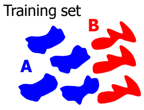
Dissimilarity Representation

Dissimilarity Representation

Not used by NN Rule

Training set



Dissimilarities d_{ij} between all training objects

$$D_T = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\ d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} \end{pmatrix}$$

Unlabeled object **x** to be classified

$$d_x = (d_{x1}\ d_{x2}\ d_{x3}\ d_{x4}\ d_{x5}\ d_{x6}\ d_{x7})$$

The traditional Nearest Neighbor rule (template matching) finds:

$\text{label}(\text{argmin}_{\text{trainset}}\{d_{xi}\})$,

without using D_T. Can we do any better?

TU Delft

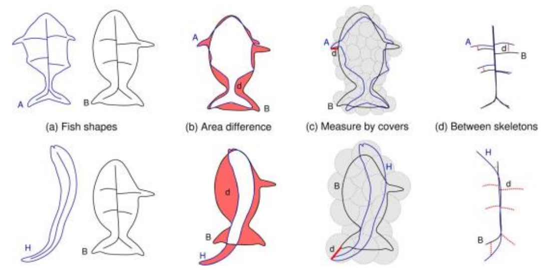
Dissimilarities – Possible Assumptions

Metric

1. Positivity: $d_{ij} \geq 0$
2. Reflexivity: $d_{ii} = 0$
3. Definiteness: $d_{ij} = 0$ objects i and j are identical
4. Symmetry: $d_{ij} = d_{ji}$
5. Triangle inequality: $d_{ij} < d_{ik} + d_{kj}$
6. Compactness: if the objects i and j are very similar then $d_{ij} < \delta$.
7. True representation: if $d_{ij} < \delta$ then the objects i and j are very similar.
8. Continuity of d.

TU Delft

Examples Dissimilarity Measures (1)

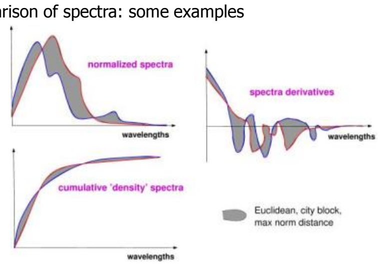


The measure should be descriptive. If there is no preference, a number of measures can be combined.

TU Delft

Examples Dissimilarity Measures (2)

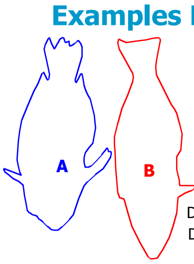
Comparison of spectra: some examples



In real applications, the dissimilarity measure should be robust to noise and small aberrations in the (raw) measurements.

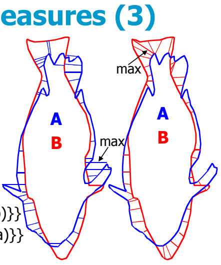
TU Delft

Examples Dissimilarity Measures (3)



Dist(**A**,**B**):
a ∈ **A**, points of **A**
b ∈ **B**, points of **B**
d(a,b): Euclidean distance

D(**A**,**B**) = max_a{min_b{d(a,b)}}
D(**B**,**A**) = max_b{min_a{d(b,a)}}



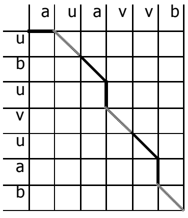
Hausdorff Distance (metric):
DH = max{max_a{min_b{d(a,b)}} , max_b{min_a{d(b,a)}}}

D(**A**,**B**) ≠ D(**B**,**A**)

Modified Hausdorff Distance (non-metric):
DM = max{mean_a{min_b{d(a,b)}},mean_b{min_a{d(b,a)}}}

TU Delft

Examples Dissimilarity Measures (4)



$D_E(X,Y)$: Σ edit operations $X \Rightarrow Y$
(insertions, deletions, substitutions)

DE(snert ,meer) = 3:
snert $\Rightarrow$ seert $\Rightarrow$ seer $\Rightarrow$ meer

