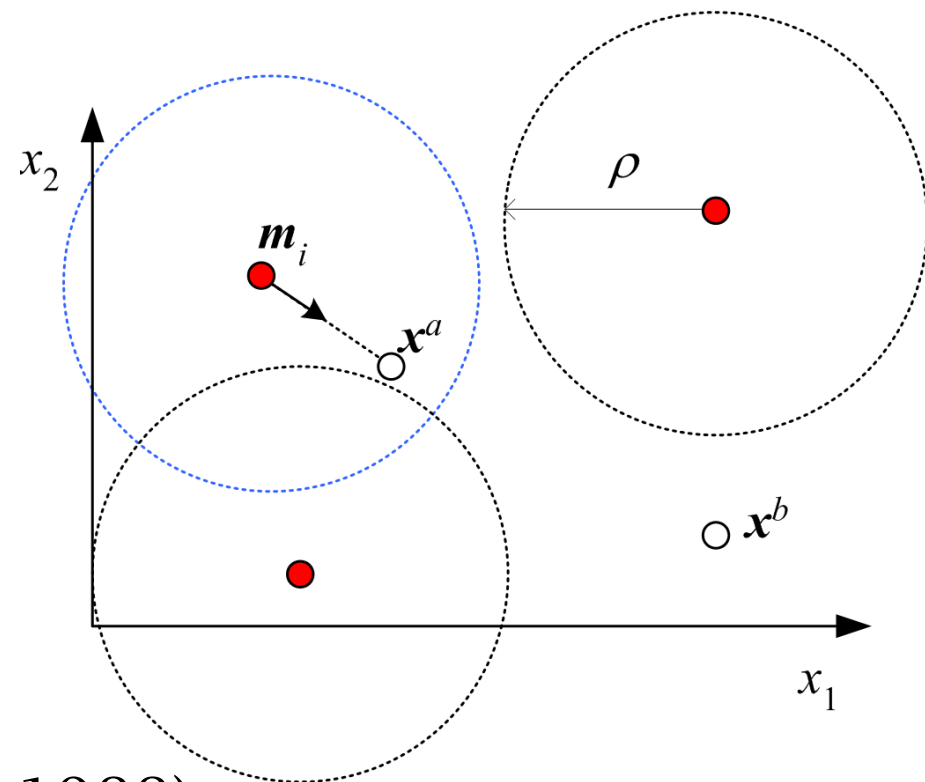


Adaptive Resonance Theory

- Incremental; add a new cluster if not covered; defined by **vigilance**, ρ

$$b_i = \|\mathbf{m}_i - \mathbf{x}^t\| = \min_{l=1}^k \|\mathbf{m}_l - \mathbf{x}^t\|$$

$$\begin{cases} \mathbf{m}_{k+1} \leftarrow \mathbf{x}^t & \text{if } b_i > \rho \\ \Delta \mathbf{m}_i = \eta(\mathbf{x} - \mathbf{m}_i) & \text{otherwise} \end{cases}$$

