Reconhecimento de Padrões

Artigo: Large Margin Gaussian Mixture Classifier With a Gabriel Graph Geometric Representation of Data Set Structure

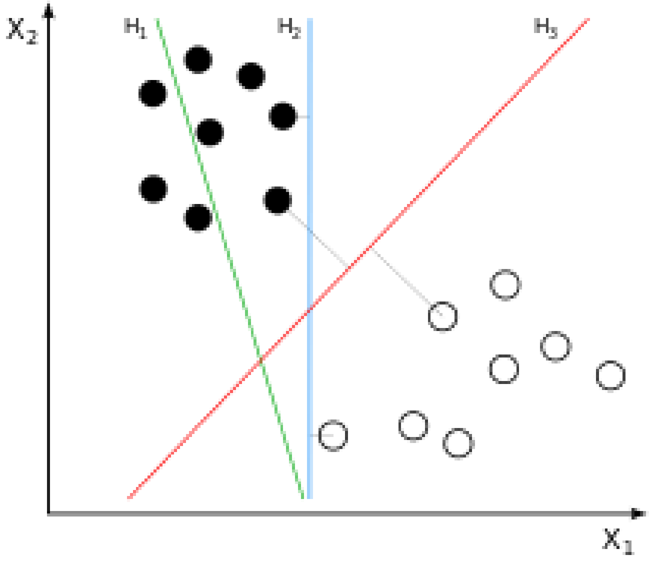
Luiz C. B. Torres , Cristiano L. Castro, Frederico Coelho , and Antônio P. Braga, Member, IEEE

Aluno: Leonam Rezende Soares de Miranda

Professor: Antônio de Pádua Braga

Support Vector Machine (SVM)

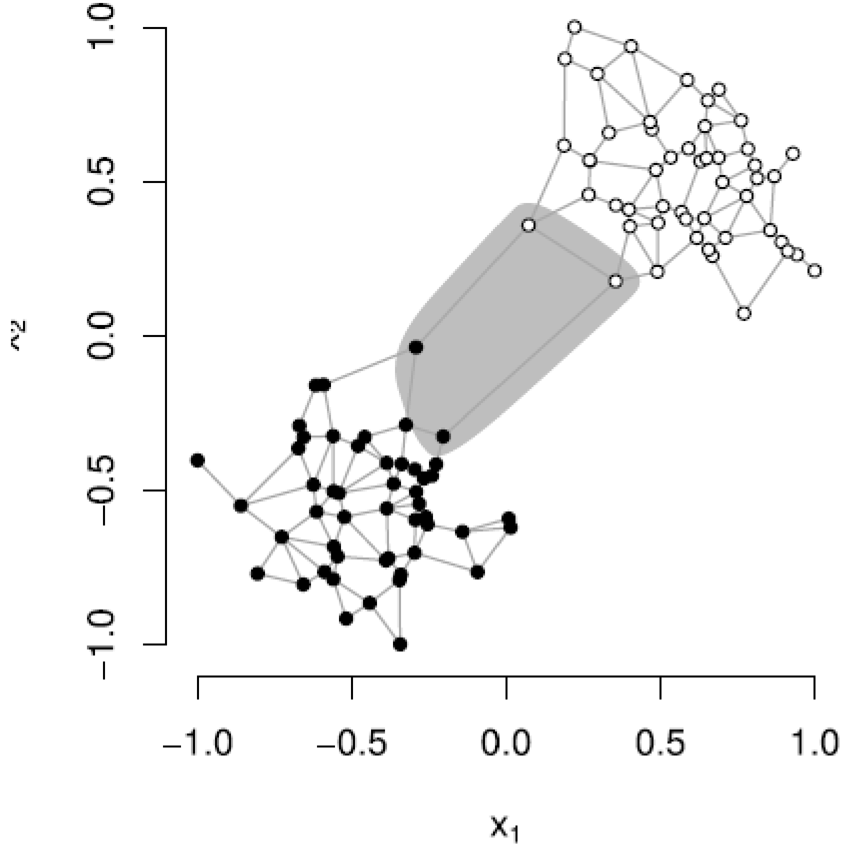
Algoritmo de aprendizado supervisionado, cujo objetivo é classificar determinado conjunto de pontos de dados que são mapeados para um espaço de características multidimensional usando uma função kernel



H_1 does not separate the classes.
H_2 does, but only with a small margin.
H_3 separates them with the maximal margin.

Introdução

- O hiperplano de margem máxima também pode ser obtido a partir da geometria do conjunto de dados;
- Algoritmo proposto não requer parâmetros do usuário e não é baseado num algoritmo de otimização.



