Data Mining in Vector Space or Metric One?

Jaromír Kukal

DSEE FNSPE CTU Prague

May 13th 2010

Basic ideas

- Pattern set in vector space
- Pattern set in metric space
- Conversions
- Sammon's Mapping for metric spaces
- Cluster Analysis in both cases
- SOM in both cases
- PCA in both cases

Pattern in vector space

$$m, n \in \mathbf{N}$$

$$\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,n}) \in \mathbf{R}^n$$

$$\mathbf{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subset \mathbf{R}^n$$

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{pmatrix}$$

Normalization in vector space

$$u_j = \max_i x_{i,j}$$

$$l_j = \min_i x_{i,j}$$

$$x^*_{i,j} = \begin{cases} 2\dfrac{x_{i,j} - l_j}{u_j - l_j} - 1 & \text{for} \quad u_j > l_j \\ 0 & \text{for} \quad u_j = l_j \\ 0 & \text{for empty } x_{i,j} \end{cases}$$

Standardization in vector space

$$\mu_j = \operatorname*{mean}_i x_{i,j}$$

$$\sigma_j = \operatorname*{std}_i x_{i,j}$$

$$x^*_{i,j} = \begin{cases} \dfrac{x_{i,j} - \mu_j}{\sigma_j} & \text{for} \quad \sigma_j > 0 \\ 0 & \text{for} \quad \sigma_j = 0 \\ 0 & \text{for empty } x_{i,j} \end{cases}$$

Robust transform in vector space

$$\eta_j = \underset{i}{\text{median}}\, x_{i,j}$$

$$\lambda_j = \underset{i}{\text{median}} \mid x_{i,j} - \eta_j \mid$$

$$x^*_{i,j} = \begin{cases} \dfrac{x_{i,j} - \eta_j}{\lambda_j} & \text{for} \quad \lambda_j > 0 \\ 0 & \text{for} \quad \lambda_j = 0 \\ 0 & \text{for empty } x_{i,j} \end{cases}$$

Minkowski distance in $\mathbf{R}^n$

$$\mathrm{d}_p(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_k | x_k - y_k |^p \right)^{1/p}, \quad p \geq 1$$

$$\lim_{p\to\infty} \mathrm{d}_p(\boldsymbol{x},\boldsymbol{y}) = \max_k | x_k - y_k |$$

- Manhattan distance $p = 1$
- Compromise distance $p = 3/2$
- Euclidean distance $p = 2$
- Maximum distance $p \to \infty$

Distance representation

$$d_{i,j} = \mathrm{d}_p(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\boldsymbol{D} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,m} \\ \vdots & \ddots & \vdots \\ d_{m,1} & \cdots & d_{m,m} \end{pmatrix}$$

special case $\quad mn > m(m-1)/2$

Metric space

$$\mathrm{card}(\mathbf{U}) > 0$$

$$\langle \mathbf{U}, \mathrm{d} \rangle$$

$$\mathrm{d} : \mathbf{U} \times \mathbf{U} \to \mathbf{R}_0^+$$

$$\forall\, x, y \in \mathbf{U} \quad \mathrm{d}(x, y) = \mathrm{d}(y, x)$$

$$\forall\, x, y \in \mathbf{U} \quad \mathrm{d}(x, y) = 0 \Leftrightarrow x = y$$

$$\forall\, x, y, z \in \mathbf{U} \quad \mathrm{d}(x, y) \leq \mathrm{d}(x, z) + \mathrm{d}(z, y)$$

Pattern in metric space

$$m \in \mathbf{N}$$

$$x_i \in \mathbf{U}$$

$$\mathbf{S} = \{x_1, \ldots, x_m\} \subseteq \mathbf{U}$$

$$d_{i,j} = \mathrm{d}(x_i, x_j)$$

$$\boldsymbol{D} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,m} \\ \vdots & \ddots & \vdots \\ d_{m,1} & \cdots & d_{m,m} \end{pmatrix}$$

