Topological approaches in machine learning

D. A. Zighed

University of Lyon (Lumière Lyon 2)

Recife - Brazil - 5..7 May 2009

Basic Concepts for machine learning

Notations

- Ω : Population concerned by the learning issue;
- $\omega \in \Omega$ individual;
- R: multidimensional feature space (p dimensions);
- Features : $X = (X_1, X_2, \ldots, X_j, \ldots, X_p)$ where $X_j : \Omega \longmapsto R_j$; R_j is any set, finite or not
- Belonging classes C; where $C : \Omega \longmapsto \{c_1, \ldots, c_k, \ldots, c_K\}$
- learning sample $\Omega_l \in \Omega$; $|\Omega_l| = n$
- test sample $\Omega_t \in \Omega$; $|\Omega_t| = t$

The Aim of the Machine Learning (ML)

Using the learning data set $(X(\Omega_l), C(\Omega_l))$ to infer a model φ that predicts with high accuracy the membership class C.
The accuracy of the model φ is evaluated on the test sample Ω_t, i.e:
$E(\Omega_t) = \sum_{\omega \in \Omega_t} I(\omega) \approx 0$;
$I(\omega) = 1 \Leftrightarrow (C(\omega) \neq \varphi(\omega))$
otherwise $I(\omega) = 0$;

Learning process

Feature space **Class attribute**

Predictive attributes (categorical)

$(X_1, X_2, X_3, \ldots, X_p)$ C

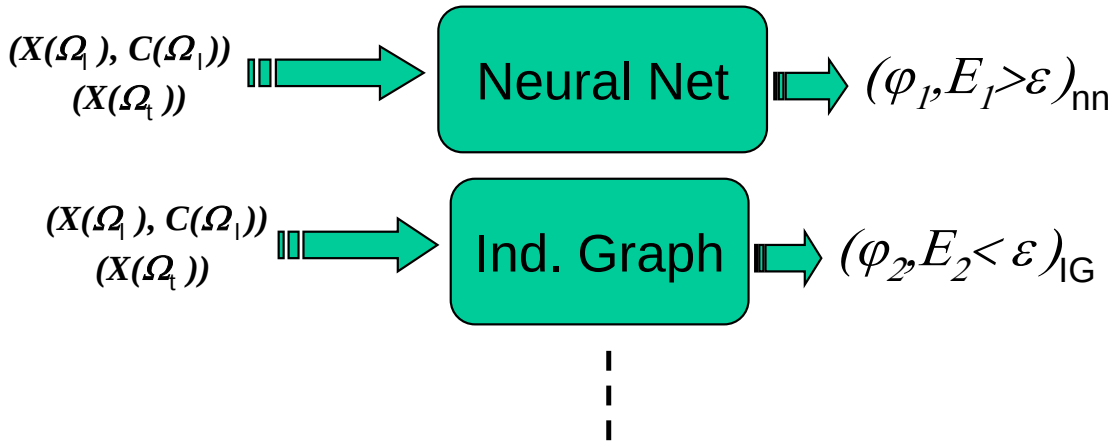
70	1	4	130	322	0	2	109	0	2.40	2	3	3	2
67	0	3	115	564	0	2	160	0	1.60	2	0	7	1
57	1	2	124	261	0	0	141	0	0.30	1	0	7	2
64	1	4	128	263	0	0	105	1	0.20	2	1	7	1
74	0	2	120	269	0	2	121	1	0.20	1	1	3	1
65	1	4	120	177	0	0	140	0	0.40	1	0	7	1
56	1	3	130	256	1	2	142	1	0.60	2	1	6	2
59	1	4	110	239	0	2	142	1	1.20	2	1	7	2
60	1	4	140	293	0	2	170	0	1.20	2	2	7	2

(Learning data set,
any type data, labeled)

- Neural Net
- Induction Graph
- Disc. Analysis
- SVM…

Machine Learning algorithm → φ, ε

Assume that we wish to find a model φ whose the error rate $E \leq \epsilon$. No matter the machine learning algorithm used for that.



What should we conclude if the screening failed ?

- all the machine learning algorithms used are not suitable, therefore we should keep hope and persevere...until when ?
- the classes are not separable, therefore they are not learnable and we should give up the screening.

The key issue

Are we able to determine which one of the two assumptions is the true ?

Proposal : a methodology to assess the **separability** of classes; to evaluate the **complexity** of the underlying patterns and appraise the **relevance** of the feature space.