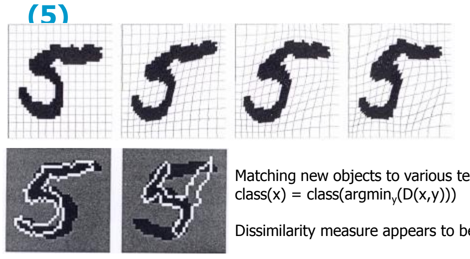
DE(ner ,meer) = 2:
ner $\Rightarrow$ mer $\Rightarrow$ meer

Possibly weighted

Triangle inequality $\Rightarrow$ computationally feasible

D(aa,bb) < D(abcdef,bcdd)

Examples Dissimilarity Measures (5)

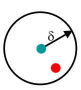


Matching new objects to various templates:
class(x) = class(argmin$_y$(D(x,y)))

Dissimilarity measure appears to be non-metric.

A.K. Jain, D. Zongker, Representation and recognition of handwritten digit using deformable templates, IEEE-PAMI, vol. 19, no. 12, 1997, 1386-1391.

Prospect of Dissimilarity based Representations: Zero



Let us assume that we deal with true representations:
$d_{ab} < \delta$ if and only if the objects a and b are very similar.

If δ is sufficiently small then a and b belong to the same class, as b is just a minor distortion of a (assuming true representations).

However, as Prob(b) > 0, there will be such an object for sufficiently large training sets $\Rightarrow$ zero classification error possible!

$\Rightarrow$ Dissimilarity representation can be a true representation

Why a Dissimilarity Representation?

Many (exotic) dissimilarity measures are used in pattern recognition
- they may solve the connectivity problem (e.g. pixel based features)
- they may offer a way to integrate structural and statistical approaches e.g. by graph distances.

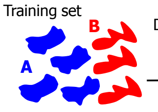
Prospect of zero-error classifiers by avoiding class overlap

Better rules than the nearest neighbor classifier appear possible (more accurate, faster)

Classification of Dissimilarity Data

Alternatives for the Nearest Neighbor Rule

Training set



Dissimilarities d_{ij} between all training objects

$$D_T = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\ d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} \end{pmatrix}$$

Unlabeled object **x** to be classified

$$d_x = (d_{x1}\ d_{x2}\ d_{x3}\ d_{x4}\ d_{x5}\ d_{x6}\ d_{x7})$$

1. Dissimilarity Space
2. Embedding

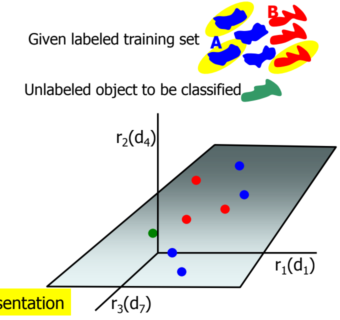
Pekalska, The dissimilarity representation for PR. World Scientific, 2005.

The Dissimilarity Space

Alternative 1: Dissimilarity Space

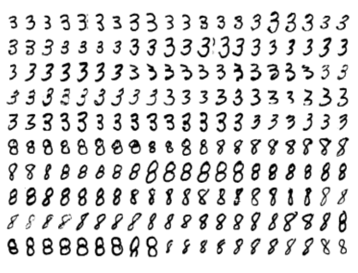
Dissimilarities

$$
\begin{array}{c}
\begin{array}{ccccccc} r_1 & & & r_2 & & & r_3 \end{array} \\
D_T = \begin{pmatrix}
d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\
d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\
d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\
d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\
d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\
d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\
d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77}
\end{pmatrix} \\
d_x = (d_{x1}\ d_{x2}\ d_{x3}\ d_{x4}\ d_{x5}\ d_{x6}\ d_{x7})
\end{array}
$$



Selection of 3 objects for representation

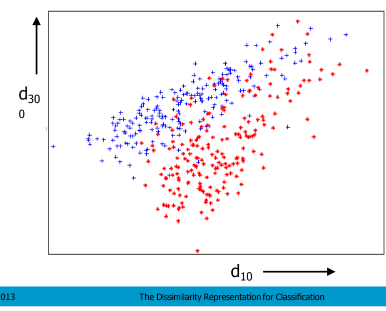
Example Dissimilarity Space: NIST Digits 3 and 8