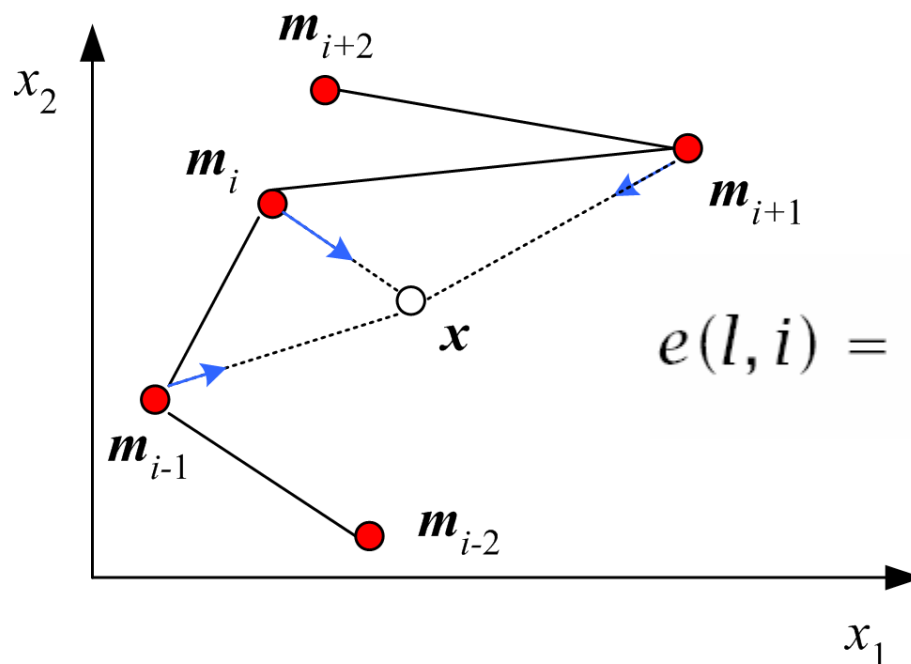


(Carpenter and Grossberg, 1988)

Self-Organizing Maps

- Units have a **neighborhood** defined; $\mathbf{m}_i$ is “between” $\mathbf{m}_{i-1}$ and $\mathbf{m}_{i+1}$, and are all updated together
- One-dim map: $\Delta \mathbf{m}_l = \eta e(l, i) (\mathbf{x}^t - \mathbf{m}_l)$

(Kohonen, 1990)

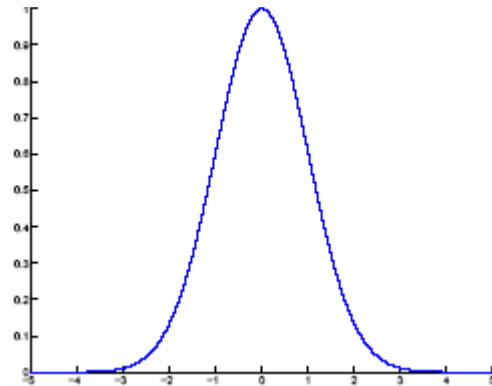


$$e(l, i) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(l-i)^2}{2\sigma^2} \right]$$