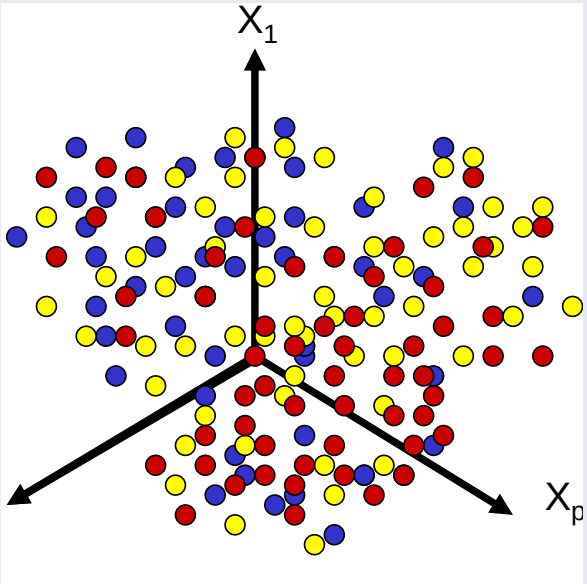
Fundamentals

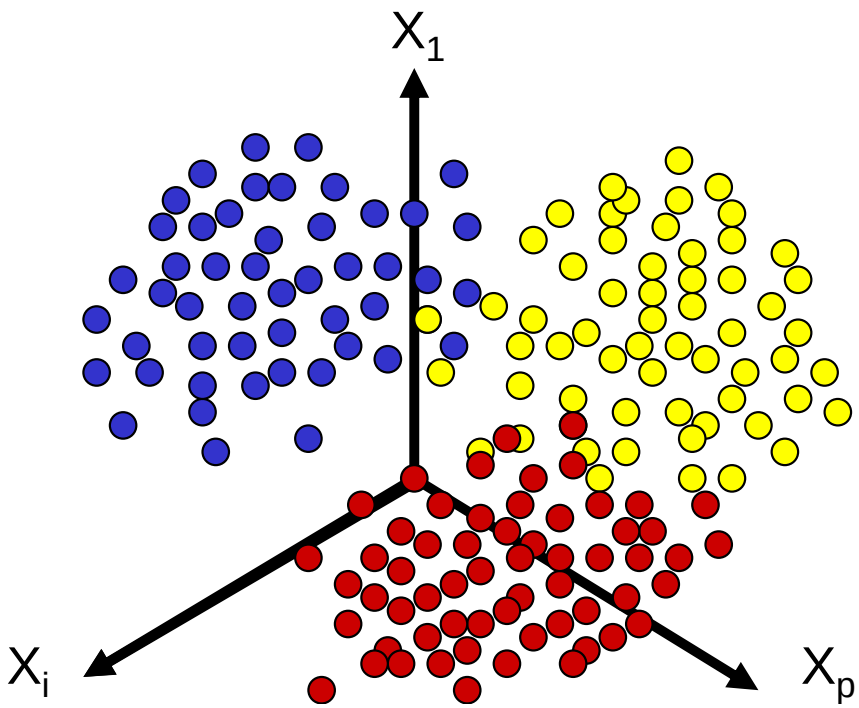
This methodology focuses on the topology of the learning data set in the feature space and exploits its properties.
The key concepts are : **Topology**, **manifolds**, **computational geometry**, **proximity measures**.

Separability

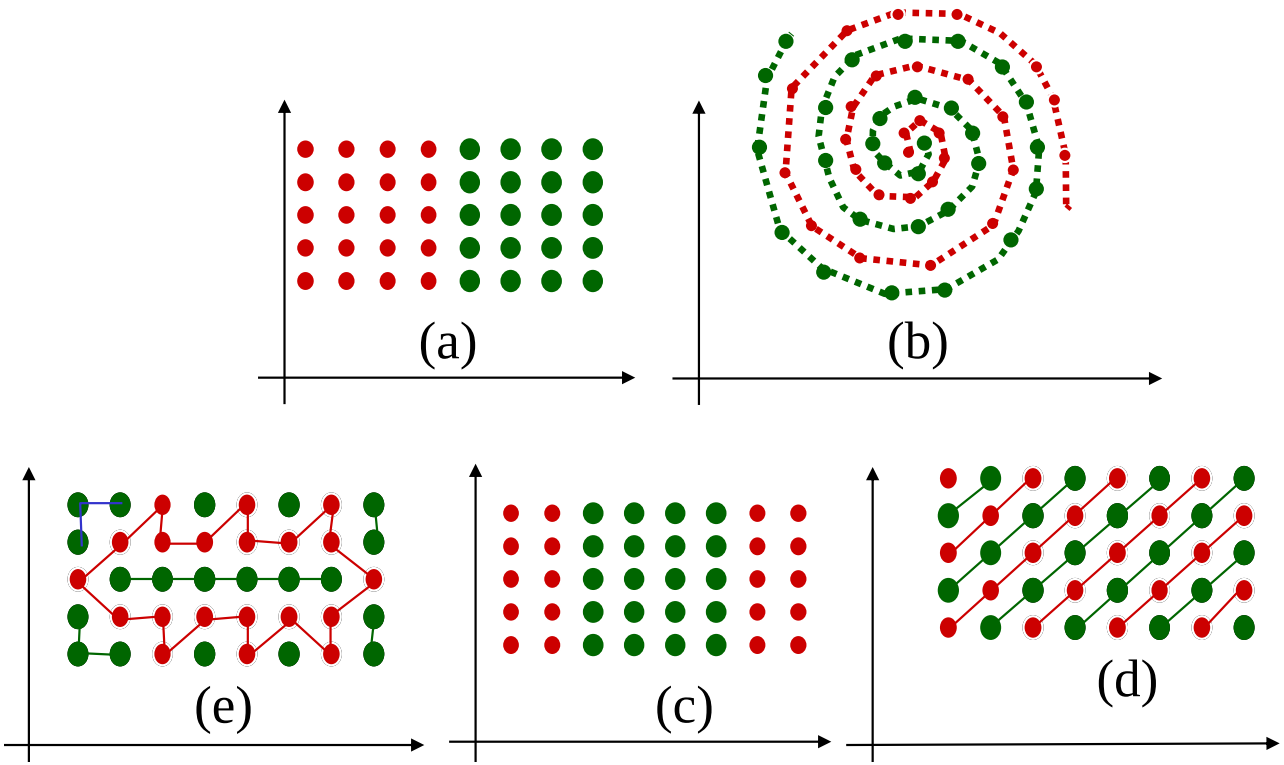
Proposition

The classes are not SEPARABLE if the learning data set in the feature space have been randomly labeled: $P(c_i/X) = P(c_i)$
Example :





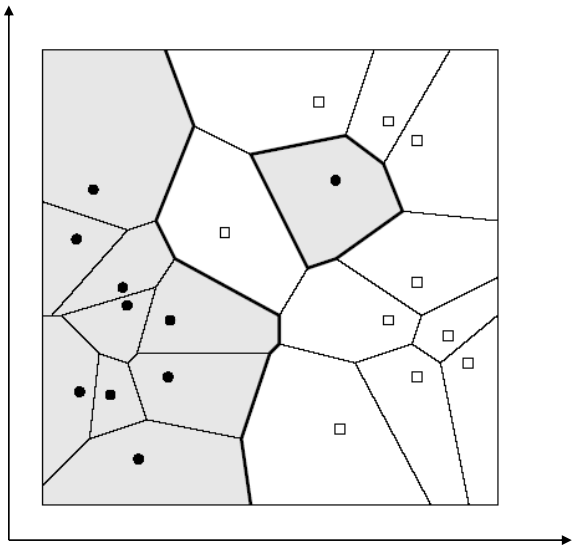
In that case, the classes are separable, therefore There exists, potentially, a machine learning algorithm capable to produce a reliable model φ, consequently, we can launch the screening process.



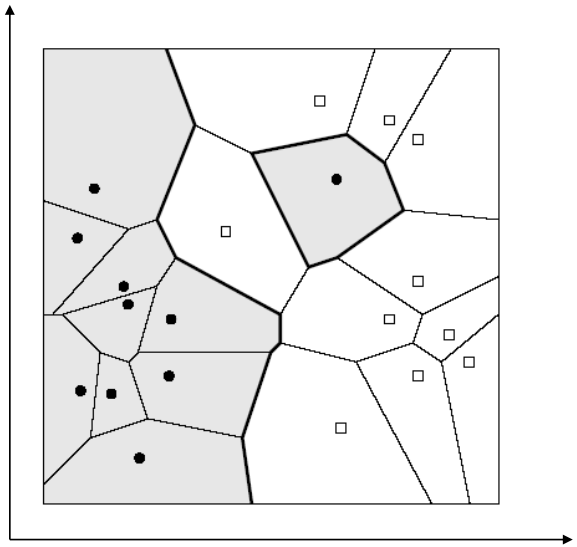
For each example, we may state that there exists an underlined model that machine learning algorithms should be able to infer.