Fonte: TORRES, L. C. B. et al

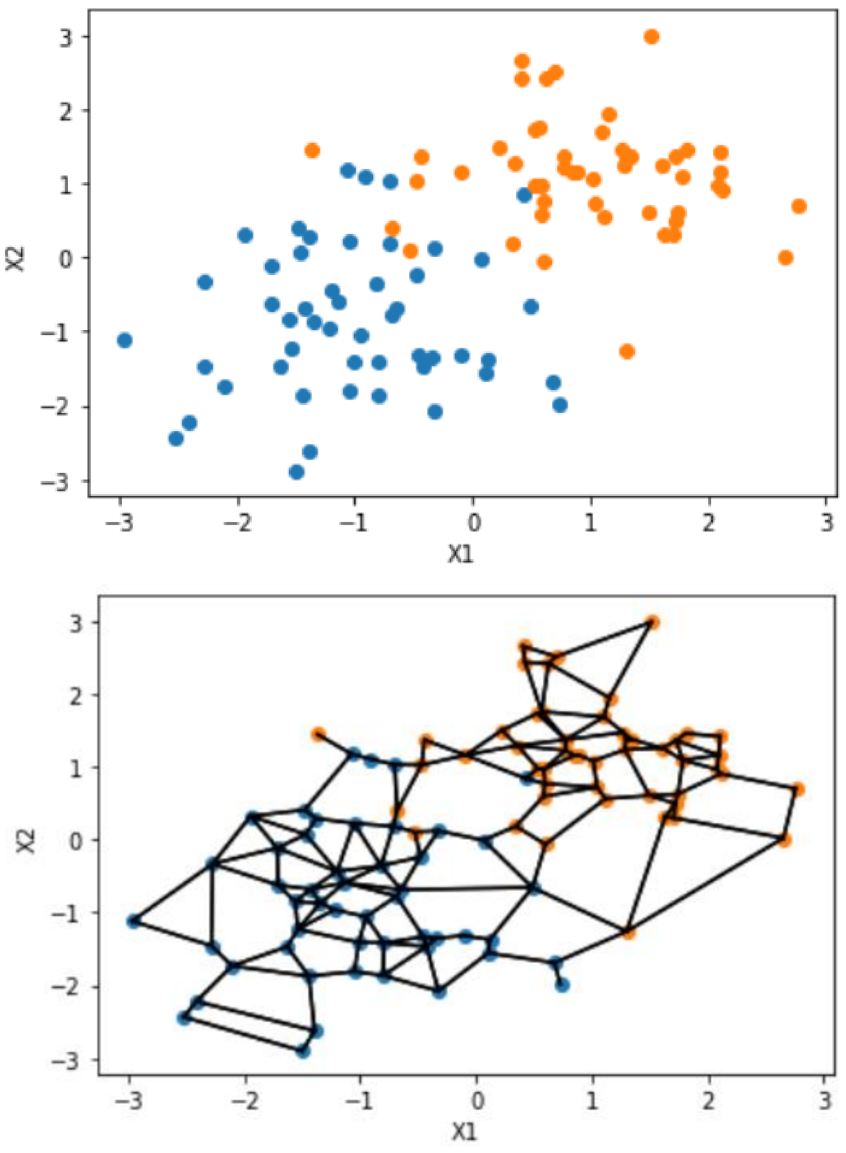
Grafo de Gabriel

A. Gabriel Graph Formulation

Considering the data set $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{C_1, C_2\}$, the Gabriel graph $\ddot{G}$ of $\mathcal{S}$ is defined as the graph with a set of vertices $\mathcal{V} = \{\mathbf{x}_i\}_{i=1}^{N}$ and edges $\mathcal{E}$ that obeys the following definition. An edge connecting the vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ from $\mathcal{V}$ belongs to $\mathcal{E}$ only, and only if

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \left(\|\mathbf{x}_i - \mathbf{x}_k\|^2 + \|\mathbf{x}_j - \mathbf{x}_k\|^2\right) \qquad (1)$$

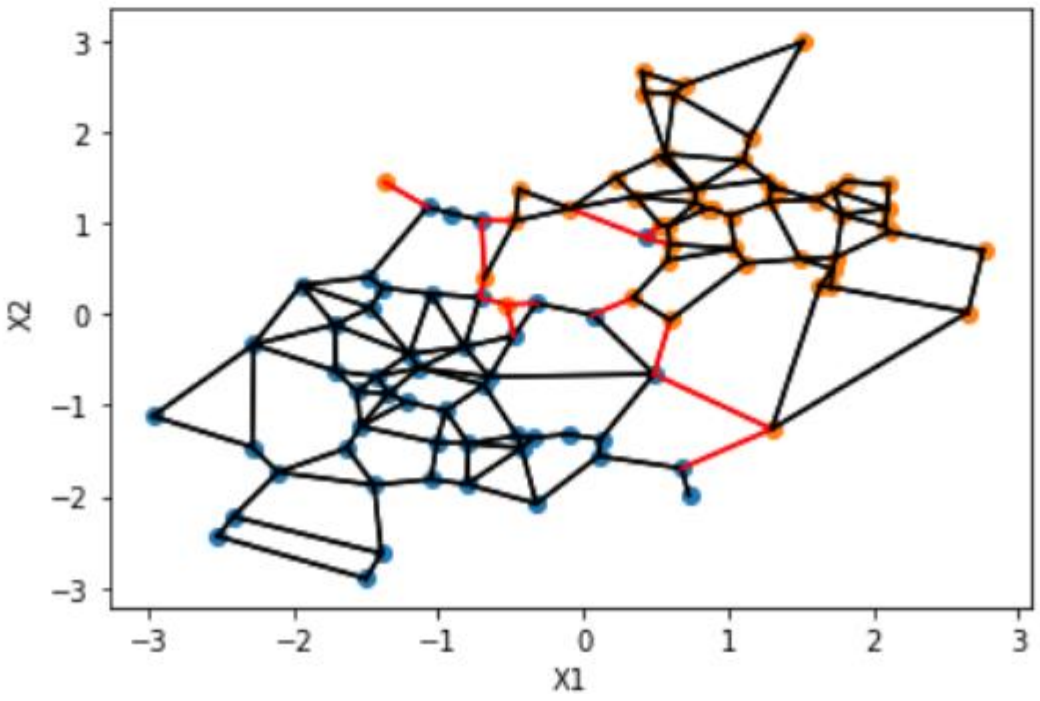
$\forall \mathbf{x}_k \in V$ and $i \neq j \neq k$, where $\|\cdot\|$ is the Euclidean distance between vertices. Fig. 2(a) shows an example of graph resulting from the previous definition.