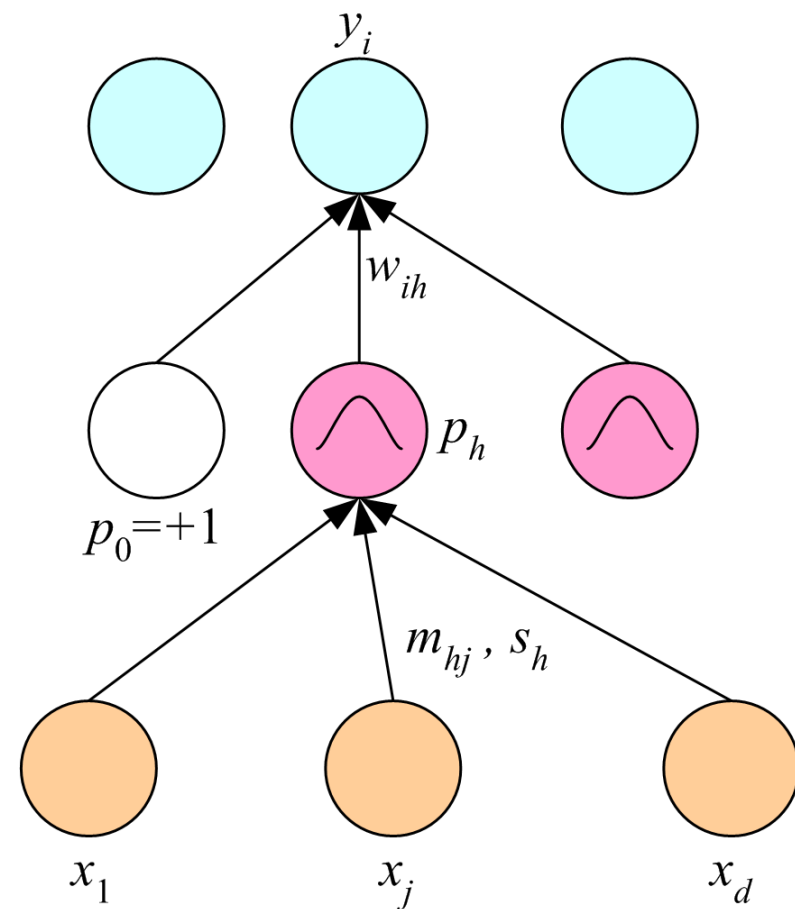
Radial-Basis Functions

- Locally-tuned units:

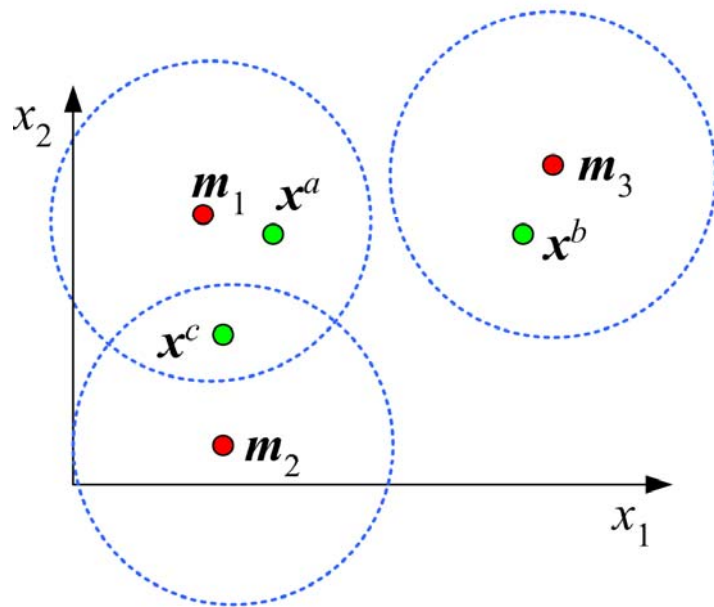
$$p_h^t = \exp \left[-\frac{\|\mathbf{x}^t - \mathbf{m}_h\|^2}{2s_h^2} \right]$$



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



Local vs Distributed Representation

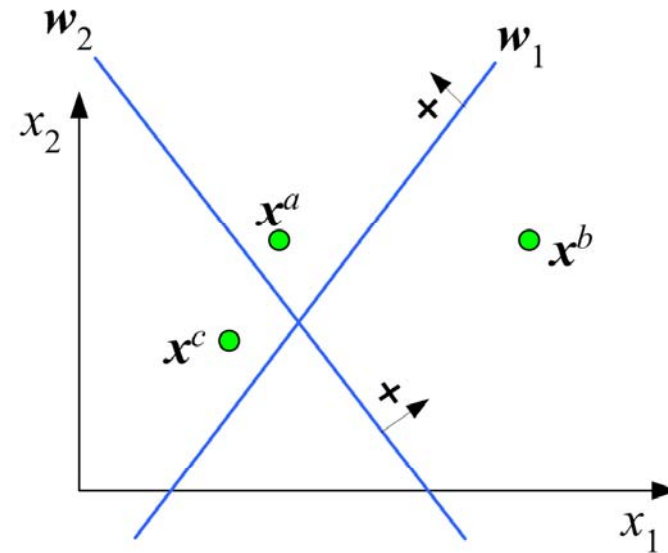


Local representation in the space of (p_1, p_2, p_3)

$x^a : (1.0, 0.0, 0.0)$

$x^b : (0.0, 0.0, 1.0)$

$x^c : (1.0, 1.0, 0.0)$



Distributed representation in the space of (h_1, h_2)