Topological Graphs

Feature space is multidimensional: Euclidien space $R = IR^p$.
There are plenty of ways to define the topology of learning the data set.

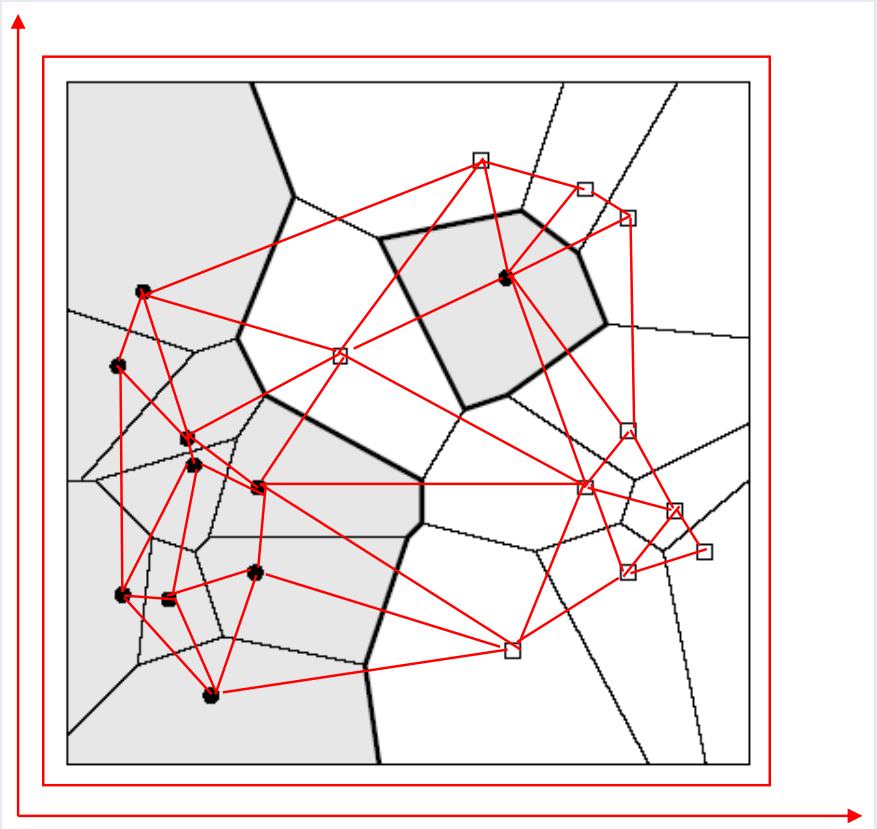


Diagram's Voronoi Topology



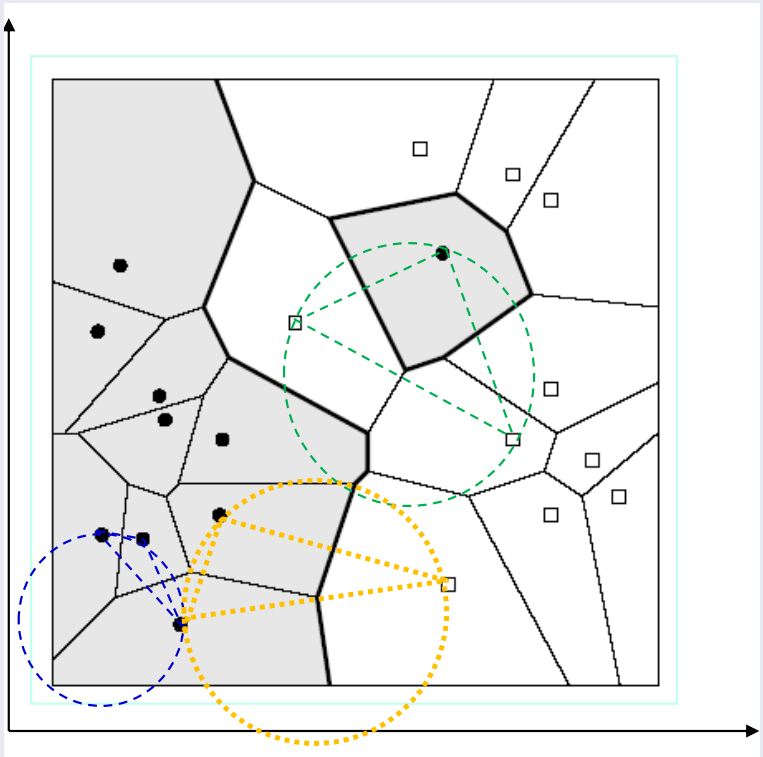
- Feature space is partitioned by the data set; each part defines the area of influence;
- Two points are neighbors if they share a common border;
- the graph brought about by the links between neighbors is the Polyhedron's Delaunay.

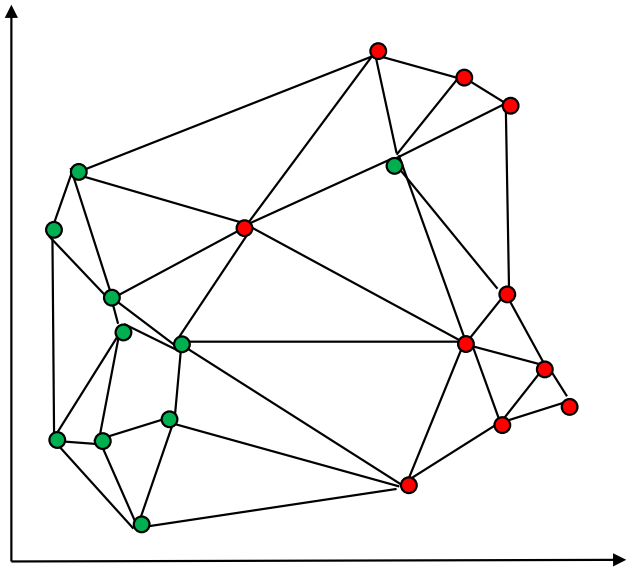
Topology of polyhedron's Delaunay



Property: all set of $P+1$ neighbors of the p-dimensional space are on tangents of an empty hypersphere.