Normalization in metric space

$$M = \begin{cases} \max\limits_{i<j} d_{i,j} \\ \operatorname{mean}\limits_{i<j} d_{i,j} \\ \operatorname{median}\limits_{i<j} d_{i,j} \end{cases}$$

$$d^*_{i,j} = d_{i,j} / M$$

$$\boldsymbol{D}^* = \boldsymbol{D} / M$$

Conversion from $\mathbf{U}$ to $\mathbf{R}^n$

$$n \in \mathbf{N}$$

$$e_k \in \mathbf{U}$$

$$\mathbf{E} = \{e_1, \ldots, e_n\} \subseteq \mathbf{U}$$

$$\boldsymbol{x}_i = \left(\mathrm{d}(x_i, e_1), \ldots, \mathrm{d}(x_i, e_n)\right)$$

$$\boldsymbol{x}_i \in \mathbf{R}^n$$

$$\mathbf{E} = \mathbf{S} \Rightarrow \boldsymbol{X} = \boldsymbol{D}$$

PCA preliminaries

$$\text{PCA} : \mathbf{R}^n \rightarrow \mathbf{R}^H, \text{ where } H < n$$

$$\boldsymbol{y} = \text{PCA}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W} \quad \boldsymbol{W} \in \mathbf{R}^{n \times H}$$

$$\boldsymbol{X} \in \mathbf{R}^{m \times n} \qquad \sum_i x_{i,j} = 0$$

$$H \leq \text{rank}(\boldsymbol{X}) \leq \min(m-1, n)$$

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$$

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}_m \quad \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}_n$$

$$\boldsymbol{S} = \text{diag}(\boldsymbol{s}) \geq \boldsymbol{0}$$

Traditional PCA

$$\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \geq \boldsymbol{0}$$

$$(\boldsymbol{A} - \lambda \boldsymbol{I}_n)\boldsymbol{v} = \boldsymbol{0} \qquad \|\boldsymbol{v}\| = 1$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$$

$$\boldsymbol{W} = \begin{pmatrix} \dfrac{\boldsymbol{v}_1}{\sqrt{\lambda_1}} & \ldots & \dfrac{\boldsymbol{v}_H}{\sqrt{\lambda_H}} \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} = \begin{pmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_H \end{pmatrix}$$

PCA for large *n*

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \geq \boldsymbol{0}$$

$$(\boldsymbol{B} - \lambda \boldsymbol{I}_m)\boldsymbol{u} = \boldsymbol{0} \qquad \|\boldsymbol{u}\| = 1$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$$

$$\boldsymbol{W} = \boldsymbol{X}^{\mathrm{T}} \begin{pmatrix} \dfrac{\boldsymbol{u}_1}{\lambda_1} & \ldots & \dfrac{\boldsymbol{u}_H}{\lambda_H} \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} = \begin{pmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_H \end{pmatrix}$$

Sammon's mapping

$$\mathrm{f}: \boldsymbol{U} \rightarrow \boldsymbol{V} \qquad y = \mathrm{f}(x)$$

$$\langle \boldsymbol{U}, \mathrm{d} \rangle \qquad d_{i,j} = \mathrm{d}(x_i, x_j)$$

$$\langle \boldsymbol{V}, \delta \rangle \qquad \delta_{i,j} = \delta(y_i, y_j)$$

$$F = \frac{\sum_{i<j} \frac{(d_{i,j} - \delta_{i,j})^2}{d_{i,j}}}{\sum_{i<j} d_{i,j}} = \min_f$$

$$\text{postulate}: d_{i,j} = \text{undefined} \Rightarrow \frac{(d_{i,j} - \delta_{i,j})^2}{d_{i,j}} = 0$$