$N_1 = N_2 = 50$

Support Edges (SEs)

São as arestas localizadas na região de separação



Class Overlapping

$$q(\mathbf{x}_i) = \frac{|\hat{\mathcal{D}}(\mathbf{x}_i)|}{|\mathcal{D}(\mathbf{x}_i)|} \qquad (2)$$

1) For all $\mathbf{x}_i \in \ddot{G}$, compute $q(\mathbf{x}_i)$ according to (2).
2) Group $q(\mathbf{x}_i)$ per class such that $\mathcal{Q}^+$ and $\mathcal{Q}^-$ holds the membership measures for the patterns with labels $+1$ and -1, respectively. In other words, $\mathcal{Q}^+$ is the set of all $q(\mathbf{x}_i)$ belonging to class $+1$ and $\mathcal{Q}^-$ for class -1.
3) Compute the class thresholds t^+ and t^- as the mean of the membership measures belonging to $\mathcal{Q}^+$ and $\mathcal{Q}^-$

$$t^+ = \frac{\sum_{q(\mathbf{x}_i)\in\mathcal{Q}^+} q(\mathbf{x}_i)}{|\mathcal{Q}^+|}, \quad t^- = \frac{\sum_{q(\mathbf{x}_i)\in\mathcal{Q}^-} q(\mathbf{x}_i)}{|\mathcal{Q}^-|}. \qquad (3)$$

4) Remove from $\ddot{G}$ all vertices whose $q(\mathbf{x}_i)$ are less than t^+ and t^-.

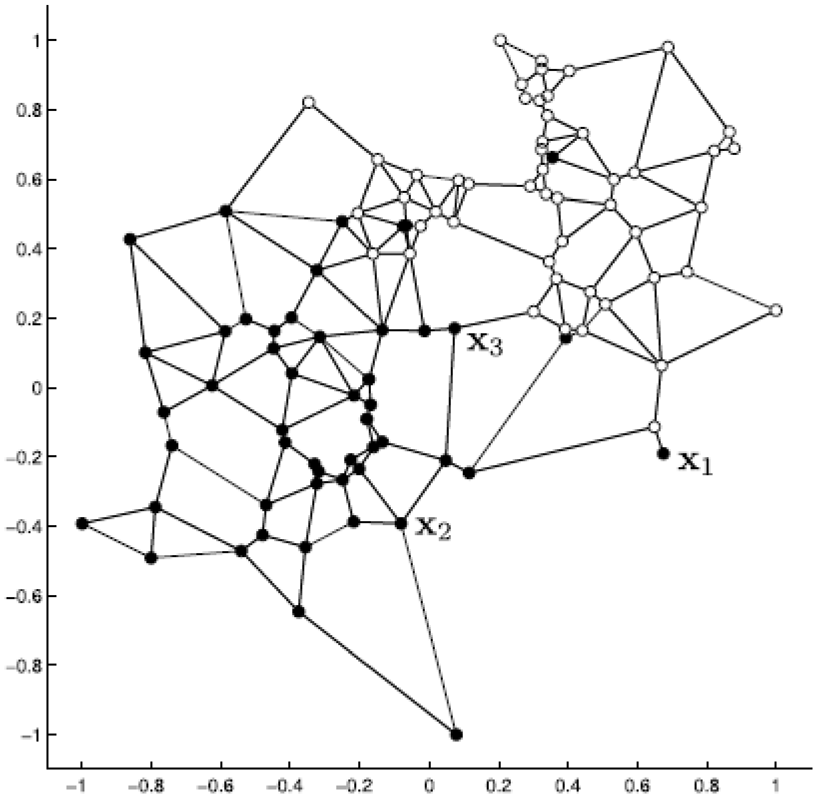
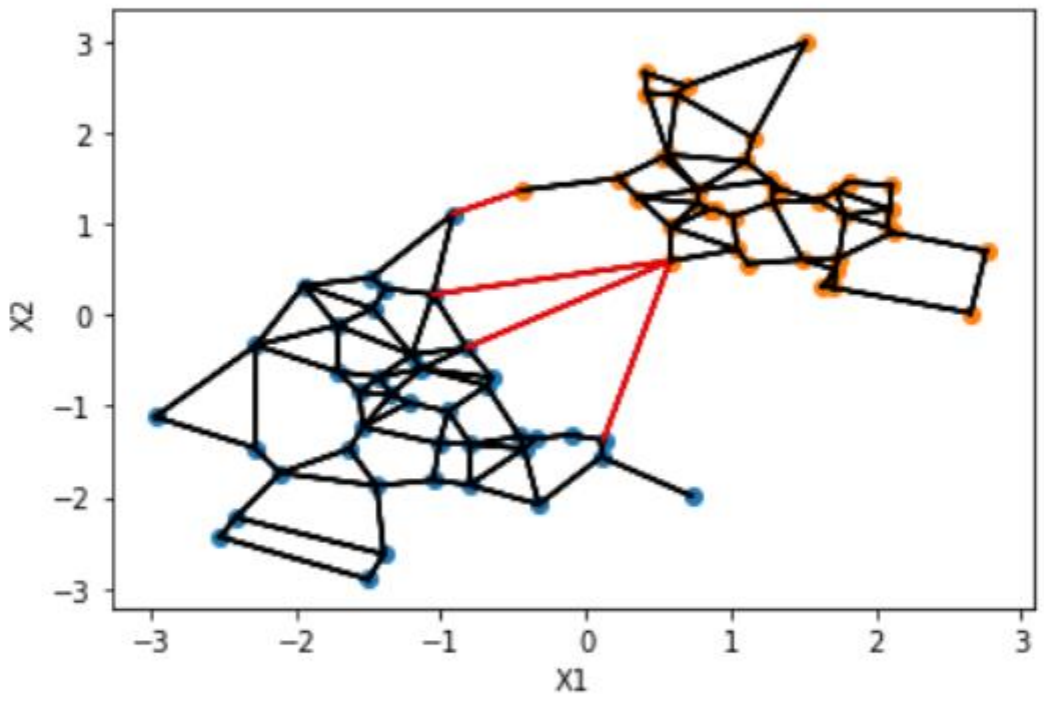