$x^a : (1.0, 1.0)$

$x^b : (0.0, 1.0)$

$x^c : (1.0, 0.0)$



Training RBF

- Hybrid learning:
 - First layer centers and spreads:
Unsupervised k -means
 - Second layer weights:
Supervised gradient-descent
- Fully supervised
- (Broomhead and Lowe, 1988; Moody and Darken, 1989)

Regression

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

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

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

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

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

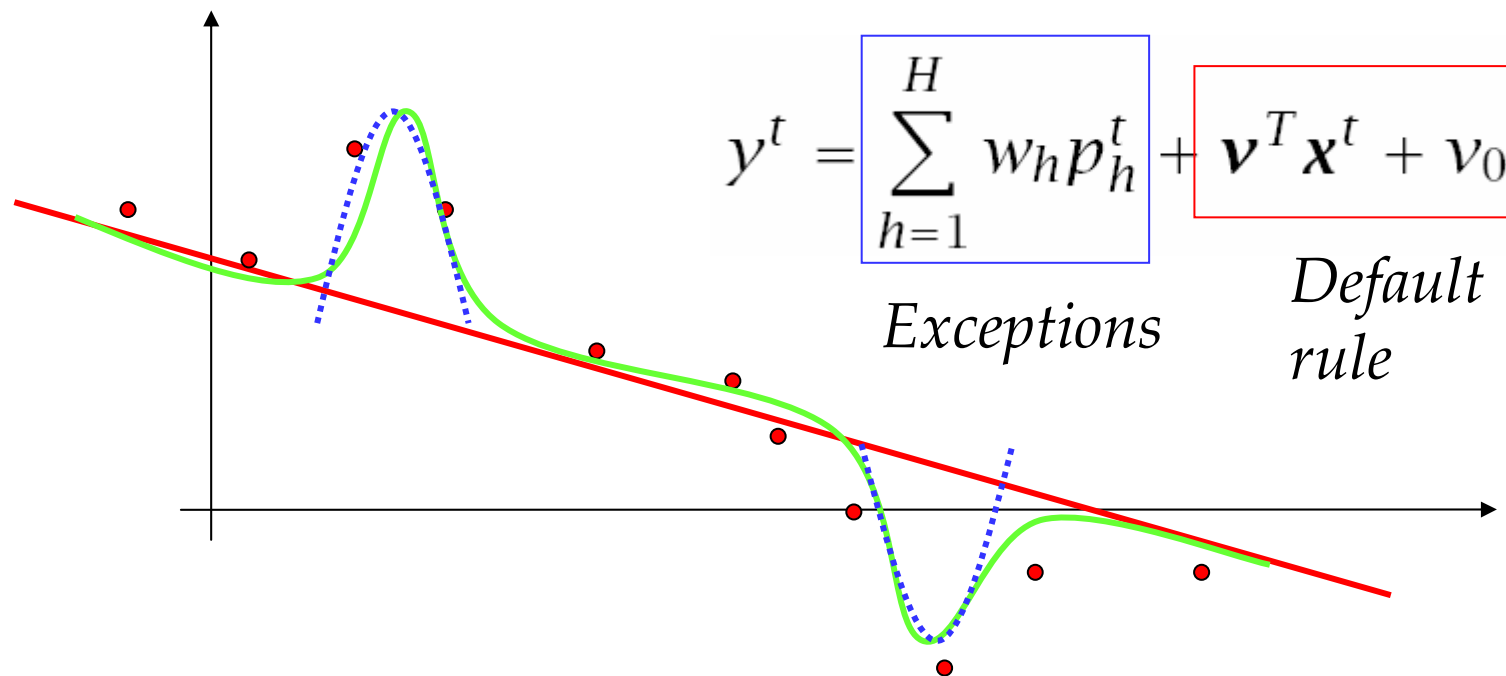


Classification

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

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

Rules and Exceptions





Rule-Based Knowledge

IF $((x_1 \approx a) \text{ AND } (x_2 \approx b)) \text{ OR } (x_3 \approx c)$ THEN $y = 0.1$