Nonlinear alternative to PCA

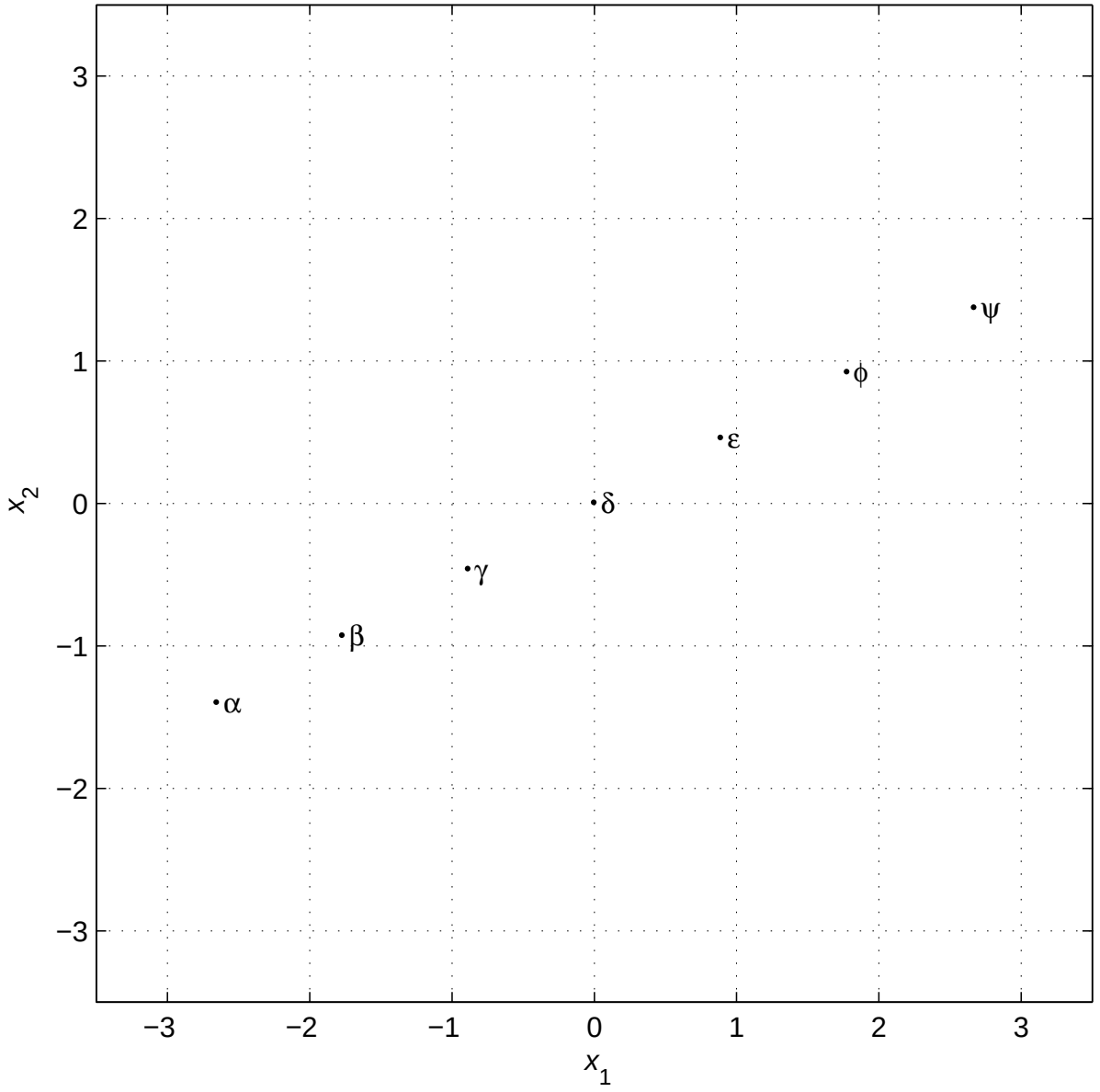
$$H \in \mathbf{N}$$

$$\mathbf{V} = \mathbf{R}^{H}$$

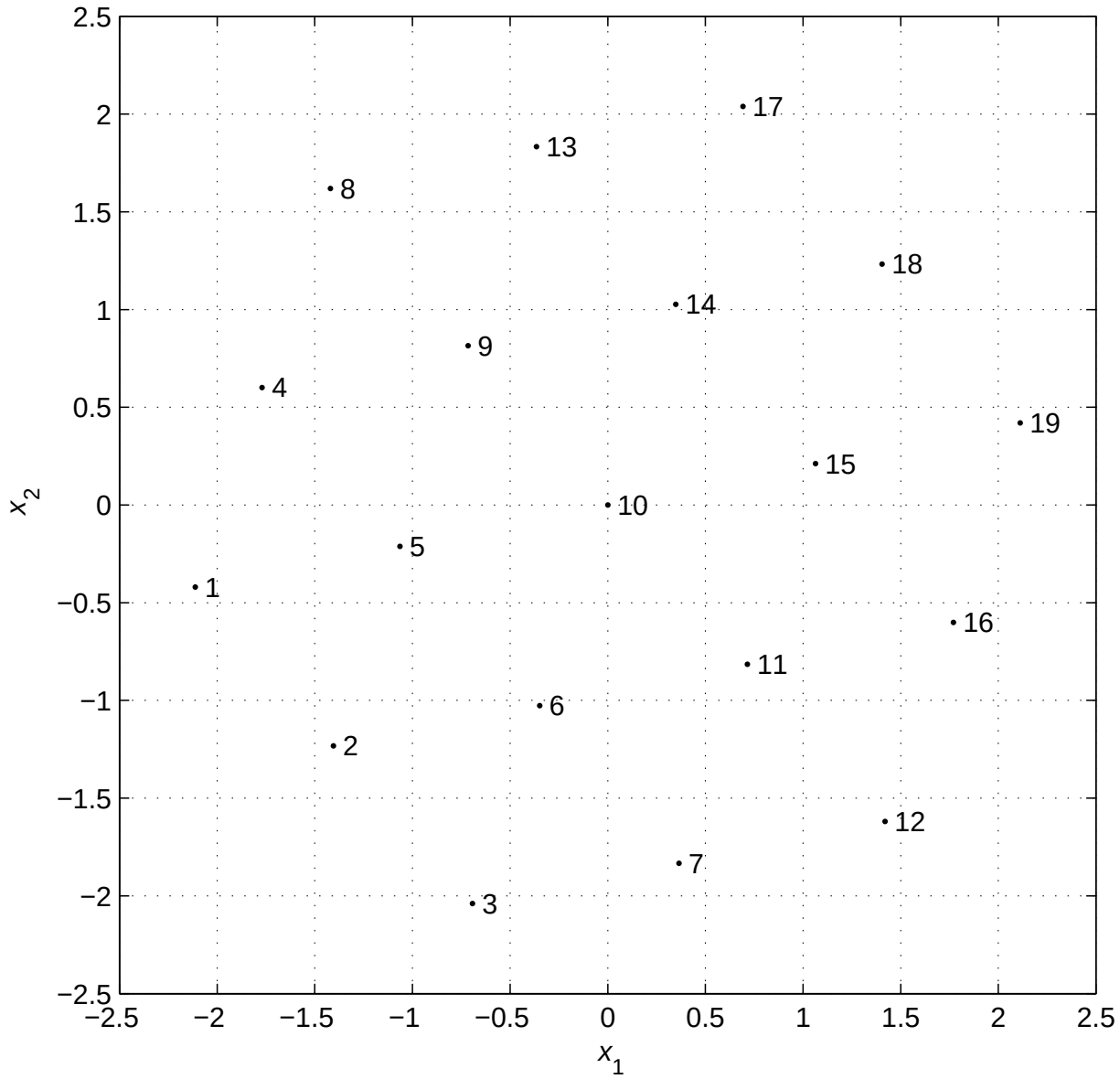
$$\delta(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_{k} (a_k - b_k)^2 \right)^{1/2}$$