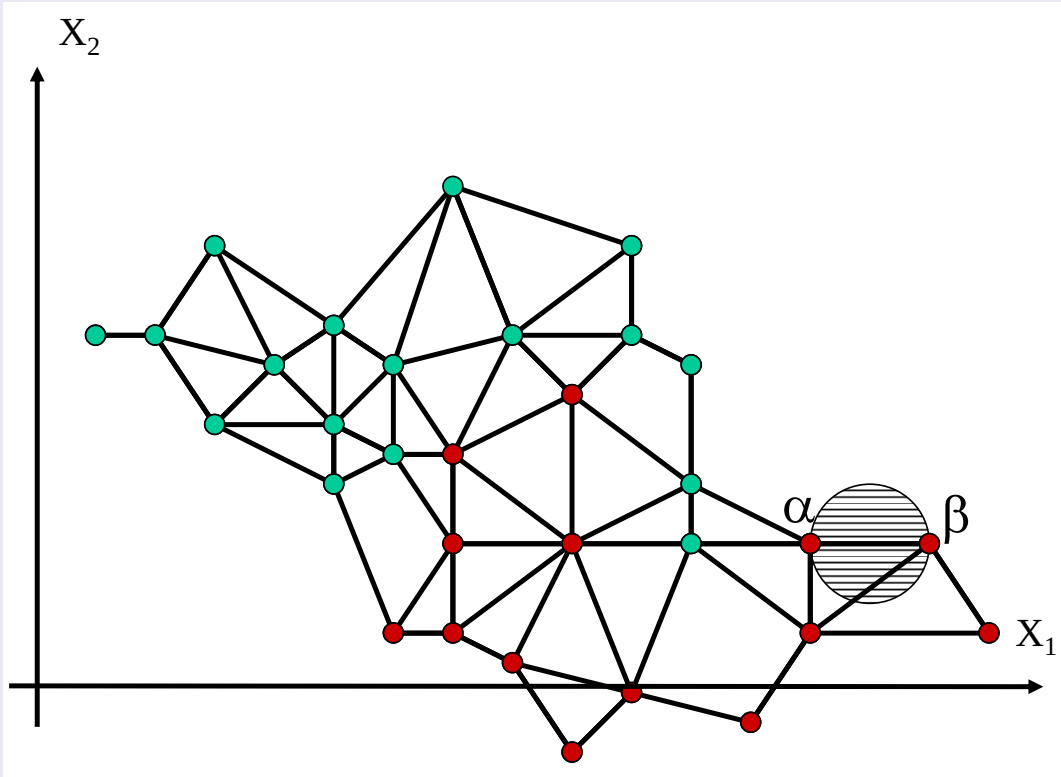
Topology of polyhedron's Delaunay





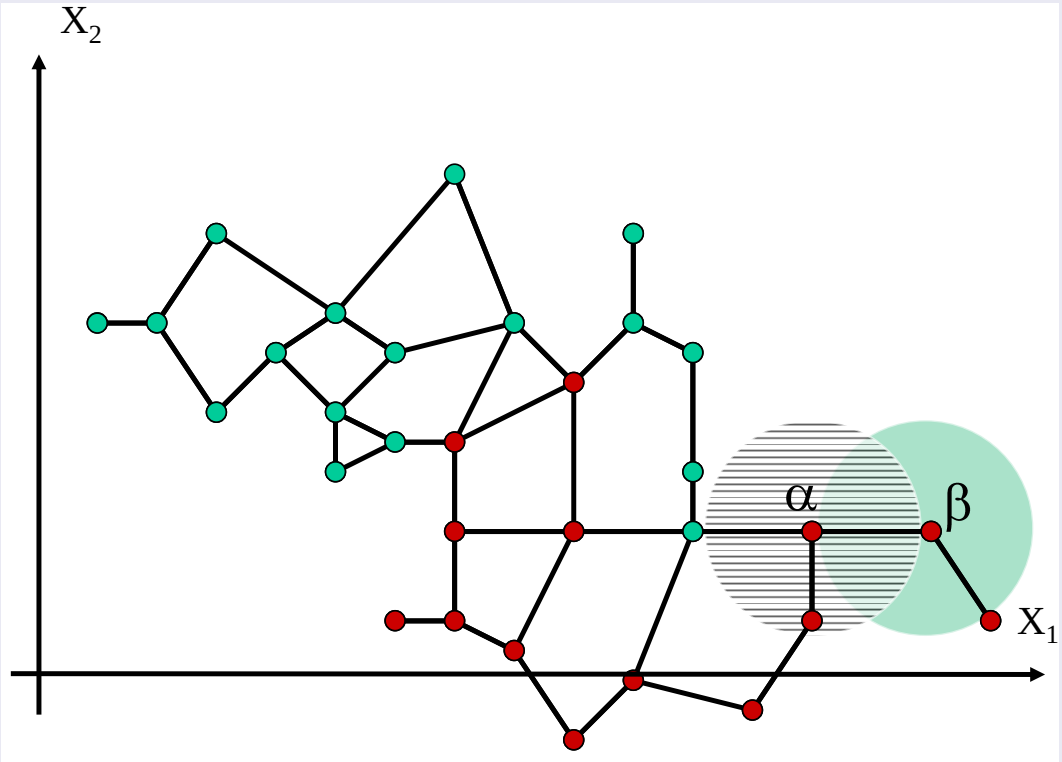
- Building Graph's Delaunay or Diagram's Vornoi is intractable in high dimension feature space
- Graph's Delaunay is a **related graph**

Gabriel Graph (GG)



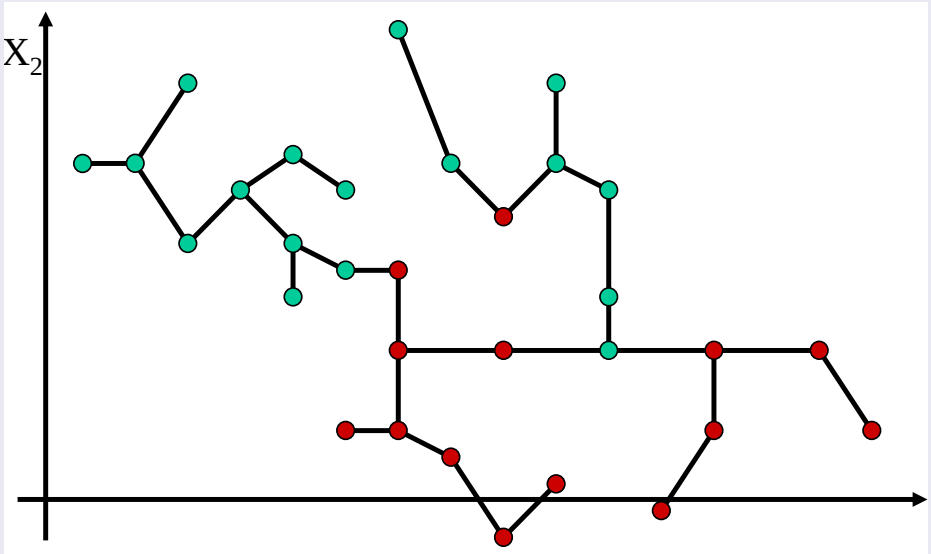
- Gabriel Graph is a **related graph**
- It feasible $O(n^2)$ even in high dimension space