$$p_1 = \exp \left[-\frac{(x_1 - a)^2}{2s_1^2} \right] \cdot \exp \left[-\frac{(x_2 - b)^2}{2s_2^2} \right] \text{ with } w_1 = 0.1$$

$$p_2 = \exp \left[-\frac{(x_3 - c)^2}{2s_3^2} \right] \text{ with } w_2 = 0.1$$

- Incorporation of prior knowledge (before training)
- Rule extraction (after training) (Tresp et al., 1997)
- Fuzzy membership functions and fuzzy rules

Normalized Basis Functions

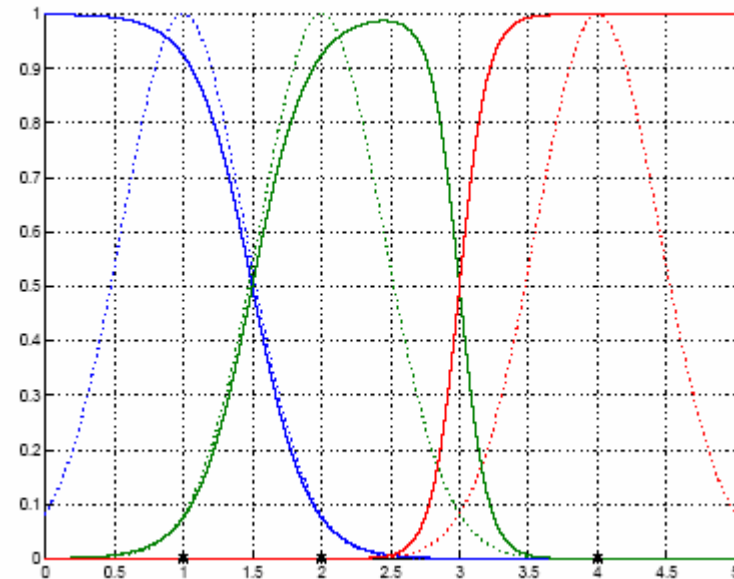
$$g_h^t = \frac{p_h^t}{\sum_{l=1}^H p_l^t}$$

$$= \frac{\exp[-\|\mathbf{x}^t - \mathbf{m}_h\|^2 / 2s_h^2]}{\sum_l \exp[-\|\mathbf{x}^t - \mathbf{m}_l\|^2 / 2s_l^2]}$$

$$y_i^t = \sum_{h=1}^H w_{ih} g_h^t$$

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

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





Competitive Basis Functions

■ Mixture model:
$$p(\mathbf{r}^t | \mathbf{x}^t) = \sum_{h=1}^H p(h | \mathbf{x}^t) p(\mathbf{r}^t | h, \mathbf{x}^t)$$

$$p(h | \mathbf{x}) = \frac{p(\mathbf{x} | h) p(h)}{\sum_l p(\mathbf{x} | l) p(l)}$$

$$g_h^t = \frac{a_h \exp[-\|\mathbf{x}^t - \mathbf{m}_h\|^2 / 2s_h^2]}{\sum_l a_l \exp[-\|\mathbf{x}^t - \mathbf{m}_l\|^2 / 2s_l^2]}$$



Regression $p(\mathbf{r}^t | \mathbf{x}^t) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(r_i^t - y_i^t)^2}{2\sigma^2} \right]$

$$\mathcal{L}(\{\mathbf{m}_h, s_h, w_{ih}\}_{i,h} | \mathcal{X}) = \sum_t \log \sum_h g_h^t \exp \left[-\frac{1}{2} \sum_i (r_i^t - y_{ih}^t)^2 \right]$$