$$\Phi_1 = \sum_{i<j} \frac{(d_{i,j} - \delta(\boldsymbol{y}_i, \boldsymbol{y}_j))^2}{d_{i,j}} = \min_{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m}$$

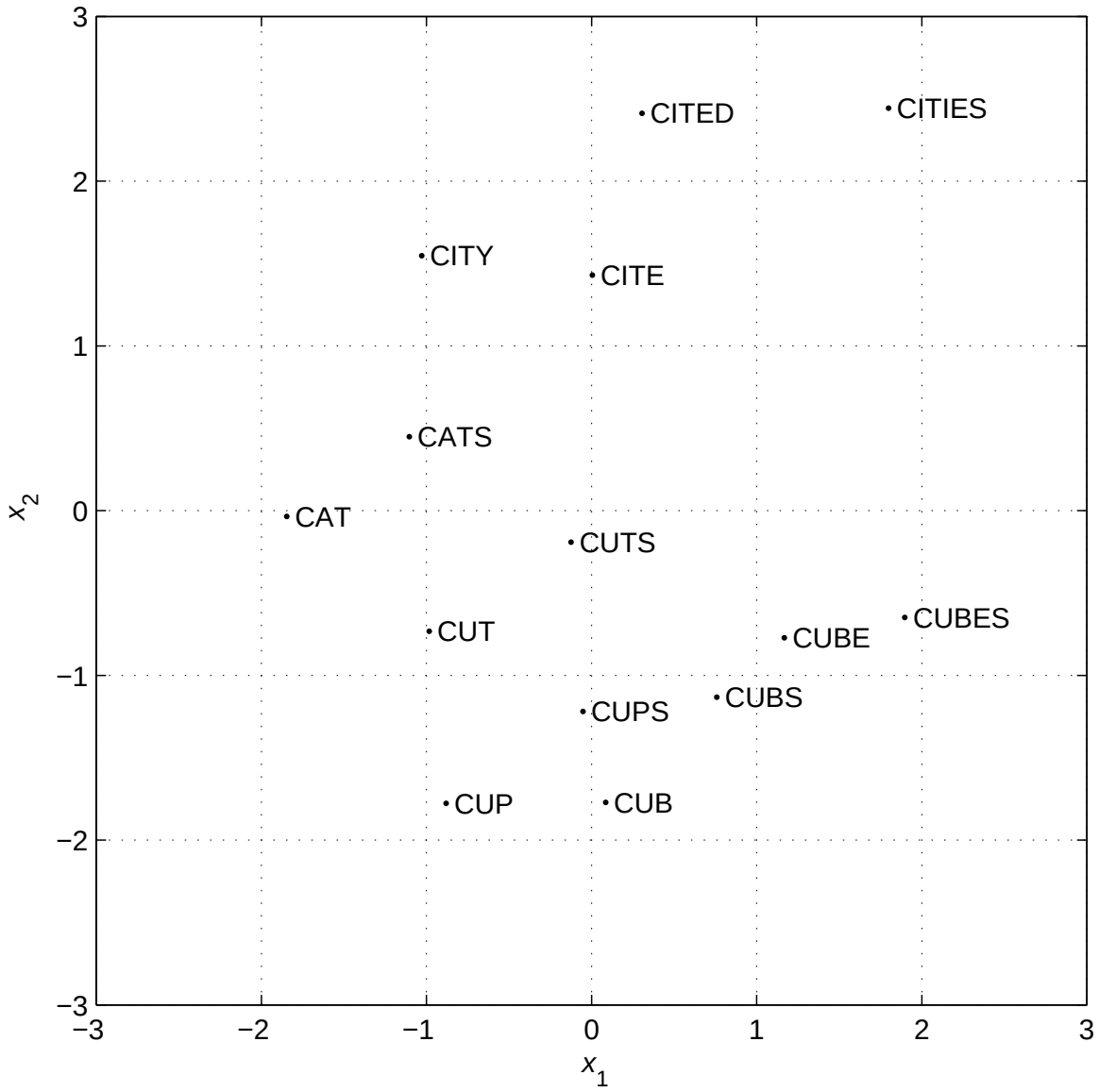
Are they sitting on the bench?