Fig. 3. Data set with overlapping.

Fonte: TORRES, L. C. B. et al.

Class Overlapping



Mistura de Gaussianas

- Cada vértice das arestas de suporte (SE) se torna o centro de uma gaussiana.
- Desvio padrão de 3σ representa 99,73% das amostras.

$$R = 3\sigma, \quad \sigma = \frac{R}{3} \tag{8}$$

$$R = \frac{1}{2}\|\mathbf{c} - \mathbf{m}\| = \frac{1}{2}\|\mathbf{d} - \mathbf{m}\| \tag{9}$$

$$\mathbf{m} = \frac{1}{2}(\mathbf{c} + \mathbf{d}), \quad (\mathbf{c}, \mathbf{d}) \in \mathcal{SE} \tag{10}$$

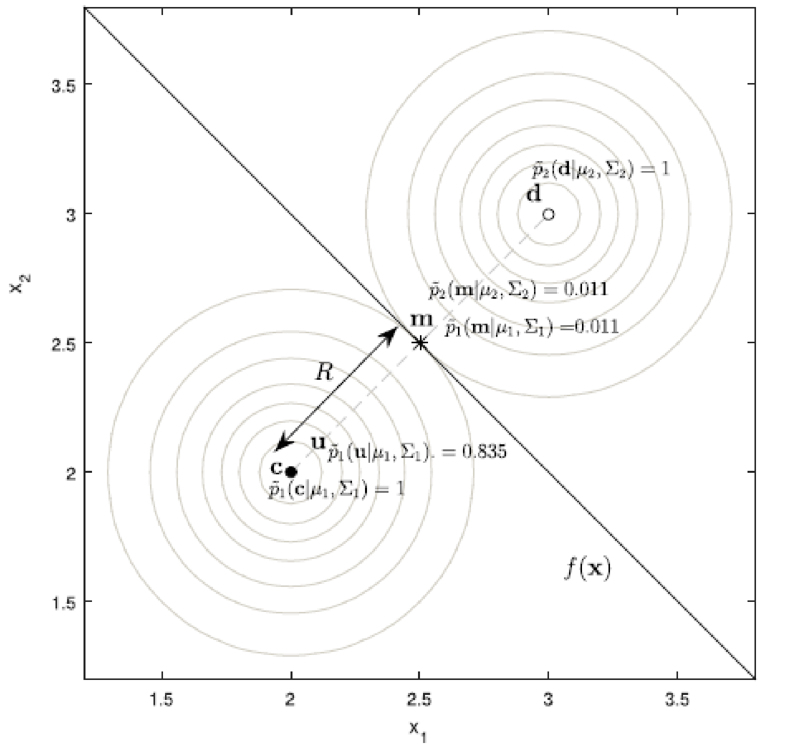


Fig. 5. Two multivariate normal distributions and a midpoint separator in the lower density region.

Fonte: TORRES, L. C. B. et al.

Mistura de Gaussianas

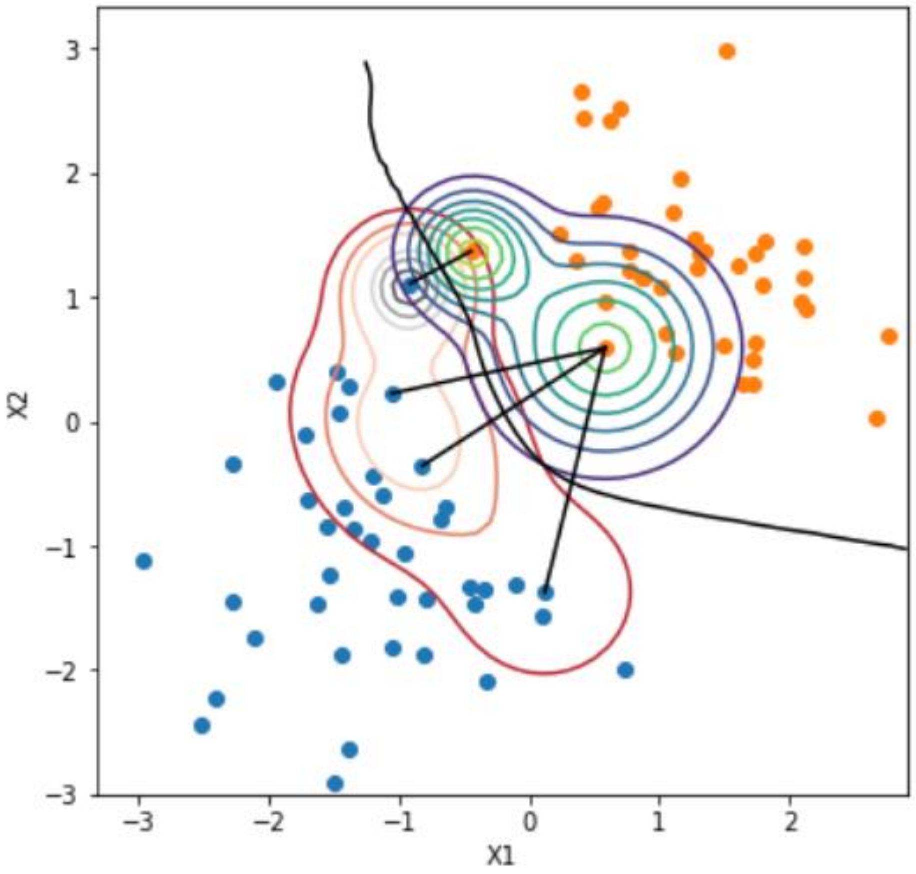
$$P(\mathbf{x}|S_1, \cdots, S_p) = \sum_{k=1}^{p} \pi_k \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_k) \right)$$