Example of raw data

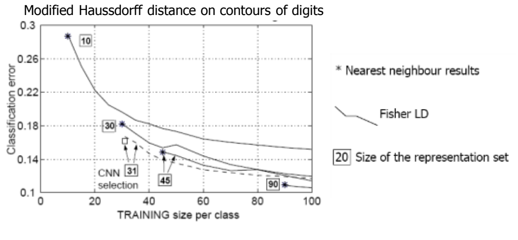
Example Dissimilarity Space: NIST Digits 3 and 8

NIST digits: Hamming distances of 2 x 200 digits



Dissimilarity Space Classification ←→ Nearest Neighbor Rule

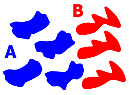
Modified Haussdorff distance on contours of digits



Dissimilarity based classification outperforms the nearest neighbor rule.

Embedding of (non-Euclidean) Dissimilarities

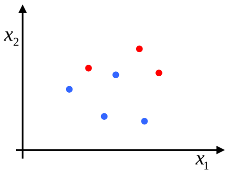
Alternative 2: Embedding



→ Dissimilarity matrix D → X

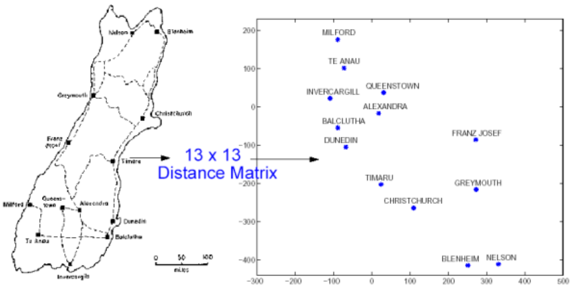
Training set

Is there a feature space for which Dist(X,X) = D ?

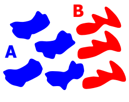


Position points in a vector space such that their Euclidean distances → D

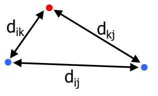
Embedding



Embedding of non-metric measurements

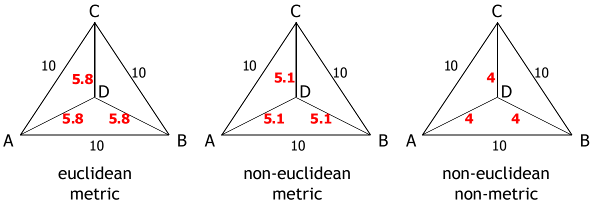


→ Dissimilarity matrix D → X



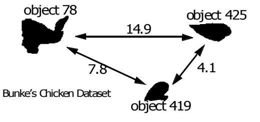
If the dissimilarity matrix cannot be explained from a vector space,
(e.g. for Hausdorff and Hamming distance of images)
or if $d_{ij} > d_{ik} + d_{kj}$ (triangle inequality not satisfied)
embedding in Euclidean space not possible
→ Pseudo-Euclidean embedding

Euclidean - Non Euclidean - Non Metric

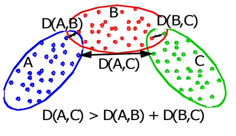


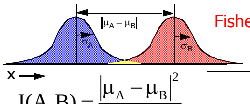
Non-metric distances

Weighted-edit distance for strings



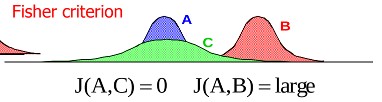
Single-linkage clustering





$$J(A,B) = \frac{|\mu_A - \mu_B|^2}{\sigma_A^2 + \sigma_B^2}$$

Fisher criterion



$$J(A,C) = 0 \quad J(A,B) = \text{large}$$

$$J(C,B) = \text{small} \neq J(A,B)$$

(Pseudo-)Euclidean Embedding

$m \times m$ D is a given, imperfect dissimilarity matrix of training objects.

Construct inner-product matrix: $B = -\frac{1}{2} J D^{(2)} J \quad J = I - \frac{1}{m}\mathbf{11}$

Eigenvalue Decomposition , $B = Q \Lambda Q^T$

Select k eigenvectors: $X = Q_k \Lambda_k^{\frac{1}{2}}$ (problem: $\Lambda_k < 0$)

Let $\Im_k$ be a k x k diag. matrix, $\Im_k(i,i)$ = sign($\Lambda_k(i,i)$)

$\Lambda_k(i,i) < 0 \rightarrow$ Pseudo-Euclidean

$n \times m$ D_z is the dissimilarity matrix between new objects and the training set.