Plot of hexagonal graph



Levensthein distance in U



Modified PCA

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_H) \in \left(\mathbf{R}^+\right)^H$$

$$\boldsymbol{y} = (PCA_1, \ldots, PCA_H) \in \mathbf{R}^H$$

$$\boldsymbol{y}_{\text{modi}} = \boldsymbol{\varepsilon} \otimes \boldsymbol{y} = (\varepsilon_1 PCA_1, \ldots, \varepsilon_H PCA_H) \in \mathbf{R}^H$$

$$\delta(\boldsymbol{a}, \boldsymbol{b}) = \left(\sum_k (a_k - b_k)^2 \right)^{1/2}$$

$$\Phi_2 = \sum_{i<j} \frac{(d_{i,j} - \delta(\boldsymbol{\varepsilon} \otimes \boldsymbol{y}_i, \boldsymbol{\varepsilon} \otimes \boldsymbol{y}_j))^2}{d_{i,j}} = \min_{\boldsymbol{\varepsilon}}$$

Alternative to cluster analysis

$$H > 1 \qquad \mathbf{V} = \{1,\ldots,H\}$$

$$\delta(a,b) = \begin{cases} 0 & \text{for} \quad a = b \\ \text{undefined} & \text{for} \quad a \neq b \end{cases}$$

$$\boldsymbol{p} = (p_1,..,p_m) \in \mathbf{V}^m$$

$$\text{postulate}: \delta_{i,j} = \text{undefined} \Rightarrow \frac{(d_{i,j} - \delta_{i,j})^2}{d_{i,j}} = 0$$

$$\Phi_3 = \sum_{\substack{i<j \\ \delta_{i,j}=0}} \frac{(d_{i,j} - \delta_{i,j})^2}{d_{i,j}} = \sum_{\substack{i<j \\ \delta_{i,j}=0}} d_{i,j} = \sum_{\substack{i<j \\ p_i = p_j}} d_{i,j} = \min_{\boldsymbol{p}}$$

Alternative to SOM

$$\mathrm{G} = \langle \mathbf{V}, \mathbf{E} \rangle \qquad \mathbf{V} = \{1, \ldots, H\} = \mathrm{V(G)}$$

$$\mathbf{E} = \mathrm{E(G)} \subset \binom{\mathbf{V}}{2} \qquad \delta(a,b) \text{ is vertex distance in G}$$

$$\boldsymbol{p} = (p_1, .., p_m) \in \mathbf{V}^m \qquad a > 0$$

$$\Phi_4 = \sum_{i<j} \frac{(d_{i,j} - a\,\delta(p_i, p_j))^2}{d_{i,j}} = \min_{\boldsymbol{p},a}$$