$y_{ih}^t = w_{ih}$ is the constant fit

$$\Delta w_{ih} = \eta \sum_t (r_i^t - y_{ih}^t) f_h^t \quad \Delta m_{hj} = \eta \sum_t (f_h^t - g_h^t) \frac{(x_j^t - m_{hj})}{s_h^2}$$

$$f_h^t = \frac{g_h^t \exp[-\frac{1}{2} \sum_i (r_i^t - y_{ih}^t)^2]}{\sum_l g_l^t \exp[-\frac{1}{2} \sum_i (r_i^t - y_{il}^t)^2]}$$

$$p(h | \mathbf{r}, \mathbf{x}) = \frac{p(h | \mathbf{x}) p(\mathbf{r} | h, \mathbf{x})}{\sum_l p(l | \mathbf{x}) p(\mathbf{r} | l, \mathbf{x})}$$

Classification

$$\begin{aligned}\mathcal{L}(\{\mathbf{m}_h, s_h, w_{ih}\}_{i,h}|\mathcal{X}) &= \sum_t \log \sum_h g_h^t \prod_i (y_{ih}^t)^{r_i^t} \\ &= \sum_t \log \sum_h g_h^t \exp \left[\sum_i r_i^t \log y_{ih}^t \right]\end{aligned}$$

$$y_{ih}^t = \frac{\exp w_{ih}}{\sum_k \exp w_{kh}}$$

$$f_h^t = \frac{g_h^t \exp[\sum_i r_i^t \log y_{ih}^t]}{\sum_l g_l^t \exp[\sum_i r_i^t \log y_{il}^t]}$$

EM for RBF (Supervised EM)

■ E-step: $f_h^t \equiv p(\mathbf{r}|h, \mathbf{x}^t)$

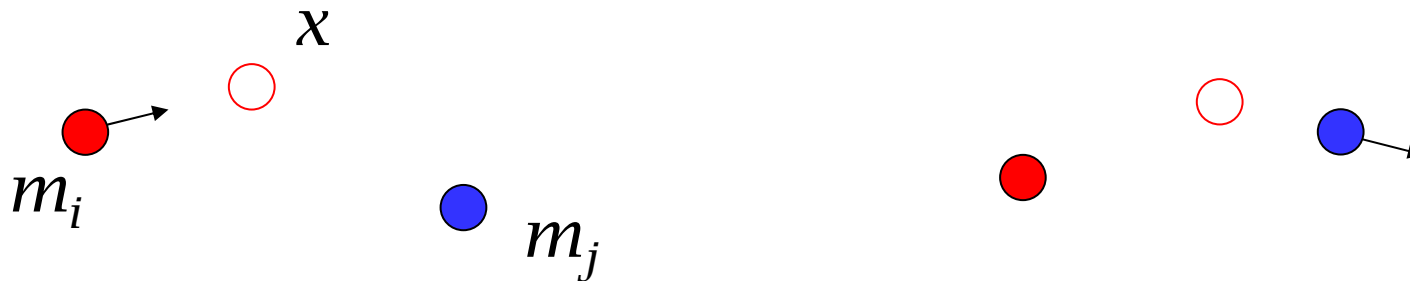
■ M-step:

$$\begin{aligned}\mathbf{m}_h &= \frac{\sum_t f_h^t \mathbf{x}^t}{\sum_t f_h^t} \\ \mathbf{S}_h &= \frac{\sum_t f_h^t (\mathbf{x}^t - \mathbf{m}_h)(\mathbf{x}^t - \mathbf{m}_h)^T}{\sum_t f_h^t} \\ w_{ih} &= \frac{\sum_t f_h^t r_i^t}{\sum_t f_h^t}\end{aligned}$$