Relative Neighborhood Graph (RNG)



- Relative Neighborhood Graph is a **related graph**
- RNG $\subset$ GG $\subset$ DG

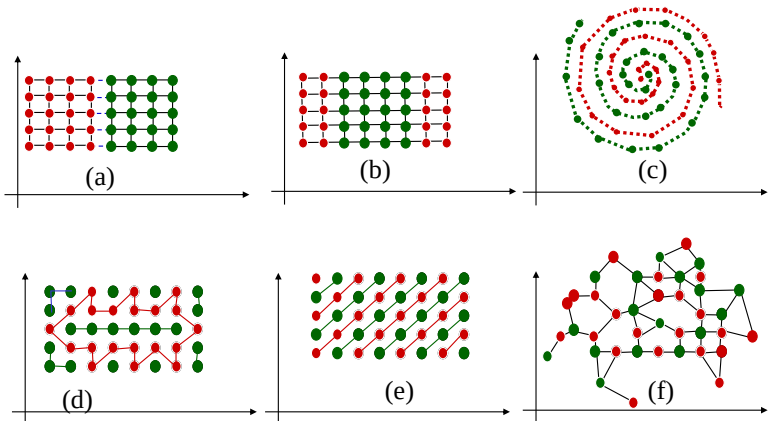
Minimum Spanning Tree (MST)



- MST is a **related graph**
- MST $\subset$ RNG $\subset$ GG $\subset$ DG

Separability of Classes

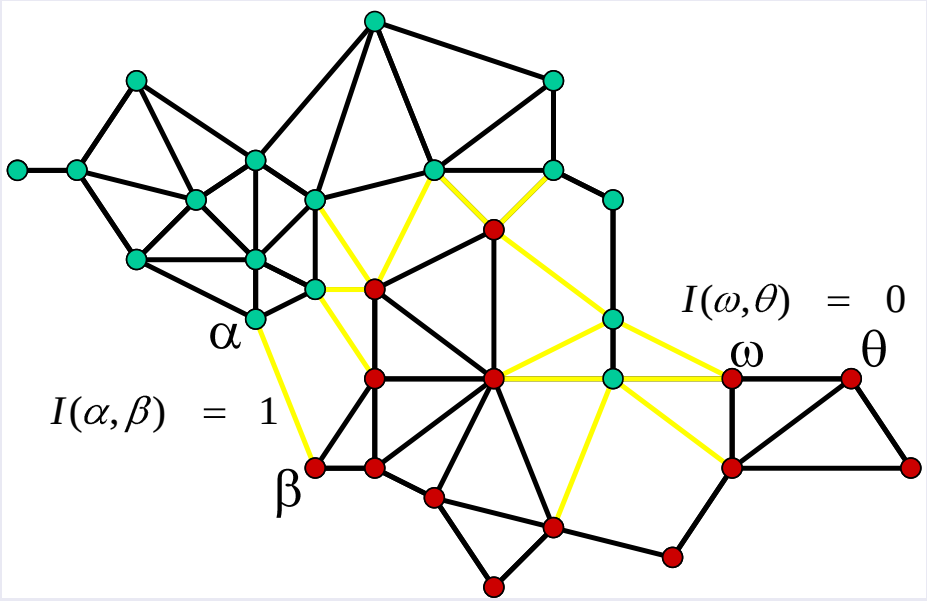
6 M.L.P. of 2 classes in IR^2 and their associated RNG.



Are those vertices of each graph have been labeled randomly ?

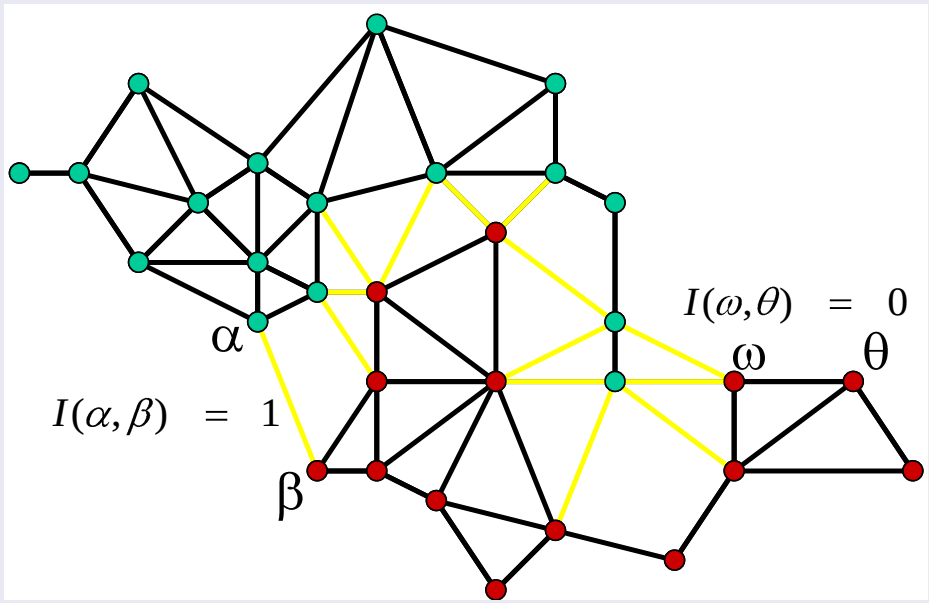
- if yes, stop there is nothing to learn !
- if not, it means that there is an underlying pattern.