Cluster and partition

$$H > 1 \qquad \mathbf{V} = \{1, \ldots, H\}$$

$$\boldsymbol{p} = (p_1, .., p_m) \in \mathbf{V}^m$$

$$\mathbf{C}_k = \{x_i \in \mathbf{S} \mid p_i = k\}$$

$$\mathbf{S} = \bigcup_{k \in \mathbf{V}} \mathbf{C}_k$$

Momentum in metric space

$$\mathrm{F}_k(y) = \sum_{x_i \in \mathbf{C}_k} \mathrm{d}^2(x_i, y)$$

$$\mathbf{S} \subseteq \mathbf{Q} \subseteq \mathbf{U}$$

$$t_k \in \arg\min_{y \in \mathbf{U}} \mathrm{F}_k(y)$$

$$t_k^+ \in \arg\min_{y \in \mathbf{Q}} \mathrm{F}_k(y)$$

$$F_k^* = \mathrm{F}_k(t_k)$$

$$F_k^* \le F_k^+ = \mathrm{F}_k(t_k^+)$$

$$\mathbf{U} = \mathbf{R}^n \text{ with Euclidean metrics } \Rightarrow \boldsymbol{t}_k = \operatorname*{mean}_{\boldsymbol{x}_i \in \mathbf{C}_k} \boldsymbol{x}_i$$

Center in metric space

$$\mathrm{G}_k(y) = \sum_{x_i \in \boldsymbol{C}_k} \mathrm{d}(x_i, y) \qquad \mathbf{S} \subseteq \mathbf{Q} \subseteq \mathbf{U}$$

$$s_k \in \arg\min_{y \in \mathbf{U}} \mathrm{G}_k(y) \qquad s_k^+ \in \arg\min_{y \in \mathbf{Q}} \mathrm{G}_k(y)$$

$$G_k^* = \mathrm{G}_k(s_k) \qquad G_k^* \le G_k^+ = \mathrm{G}_k(s_k^+)$$

$$\mathbf{U} = \mathbf{R}^n \text{ with Manhattan metrics } \Rightarrow \boldsymbol{s}_k = \operatorname*{median}_{\boldsymbol{x}_i \in \mathbf{C}_k} \boldsymbol{x}_i$$

ISODATA in metric space

Random initialization $\boldsymbol{p} = (p_1,.., p_m) \in \mathbf{V}^m \quad \mathrm{card}(\mathbf{C}_k) > 0$

Momentum revision $t_k^+ \in \arg\min_{y \in \mathbf{Q}} \mathrm{F}_k(y)$

Cluster revision $x_i \in \mathbf{C}_k \Rightarrow k \in \arg\min_{1 \le j \le H} \mathrm{d}(x_i, t_j^+)$

Objective function $\Phi_5 = \sum_k F_k^+ = \min_{\boldsymbol{p}}$

ISODATA in vector space

Random initialization $\boldsymbol{p} = (p_1,.., p_m) \in \mathbf{V}^m \quad \text{card}(\mathbf{C}_k) > 0$

Momentum revision $\boldsymbol{t}_k = \underset{\boldsymbol{x}_i \in \mathbf{C}_k}{\text{mean}}\, \boldsymbol{x}_i$

Cluster revision $\boldsymbol{x}_i \in \mathbf{C}_k \Rightarrow k \in \arg\min_{1 \le j \le H} \text{d}(\boldsymbol{x}_i, \boldsymbol{t}_j)$

Objective function $\Phi_6 = \sum_k F_k^* = \min_{\boldsymbol{p}}$

SOM preliminaries

$$\mathrm{G} = \langle \mathbf{V}, \mathbf{E} \rangle \qquad \mathbf{V} = \{1,\ldots,H\} = \mathrm{V(G)}$$

$$\mathbf{E} = \mathrm{E(G)} \subset \binom{\mathbf{V}}{2} \qquad \delta(a,b) \text{ is vertex distance in G}$$

$$R > 0 \text{ is learning radius}$$

$$\boldsymbol{p} = (p_1,..,p_m) \in \mathbf{V}^m \qquad \boldsymbol{w} = (w_1,..,w_H) \in \mathbf{U}^H$$

$$\chi : \mathbf{N}_0 \to [0;1]$$

$$\forall\, k \in \mathbf{N}_0 : \chi(k) \geq \chi(k+1)$$

$$\chi(0) = 1 \qquad \chi(M) < 1 \qquad M = \max_{a,b \in \mathbf{V}} \delta(a,b)$$

SOM characteristic

$$\chi(\delta) = \begin{cases} 1 & \text{for} \quad \delta \le R \\ 0 & \text{for} \quad \delta > R \end{cases}$$

$$\chi(\delta) = \exp\left(-\frac{\delta^2}{2R^2}\right)$$

$$\chi(\delta) = \left(1 + \frac{\delta^2}{R^2}\right)^{-1}$$

Batch SOM in metric space

Random initialization $\boldsymbol{p} = (p_1,.., p_m) \in \mathbf{V}^m \qquad \mathrm{card}(\mathbf{C}_k) > 0$