The inner-product matrix: $B_z = -\frac{1}{2}(D_z^{(2)} J - \frac{1}{n}\mathbf{11}^T D^{(2)} J)$

The embedded objects: $Z = B_z Q_k |\Lambda_k|^{-\frac{1}{2}} \Im_k$

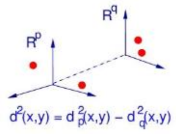
PES: Pseudo-Euclidean Space (Krein Space)

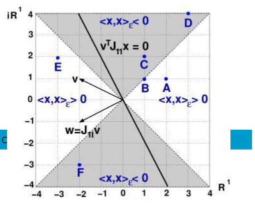
If D is non-Euclidean, B has p positive and q negative eigenvalues.

A pseudo-Euclidean space $\mathcal{E}$ with signature (p,q), k =p+q, is a non-degenerate inner product space $\Re_k = \Re_p \oplus \Re_q$ such that:

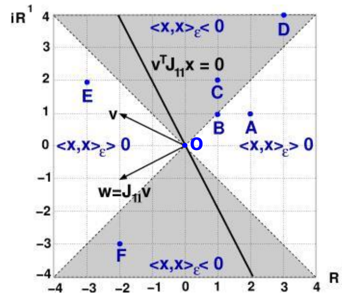
$$\langle x,y\rangle_\varepsilon = x^T \Im_{pq} y = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{q} x_j y_j \qquad \Im_{pq} = \begin{bmatrix} I_{p\times p} & 0 \\ 0 & -I_{q\times q}\end{bmatrix}$$

$$d_\varepsilon^2(x,y) = \langle x-y, x-y\rangle_\varepsilon = d_p^2(x,y) - d_q^2(x,y)$$





Distances in PES



$d^2(O,A) > 0$

$d^2(O,E) > 0$

$d^2(O,B) = 0$

$d^2(O,D) < 0$

All points in the grey area are closer to O than O itself !?

Any point has a negative square distance to some points on the line $v^T \Im x = 0$.

Can it be used as a classifier?

Can we define a margin as in the SVM?

PE Space ←→ Kernels

$$K(x,y) = -\tfrac{1}{2} J D(x,y)^{(2)} J \qquad J = I - \tfrac{1}{m}\mathbf{11}$$

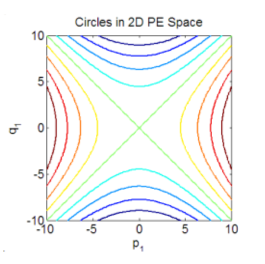
may by considered as a kernel. If

$$K(x,y) = < L(x), L(y) >$$

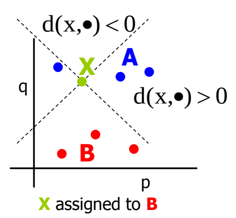
- The **kernel trick** may be used: operations defined on inner products in kernel space can be operated directly on K(x,y) **without embedding**!
- True for **Mercer kernels** (all eigenvalues ≥ 0).
- Difficult for **indefinite kernels**.
- Studying classifiers in **PE space** is studying the indefinite **kernel space**.
- Dissimilarities are more informative than kernels (due to normalization).

Classifiers in Pseudo-Euclidean Space

Distance based classifiers in PE Space



Metric in PE Space.
Equidistant points to the origin.



X assigned to **B**

Nearest Neighbour and
Nearest Mean can be properly defined.
SVM ? What is the distance to a line?

SVM in PE Space

- SVM on indefinite kernels may not converge as Mercer's conditions are not fulfilled.
- However, if it converges the solution is proper:

$$| w^T \Im w |$$

is minimized.

- *See also: B. Haasdonk, Feature Space Interpretation of SVMs with Indefinite Kernels, IEEE PAMI, 24, 482-492, 2005.*

Densities in PE Space

- Densities can be defined in a vector space on de basis of volumes, without the need of a metric.
- Density estimates however, often need a metric. E.g. the Parzen estimator:

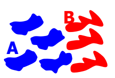
$$\hat{f}(x) = \frac{1}{n}\sum_{y_i} c\exp\left(-\frac{d(x,y_i)^2}{2h^2}\right)$$

needs a distance definition $d(x,y)$.
- There is no problem, however, in case for all objects $d(x,y) > 0$.
- How can Gaussian densities be defined?
- Note that QDA in PES is identical to the QDA in AES as the signature cancels. The relation with a Gaussian distribution, however, is lost.