$$f(\mathbf{x}_i) = \begin{cases} +1, & \text{if } \tilde{p}(\mathbf{x}_i, \theta_1 | C_1) P(C_1) \geq \tilde{p}(\mathbf{x}_i, \theta_2 | C_2) P(C_2) \\ -1, & \text{if } \tilde{p}(\mathbf{x}_i, \theta_1 | C_1) P(C_1) < \tilde{p}(\mathbf{x}_i, \theta_2 | C_2) P(C_2) \end{cases}$$

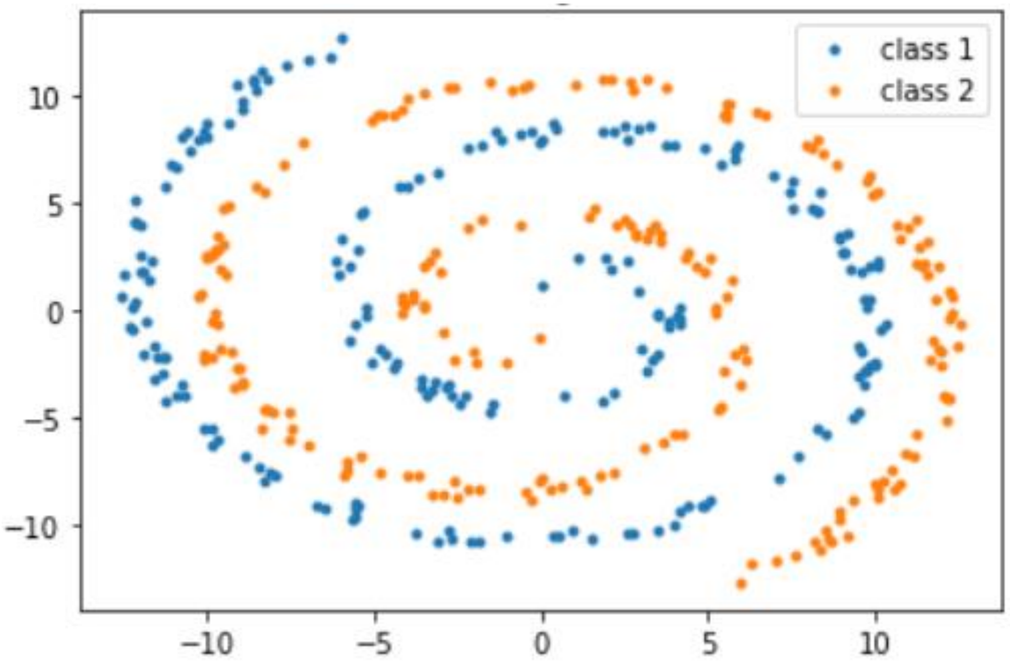
where $\tilde{p}(\mathbf{x}_i, \theta_1 | C_1)$ and $\tilde{p}(\mathbf{x}_i, \theta_1 | C_2)$ are the likelihoods of the positive and negative classes, respectively, estimated with $\tilde{\mathrm{SV}}$ only, θ_1 and θ_2 their vectors of parameters.

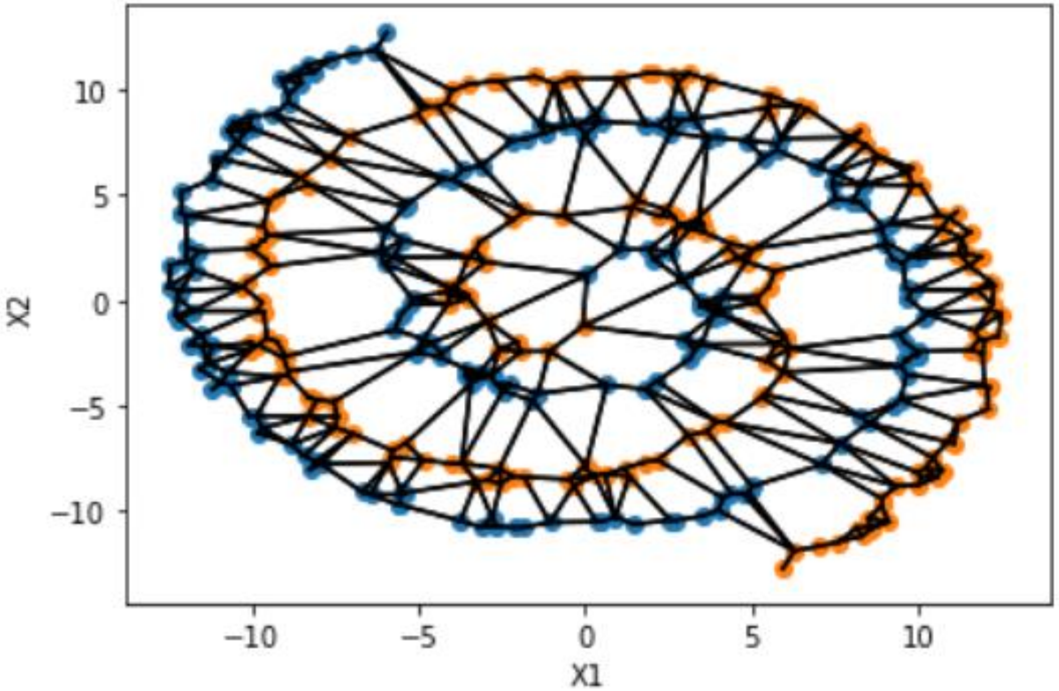
Fonte: TORRES, L. C. B. et al.