Statistic of the cut edges



- $I = 14$ couples belonging to two different classes
- $J = 61$ couples belonging to the same class
- $P_J = \frac{I}{I+J} = 18,6\%$; $1 \leq P_J < 7n$

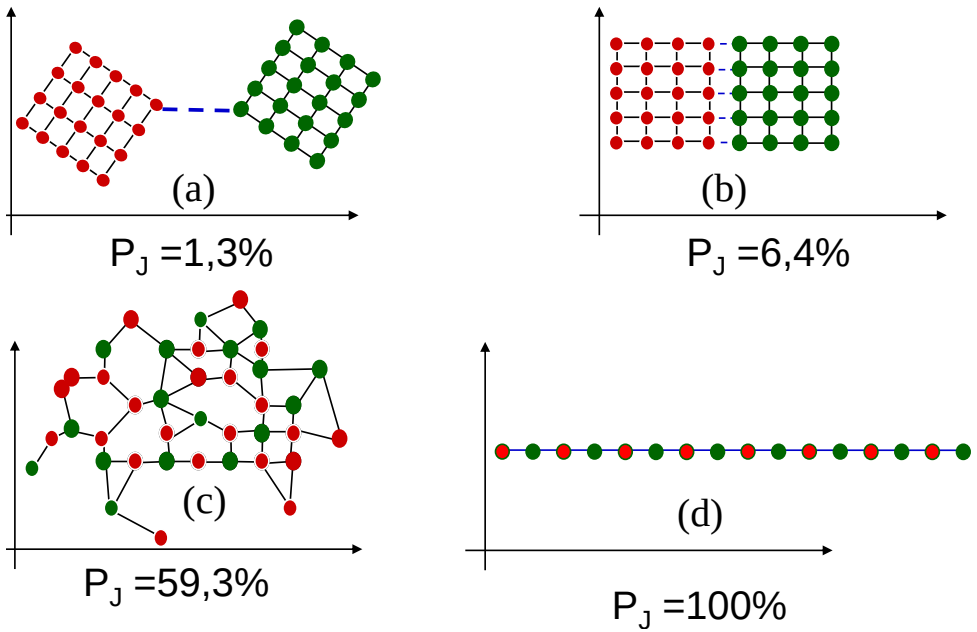
Statistic of the cut edges



- $P_J = \frac{I}{I+J} = 18,6\%$; $1 \leq P_J < 7n$

What would be this proportion in case of random labeling ?

Fewer is this proportion, better is the separability.



If this proportion was much higher than the one expected value in case of random labeling, the learning would be much harder.

Distribution of I and J under the null hypothesis

H_0: The vertices of the graph are randomly labeled according to the same probability π_k for the class k, $k = 1, \ldots, K$. We have established in

- Zighed et al. (2002) "Separability Index in Supervised Learning", LNAI 2431, pp. 475-487, .
- Zighed et al. (2005) "A statistical approach of class separability", App. Stochastic Models in Bus. and Ind., Vol. 21, No. 2, , pp. 187-197.

the law of I and J for K classes.

Boolean case

Two classes: c_1 with proportion π_1 and c_2 with π_2
Under H_0

- Mean: $m_J = S_0 \pi_1 \pi_2$
- Variance: $V_J = S_1 \pi_1{}^2 \pi_2{}^2 + S_2 \pi_1 \pi_2 (\frac{1}{4} - \pi_1 \pi_2)$

Where

- $S_0 = \sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} w_{ij}$
- $S_1 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1; j \neq i}^{n} (w_{ij} + w_{ji})^2$
- $S_2 = \sum_{i=1} n (w_{i+} + w_{+i})^2$

$w_{i+} = \sum_{j}^{n} w_{ij}$ and $w_{+i} = \sum_{i}^{n} w_{ij}$
w_{ij} is the weight of the edge (i, j) connecting vertices i and j;

Critical values of J for a threshold α_0

- $J_{\alpha_0/2} = S_0\pi_1\pi_2 - u_{1-\alpha_0/2}\sqrt{S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2(\frac{1}{4} - \pi_1\pi_2)}$
- $J_{1-\alpha_0/2} = S_0\pi_1\pi_2 + u_{1-\alpha_0/2}\sqrt{S_1\pi_1^2\pi_2^2 + S_2\pi_1\pi_2(\frac{1}{4} - \pi_1\pi_2)}$

The p-value is calculated from the normal distribution after standardisation.

Some illustrations

Breiman et al. Waves problem

Domain name	n	p	k	J / (I + J)	J^s	p-value
Waves-20	20	21	3	0.400	-0.44	0.6635
Waves-50	50	21	3	0.375	-4.05	5.0E-05
Waves-100	100	21	3	0.301	-8.44	3.3E-17
Waves-1000	1000	21	3	0.255	-42.75	0

On 13 benchmarks

- 13 benchmarks of the UCI Machine Learning Repository
- Graph: Relative Neighborhood Graph of Toussaint
- Weights: connection, distance and rank

General information					weighting: connection			weighting: distance			weighting: rank		
Domain name	n	p	k	error r.	J / (I + J)	J^s	p-value	J / (I + J)	J^s	p-value	J / (I + J)	J^s	p-value
Wine recognition	178	13	3	0.0389	0.093	-19.32	0	0.054	-19.40	0	0.074	-19.27	0
Breast Cancer	683	9	2	0.0409	0.008	-25.29	0	0.003	-24.38	0	0.014	-25.02	0
Iris (Bezdek)	150	4	3	0.0533	0.090	-16.82	0	0.077	-17.01	0	0.078	-16.78	0
Iris plants	150	4	3	0.0600	0.087	-17.22	0	0.074	-17.41	0	0.076	-17.14	0
Musk "Clean1"	476	166	2	0.0650	0.167	-17.53	0	0.115	-7.69	2E-14	0.143	-18.10	0
Image seg.	210	19	7	0.1238	0.224	-29.63	0	0.141	-29.31	0	0.201	-29.88	0
Ionosphere	351	34	2	0.1397	0.137	-11.34	0	0.046	-11.07	0	0.136	-11.33	0
Waveform	1000	21	3	0.1860	0.255	-42.75	0	0.248	-42.55	0	0.248	-42.55	0
Pima Indians	768	8	2	0.2877	0.310	-8.74	2E-18	0.282	-9.86	0	0.305	-8.93	4E-19
Glass Ident.	214	9	6	0.3169	0.356	-12.63	0	0.315	-12.90	0	0.342	-12.93	0
Haberman	306	3	2	0.3263	0.331	-1.92	0.054	0.321	-2.20	0.028	0.331	-1.90	0.058
Bupa	345	6	2	0.3632	0.401	-3.89	1E-04	0.385	-4.33	1E-05	0.394	-4.08	5E-05
Yeast	1484	8	10	0.4549	0.524	-27.03	0	0.512	-27.18	0	0.509	-28.06	0

Nb cases

Nb variables

Nb of classes

error rate on a 1-NN
(in a 10-fold cross validation)