Dissimilarity based classifiers compared

Dissimilarity based classification procedured compared

Training set



→ Dissimilarity matrix D

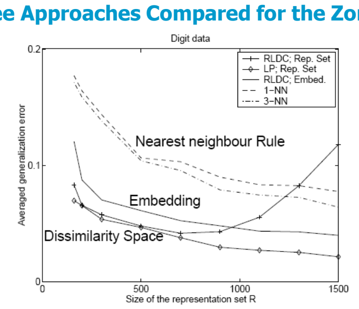
Test object x

→ Dissimilarities d_x with training set

1. Nearest Neighbour Rule
2. Reduce training set to representation set
$\Rightarrow$ dissimilarity space
3. Embedding:Select large $\Lambda_{ii} > 0 \Rightarrow$ Euclidean space
Select large $|\Lambda_{ii}| > 0 \rightarrow$ pseudo-Euclidean space

} discriminant function

Three Approaches Compared for the Zongker Data

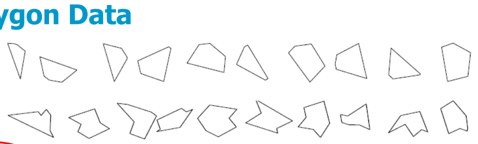


Dissimilarity Space equivalent to Embedding better than Nearest Neighbour Rule

Polygon Data

Convex Pentagons

Heptagons

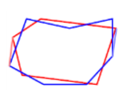


no class overlap zero error

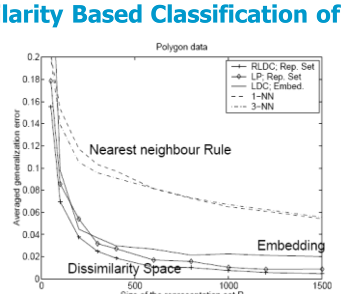
Minimum edge length: 0.1 of maximum edge length

Distance measures: Hausdorff $D = \max \{ \max_i(\min_j(d_{ij}))\ ,\ \max_j(\min_i(d_{ij})) \}$.
Modified Hausdorff $D = \max \{\mathrm{mean}_i(\min_j(d_{ij})),\ \mathrm{mean}_j(\min_i(d_{ij})) \}$. (no metric!)
d_{ij} = distance between vertex i of polygon_1 and vertex j of polygon_2.
Polygons are scaled and centered.

Find the largest of the smallest vertex distances



Dissimilarity Based Classification of Polygons



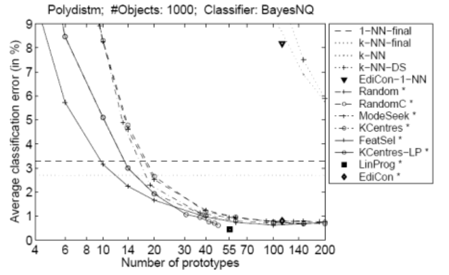
Zero error difficult to reach!

Prototype Selection

Assume D(T,R) are the distances between a training set T and a representation set R.
A classifier can be trained:
- on the distances directly
- in dissimilarity spaces
- in embedded spaces defined by D(R,R)

Selection of prototypes $R \subset T$:
- Random
- k-centres, mode seeking or some clustering procedure
- Feature selection methods
- Editting and condensing methods
- Sparse linear programming methods (L_1-norm SVM)

Prototype Selection: Polygon Dataset



The classification performance of the quadratic Bayes Normal classifier and the k-NN in dissimilarity spaces and the direct k-NN, as a function of the number of selected prototypes. Note that for 10-20 prototypes already better results are obtained than by using 1000 objects in the NN rules.

Dissimilarity Representation

- Based on a pairwise comparison of objects
- Alternative to features for using expert knowledge
- Various ways of construction vector spaces, useful for traditional classifiers.
- May show good performances compared to nearest neigbour rule