Learning Vector Quantization

- H units per class prelabeled (Kohonen, 1990)
- Given $\mathbf{x}$, $\mathbf{m}_i$ is the closest:

$$\begin{cases} \Delta \mathbf{m}_i = \eta (\mathbf{x}^t - \mathbf{m}_i) & \text{if } \mathbf{x}^t \text{ and } \mathbf{m}_i \text{ have the same class label} \\ \Delta \mathbf{m}_i = -\eta (\mathbf{x}^t - \mathbf{m}_i) & \text{otherwise} \end{cases}$$

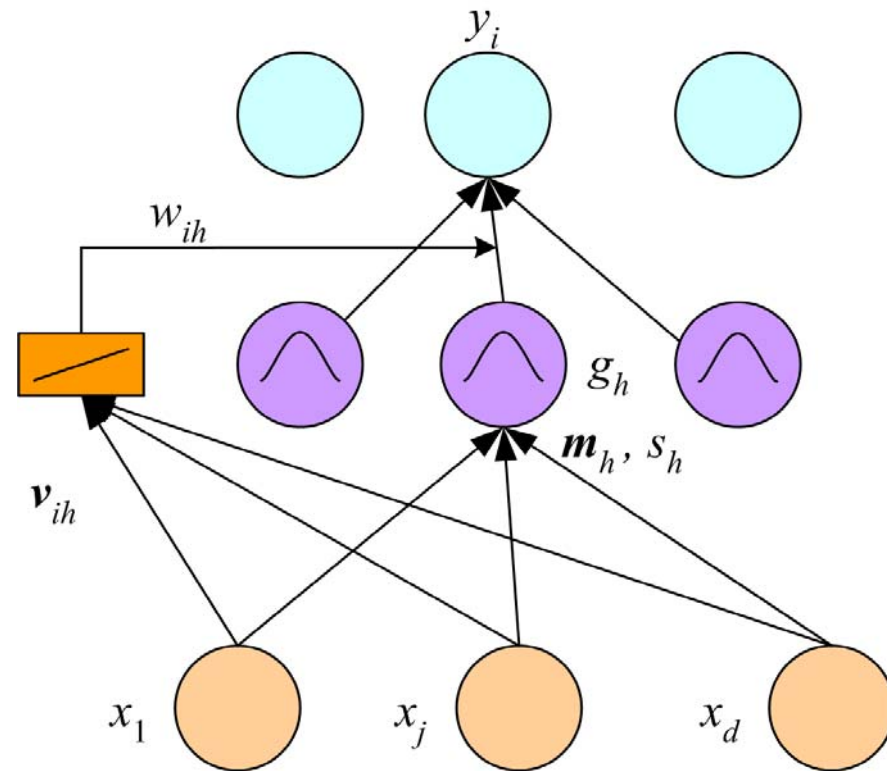


Mixture of Experts

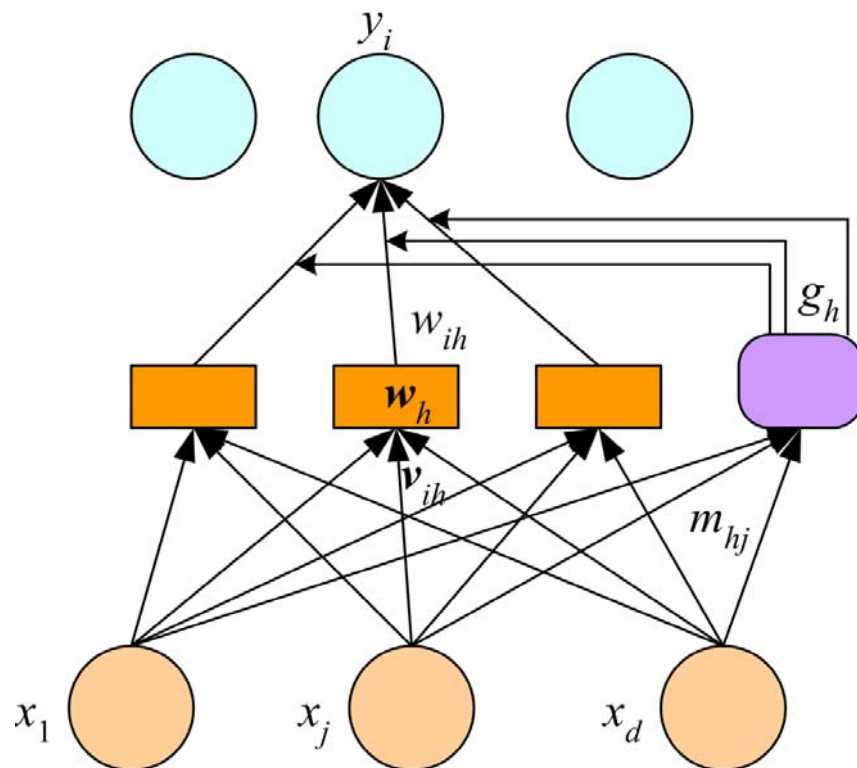
- In RBF, each local fit is a constant, w_{ih} , second layer weight
- In MoE, each local fit is a linear function of x , a local expert:

$$w_{ih}^t = \mathbf{v}_{ih}^T \mathbf{x}^t$$

(Jacobs et al., 1991)



MoE as Models Combined



- Radial gating:

$$g_h^t = \frac{\exp[-\|\mathbf{x}^t - \mathbf{m}_h\|^2 / 2s_h^2]}{\sum_l \exp[-\|\mathbf{x}^t - \mathbf{m}_l\|^2 / 2s_l^2]}$$

- Softmax gating:

$$g_h^t = \frac{\exp[\mathbf{m}_h^T \mathbf{x}^t]}{\sum_l \exp[\mathbf{m}_l^T \mathbf{x}^t]}$$



Cooperative MoE

■ Regression

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

$$\Delta \mathbf{v}_{ih} = \eta \sum_t (r_i^t - y_i^t) g_h^t \mathbf{x}^t$$

$$\Delta m_{hj} = \eta \sum_t \sum_i (r_i^t - y_i^t) (w_{ih}^t - y_i^t) g_h^t x_j^t$$

Competitive MoE: Regression

$$\mathcal{L}(\{\mathbf{m}_h, s_h, w_{ih}\}_{i,h} | \mathcal{X}) = \sum_t \log \sum_h g_h^t \exp \left[-\frac{1}{2} \sum_i (r_i^t - y_{ih}^t)^2 \right]$$

$$y_{ih}^t = w_{ih}^t = \mathbf{v}_{ih} \mathbf{x}^t$$

$$\Delta \mathbf{v}_{ih} = \eta \sum_t (r_i^t - y_{ih}^t) f_h^t \mathbf{x}^t$$

$$\Delta \mathbf{m}_h = \eta \sum_t (f_h^t - g_h^t) \mathbf{x}^t$$



Competitive MoE: Classification

$$\begin{aligned}\mathcal{L}(\{\mathbf{m}_h, s_h, w_{ih}\}_{i,h}|\mathcal{X}) &= \sum_t \log \sum_h g_h^t \prod_i (y_{ih}^t)^{r_i^t} \\ &= \sum_t \log \sum_h g_h^t \exp \left[\sum_i r_i^t \log y_{ih}^t \right]\end{aligned}$$

$$y_{ih}^t = \frac{\exp w_{ih}^t}{\sum_k \exp w_{kh}^t} = \frac{\exp[\mathbf{v}_{ih}\mathbf{x}^t]}{\sum_k \exp[\mathbf{v}_{kh}\mathbf{x}^t]}$$



Hierarchical Mixture of Experts

- Tree of MoE where each MoE is an expert in a higher-level MoE
- **Soft decision tree**: Takes a weighted (gating) average of all leaves (experts), as opposed to using a single path and a single leaf
- Can be trained using EM (Jordan and Jacobs, 1994)