Weight revision $w_i \in \arg\min_{y \in \mathbf{Q}} \sum_{k=1}^{H} \sum_{x_j \in \mathbf{C}_k} \mathrm{d}^2(x_j, y)\, \chi(\delta(i,k))$

Vertex revision $x_i \in \mathbf{C}_k \Rightarrow k \in \arg\min_{1 \le k \le H} \sum_{i=1}^{H} \mathrm{d}^2(x_j, w_i)\, \chi(\delta(i,k))$

Objective function $\Phi_7 = \sum_{k=1}^{H} \sum_{i=1}^{H} \sum_{x_j \in \mathbf{C}_k} \mathrm{d}^2(x_j, w_i)\, \chi(\delta(i,k)) = \min_{\boldsymbol{p},\boldsymbol{w}}$

Batch SOM in vector space

Random initialization $\quad \boldsymbol{p} = (p_1, .., p_m) \in \mathbf{V}^m \quad \mathrm{card}(\mathbf{C}_k) > 0$

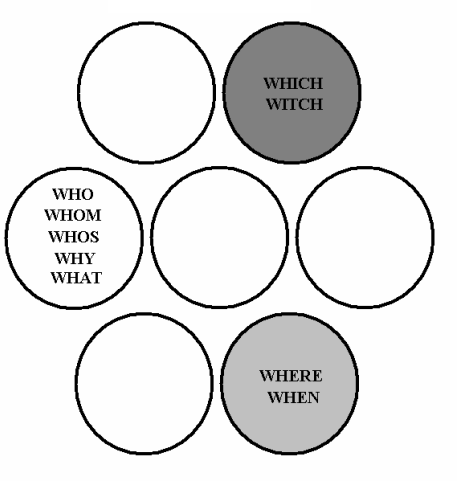
Weight revision $\quad \boldsymbol{w}_i = \dfrac{\sum_{k=1}^{H} \sum_{x_j \in \mathbf{C}_k} \boldsymbol{x}_j \, \chi(\delta(i,k))}{\sum_{k=1}^{H} \sum_{x_j \in \mathbf{C}_k} \chi(\delta(i,k))}$

Vertex revision $\quad \boldsymbol{x}_i \in \mathbf{C}_k \Rightarrow k \in \arg\min_{1 \le k \le H} \sum_{i=1}^{H} \mathrm{d}^2(\boldsymbol{x}_j, \boldsymbol{w}_i) \, \chi(\delta(i,k))$

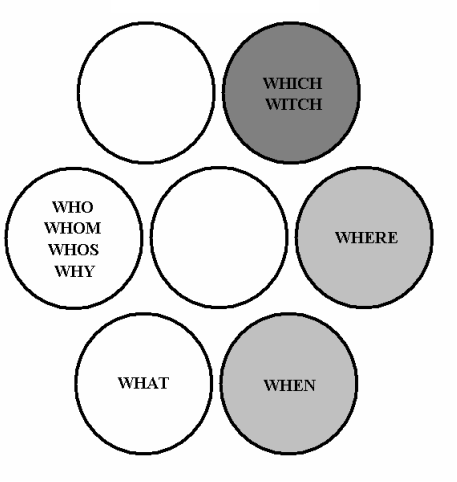
Objective function $\quad \Phi_8 = \sum_{k=1}^{H} \sum_{i=1}^{H} \sum_{\boldsymbol{x}_j \in \mathbf{C}_k} \mathrm{d}^2(\boldsymbol{x}_j, \boldsymbol{w}_i) \, \chi(\delta(i,k)) = \min_{\boldsymbol{p}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_H}$