Resultados



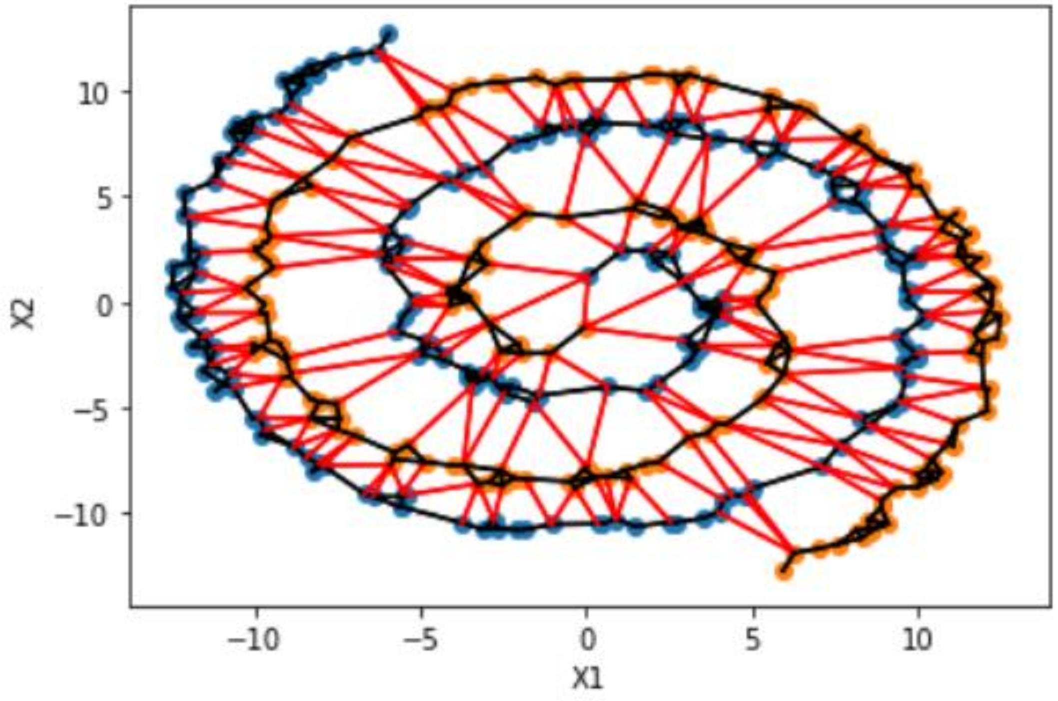
Duas Espirais

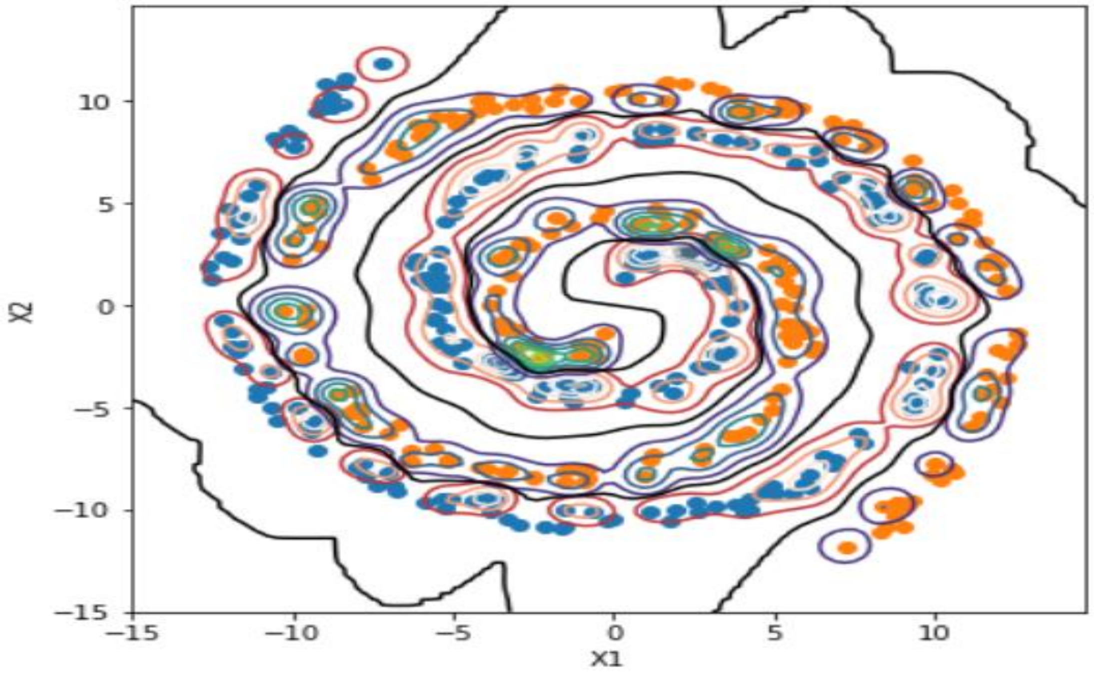




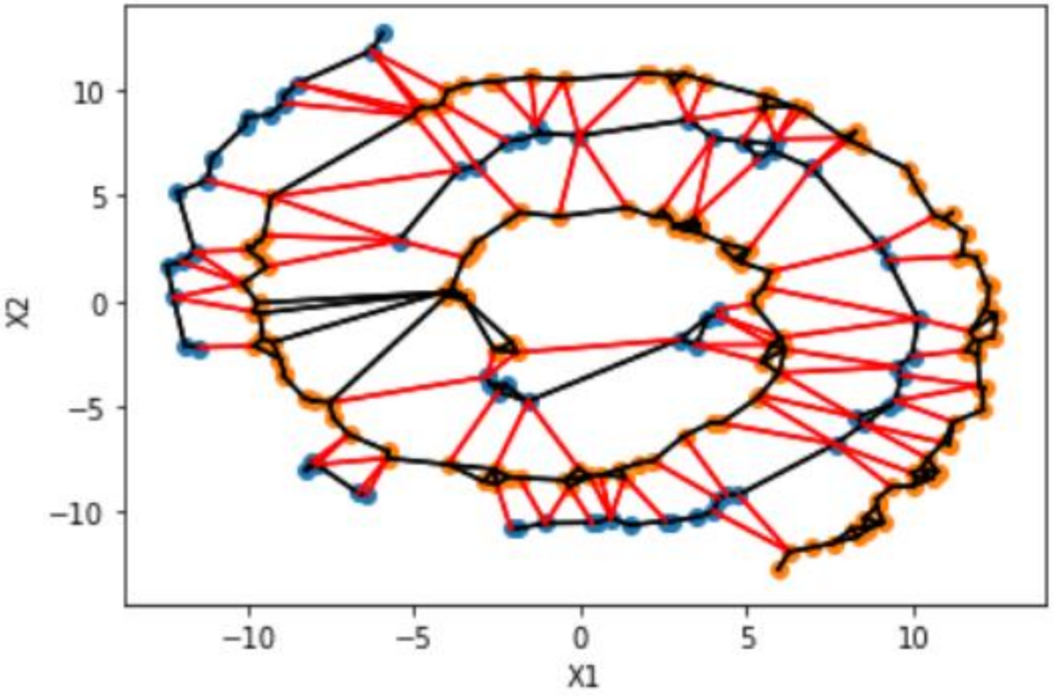
$N_1 = N_2 = 200$

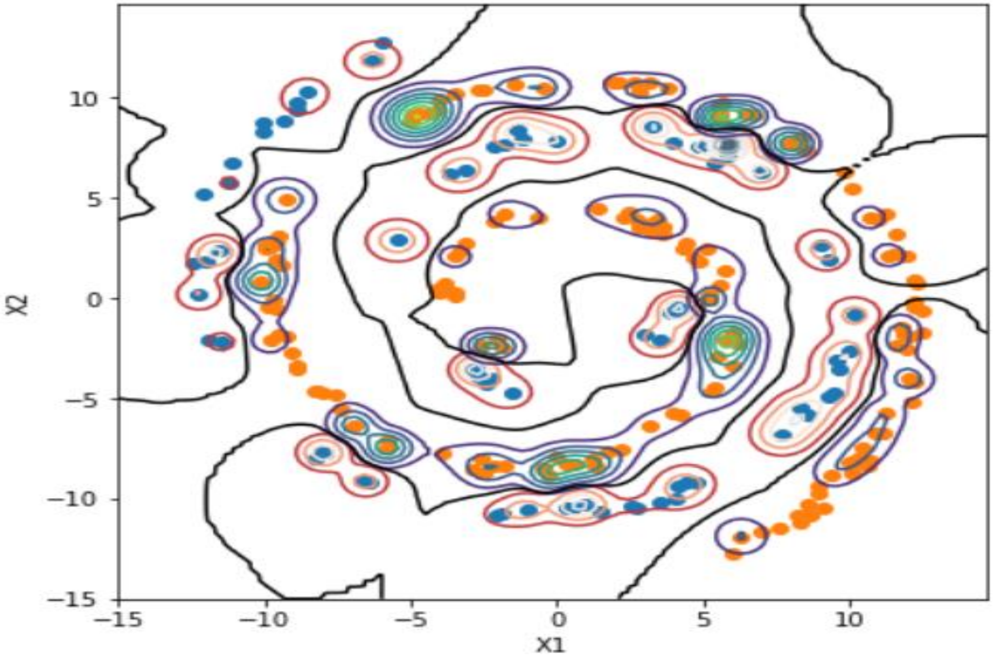
Duas Espirais





Duas Espirais – Removendo Sobreposição





Resultados

TABLE I

AVERAGE VALUES OF AUC, TRAINING TIME, AND CHARACTERISTICS OF THE DATA SETS

	New Method			**SVM-RBF**			**SVM-Poly**			**SVM-Linear**			N_d	N	N^+	N^-
Data Set	AUC	Ngv	T(s)	AUC	Nsv	T(s)	AUC	Nsv	T(s)	AUC	Nsv	T(s)				
Appendicitis	**0.792±0.165**	8	0.002	0.712±0.226	50.7	56.04	0.766±0.193	32.5	168.5	0.652±0.203	31.2	45.66	7	106	21	85
Stalog Australian Credit	0.836±0.040	251.9	0.118	0.864±0.040	312	105.2	**0.872±0.048**	298.2	571.66	0.857±0.038	198.4	63.95	14	690	307	383
Banknote Authentication	0.997±0.005	177.8	0.177	**1.000±0.000**	193.3	111.8	0.999±0.003	195.9	986.0	0.991±0.011	69.1	58.29	4	1372	610	762
The Wisconsin Breast Cancer	0.959±0.019	51.7	0.047	**0.968±0.020**	262.3	96.99	0.967±0.021	83.7	361.7	0.960±0.028	46.2	51.40	9	683	444	239
Breast Cancer Hess Probes	**0.814±0.115**	45.7	0.047	0.736±0.176	75.2	62.02	0.670±0.165	60.8	211.08	0.555±0.110	47.4	47.56	30	133	99	34
Climate Model Simulation Craches	0.704±0.173	235.2	0.195	0.510±0.032	112.3	113.0	**0.759±0.172**	85.9	364.78	0.751±0.100	56.3	53.27	18	540	494	46
Pima Indian Diabetes	**0.727±0.056**	213.5	0.067	0.717±0.065	424.3	116.6	0.706±0.052	393.2	606.25	0.717±0.050	361.5	59.72	8	768	500	268