J/(I+J): relative cut edge weight
J^s: standardized cut edge weight

On 13 benchmarks

General information					weighting: connection			weighting: distance			weighting: rank		
Domain name	n	p	k	error r.	J / (I + J)	J^s	p-value	J / (I + J)	J^s	p-value	J / (I + J)	J^s	p-value
Wine recognition	178	13	3	0.0389	0.093	-19.32	0	0.054	-19.40	0	0.074	-19.27	0
Breast Cancer	683	9	2	0.0409	0.008	-25.29	0	0.003	-24.38	0	0.014	-25.02	0
Iris (Bezdek)	150	4	3	0.0533	0.090	-16.82	0	0.077	-17.01	0	0.078	-16.78	0
Iris plants	150	4	3	0.0600	0.087	-17.22	0	0.074	-17.41	0	0.076	-17.14	0
Musk "Clean1"	476	166	2	0.0650	0.167	-17.53	0	0.115	-7.69	2E-14	0.143	-18.10	0
Image seg.	210	19	7	0.1238	0.224	-29.63	0	0.141	-29.31	0	0.201	-29.88	0
Ionosphere	351	34	2	0.1397	0.137	-11.34	0	0.046	-11.07	0	0.136	-11.33	0
Waveform	1000	21	3	0.1860	0.255	-42.75	0	0.248	-42.55	0	0.248	-42.55	0
Pima Indians	768	8	2	0.2877	0.310	-8.74	2E-18	0.282	-9.86	0	0.305	-8.93	4E-19
Glass Ident.	214	9	6	0.3169	0.356	-12.63	0	0.315	-12.90	0	0.342	-12.93	0
Haberman	306	3	2	0.3263	0.331	-1.92	0.054	0.321	-2.20	0.028	0.331	-1.90	0.058
Bupa	345	6	2	0.3632	0.401	-3.89	1E-04	0.385	-4.33	1E-05	0.394	-4.08	5E-05
Yeast	1484	8	10	0.4549	0.524	-27.03	0	0.512	-27.18	0	0.509	-28.06	0

Nb cases

Nb variables

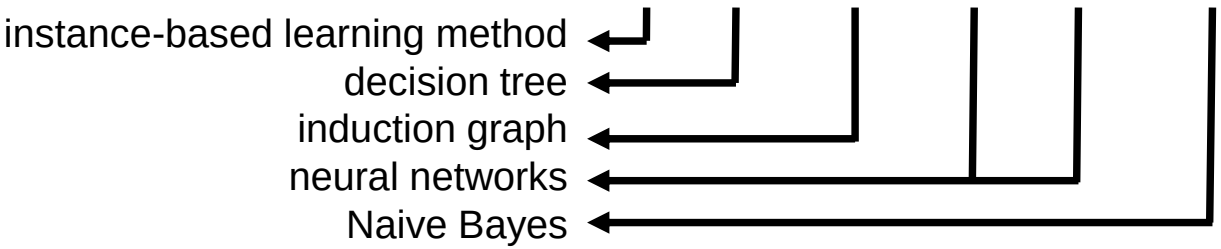
Nb of classes

error rate on a 1-NN
(in a 10-fold cross validation)

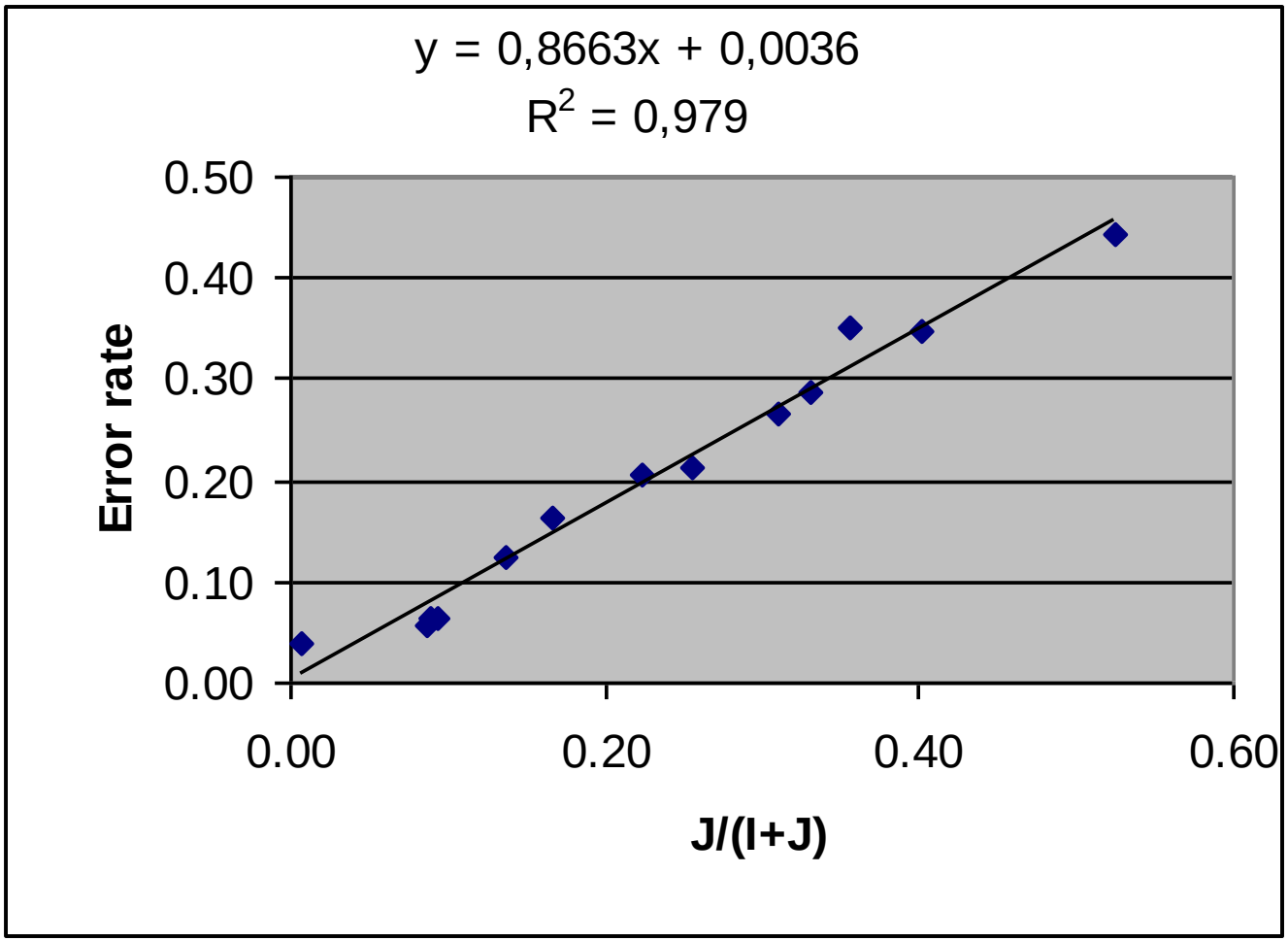
J/(I+J): relative cut edge weight
J^s: standardized cut edge weight