SOM of written words

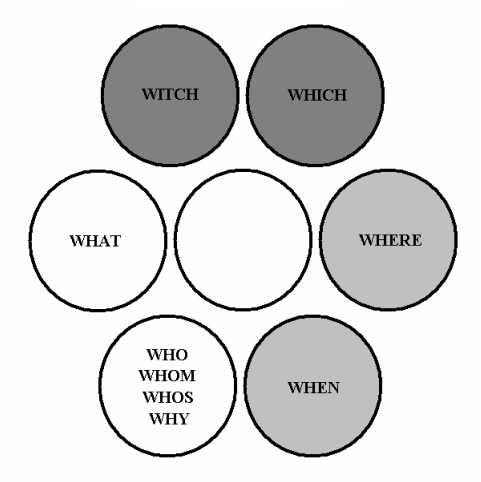
$R = 1.0$



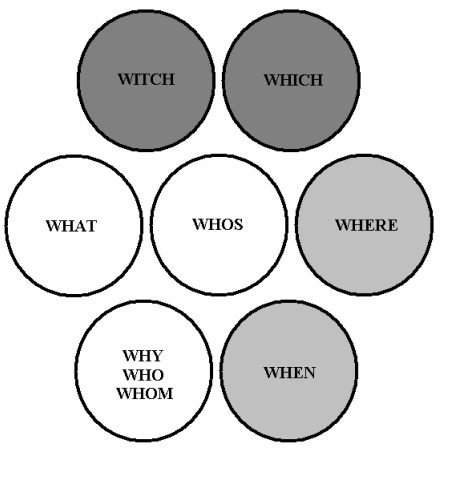
$R = 0.5$



$R = 0.2$

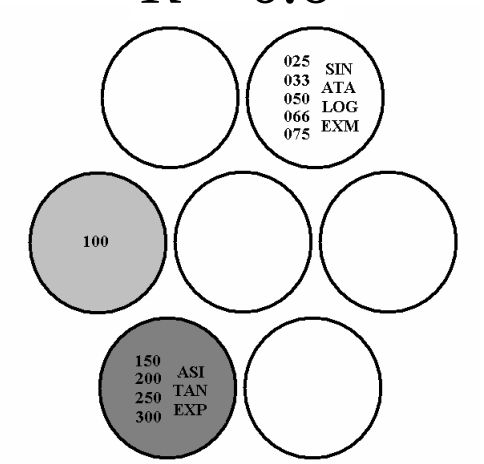


$R = 0.1$

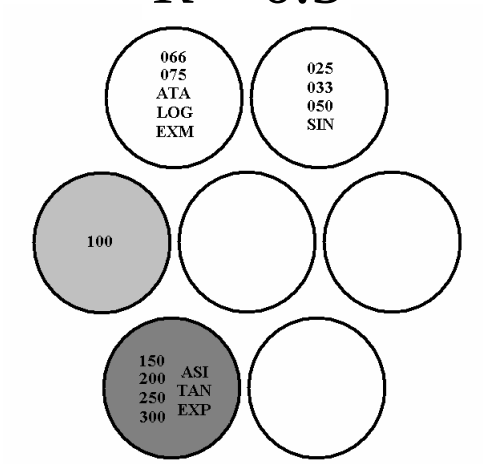


SOM of functions

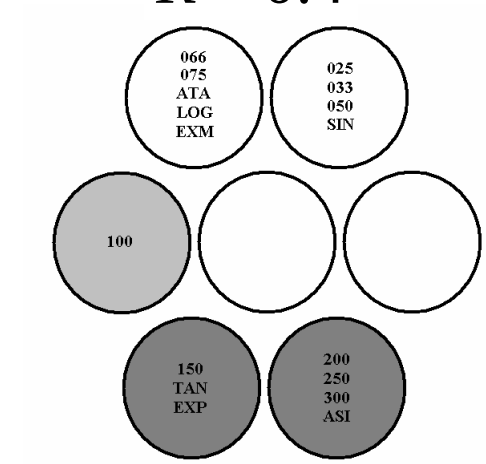
$R = 0.6$



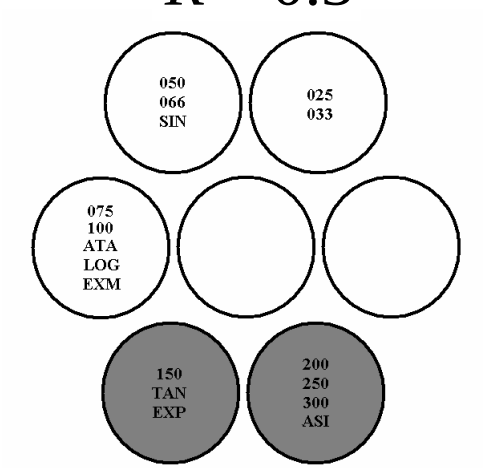
$R = 0.5$



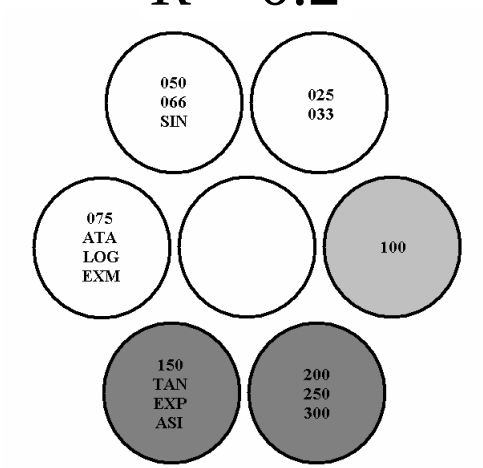
$R = 0.4$



$R = 0.3$



$R = 0.2$



$R = 0.1$