EEG Eye State	**0.802±0.014**	4805.5	44.26	0.797±0.036	6629.2	401.9	0.643±0.062	8732.2	2494.02	0.581±0.015	11637.5	307.05	14	14980	6723	8257
Fertility	**0.643±0.282**	34.9	0.004	0.500±0	39.1	56.39	0.500±0	34.2	1.94	0.5±0	35.3	46.06	9	100	12	88
Stalog German Credit	**0.676±0.049**	459.4	0.966	0.649±0.046	564.2	202.03	0.662±0.046	516.4	1266.05	0.668±0.054	477.7	98.34	24	1000	700	300
Glass Identification	**0.924±0.106**	26.8	0.007	0.880±0.103	72.5	60.34	0.896±0.097	30.1	193.11	0.874±0.175	19.3	46.68	9	214	29	185
Haberman's Survival	0.550±0.118	56.7	0.010	0.534±0.052	165.1	65.14	0.497±0.007	147.4	249.38	0.494±0.010	149.7	48.80	3	306	225	81
Stalog Heart	0.804±0.103	95	0.032	0.828±0.075	133.9	66.15	**0.831±0.087**	140	250.50	0.824±0.097	88.9	49.37	13	270	150	120
Indian Liver Patient	**0.622±0.083**	146.2	0.035	0.498±0.011	356.6	97.71	0.497±0.008	315.9	542.03	0.499±0.004	323.3	58.26	10	579	414	165
Ionosphere	0.893±0.049	105.3	0.045	**0.938±0.039**	153.6	80.94	0.886±0.049	90.4	351.64	0.831±0.066	77.4	53.45	33	351	225	126
Parkinsons	0.792±0.125	31.5	0.008	0.790±0.151	89	63.99	**0.867±0.114**	56.9	223.69	0.753±0.063	58	48.16	22	195	147	48
Breast Cancer WP	0.566±0.162	76.9	0.036	0.493±0.015	115.5	67.38	**0.594±0.127**	92.1	257.28	0.591±0.112	78.1	54.10	33	194	46	148
Letter Recognition A Vs All	0.956±0.22	1985	123.055	0.956±0.030	391.5	2600	**0.990±0.009**	226.8	485.91	0.925±0.023	469.0	879.03	16	20000	789	19211