•Error rate in machine learning and weight of the cut edges

	General information					Statistical value			Error rate						
Domain name	**n**	**p**	**k**	**clust.**	**edges**	**J / (I + J)**	**J^s**	**p-value**	**1-NN**	**C4.5**	**Sipina**	**Perc.**	**MLP**	**N. Bayes**	**Mean**
Breast Cancer	683	9	2	10	7562	0.008	-25.29	0	0.041	0.059	0.050	0.032	0.032	0.026	0.040
BUPA liver	345	6	2	50	581	0.401	-3.89	0.0001	0.363	0.369	0.347	0.305	0.322	0.380	0.348
Glass Ident.	214	9	6	52	275	0.356	-12.63	0	0.317	0.289	0.304	0.350	0.448	0.401	0.352
Haberman	306	3	2	47	517	0.331	-1.92	0.0544	0.326	0.310	0.294	0.241	0.275	0.284	0.288
Image seg.	210	19	7	27	268	0.224	-29.63	0	0.124	0.124	0.152	0.119	0.114	0.605	0.206
Ionosphere	351	34	2	43	402	0.137	-11.34	0	0.140	0.074	0.114	0.128	0.131	0.160	0.124
Iris (Bezdek)	150	4	3	6	189	0.090	-16.82	0	0.053	0.060	0.067	0.060	0.053	0.087	0.063
Iris plants	150	4	3	6	196	0.087	-17.22	0	0.060	0.033	0.053	0.067	0.040	0.080	0.056
Musk "Clean1"	476	166	2	14	810	0.167	-17.53	0	0.065	0.162	0.232	0.187	0.113	0.227	0.164
Pima Indians	768	8	2	82	1416	0.310	-8.74	2.4E-18	0.288	0.283	0.270	0.231	0.266	0.259	0.266
Waveform	1000	21	3	49	2443	0.255	-42.75	0	0.186	0.260	0.251	0.173	0.169	0.243	0.214
Wine recognition	178	13	3	9	281	0.093	-19.32	0	0.039	0.062	0.073	0.011	0.017	0.186	0.065
Yeast	1484	8	10	401	2805	0.524	-27.03	0	0.455	0.445	0.437	0.447	0.446	0.435	0.444
								Mean	0.189	0.195	0.203	0.181	0.187	0.259	0.202
								R^2 (J/(I+J) ; error rate)	0.933	0.934	0.937	0.912	0.877	0.528	0.979
								R^2 (J^s ; error rate)	0.076	0.020	0.019	0.036	0.063	0.005	0.026



•Relative cut edge weight and mean of the error rates



Evaluation of Kernel Matrix

Kernel methods are based on

- mapping data into high dimension feature space
- using linear separator in the new feature space

The kernelisation process plays a major role in the success of learning.

kernelisation

$\varphi : X \longmapsto H$ where X is the initial feature space and H the new one. The kernel matrix K is defined by :
$K = < \varphi(x_i), \varphi(x_j) >_{i=1,...,n;j=1,...,n}$ For instance :

- $K(x_i, x_j) = exp(-\gamma \left\| x_i - x_j \right\|^2); \gamma > 0$
- $K(x_i, x_j) = (x_i x_j + 1)^p; p \in I\!N^*$
- ...

Which kernel is the best for a specific application ?

From the Kermal matrix to topological graph

Let's denote by $D(x_i, x_j)$ the original distance matrix in the original space X and $K(x_i, x_j)$ in the new feature space. To build one of the topological graphs cited previously, we just need to Know the distance matrix.

- GG_D is the Gabriel Graph derived from the distance matrix $D(x_i, x_j)$
- GG_K is the Gabriel Graph derived from the kernel Matrix $K(x_i, x_j)$

Using the statistical test of separability we have introduced before, we can assess if the kernel K leads to better feature space than the kernel K'.

Experiments

Experiment design

Data Sets	Gold Standard Error rate SVM 5 folds cross validatios			**J index of separability with GG**			Result : GG
	Linear K	Polynomial K ° 4	RBF K	Linear K	Polynomial K ° 4	RBF K	
Ionosphere		X		2	1	3	1
Heart	X			1	3	2	1
Diabetes		X		1	2	3	2
German			X	3	1	2	2
Mushrooms	X			1	3	2	1
Vehicle		X		2	1	3	1
Breast-Cancer		X		1	2	3	2
Australian	X			2	1	3	2
	Identification of the best Kernal			Rank of each Kernal			12 Score

- **J index of separability with Gabriel Graph (GG)**
- **J index of separability with MST Graph (MST)**
- **Kernal Target Alignment (KTA)***
- **Feature Space Measure (FSM)***

(*) Cristianini et al. (2001)

Results

Data Sets	Result : GG	Result : MST	Result : KTA	Result : FSM
Ionosphere	1	3	3	1
Heart	1	2	3	1
Diabetes	2	1	3	1
German	2	2	2	2
Mushrooms	1	2	1	1
Vehicle	1	3	2	2
Breast-Cancer	2	2	2	2
Australian	2	3	1	3
	12 Score	18 Score	17 Score	13 Score

We have applied this approach in various application domains :

- Muhlenbach et al. “Outlier Handling in the Neighbourhood-based Learning of a Continuous Class", LNAI Springer-Verlag, pp pp 314-321 (2004).
- Lallich et al. "Improving classification by removing or relabeling mislabeled instances", LNAI 2366 Springer-Verlag , pp.5-15,(2002).
- Scuturici et al., "Topological Query in Image Databases", Proceedings of 8th Ibero-American Congres on Pattern Recognition (CIARP 2003), Havana, Cuba. pp.144-151(2003).
- Scuturici et al., "Topological Query in Image Databases", LNCS Springer Verlag, Vol. 2905, 2004, 145-152.
- Scuturici M. et al. "Topological representation model for image databases query", Journal of Experimental and Theoretical Artificial Intelligence, Vol. 17, No. 1-2, (2005), 145-160.
- Muhlenbach F. et al., "Identifying and Handling Mislabelled Instances", Journal of Intelligent Information Systems, Vol. 22, No. 1, January (2004), 89-109.