Mnist 0 Vs All	0.982±0.01	847.3	10516.18	**0.992±0.002**	1574.9	1160.9	**0.992±0.001**	976.2	206.86	0.967±0.007	1802.0	1679.01	40	70000	6903	63097
Staglog Shuttlest	0.962±0.02	355	2497.32	**0.998±0.001**	653.6	202.23	0.974±0.012	3165.1	148.13	0.952±0.001	4903.3	270.0124	9	58000	45586	12414
Av. Rank	1.9750			2.175			2.575			3.275						

Fontes: M. Lichman(2013),
J. Alcalá-Fdez, et al.,
K. R. Hess *et al.*,

Conclusão

- O método do artigo performa melhor em conjunto de dados menores quando comparado ao SVM. A construção do grafo de Gabriel possui complexidade $O(dn\hat{}3)$;
- Nem sempre é positivo aplicar a remoção de sobreposição, pois quando há vértices sem uma vizinhança povoada, estes são removidos;

Referências Bibliográficas

- TORRES, Luiz CB et al. Large Margin Gaussian Mixture Classifier With a Gabriel Graph Geometric Representation of Data Set Structure. IEEE Transactions on Neural Networks and Learning Systems, 2020.
- M. Lichman. (2013). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml
- J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, and S. García, “KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework,” *J. Multiple-Valued Logic Soft Comput.*, vol. 17, nos. 2–3, pp. 255–287, 2011.
- K. R. Hess *et al.*, “Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer,” *J. Clin. Oncol.*, vol. 24, no. 26, pp. 4236–4244, 2006.