Classification with stylized betweenness-relations allowing for regularization with uniform Vapnik-Chervonenkis-guarantees

Georg Schollmeyer

21.03.2019

General starting point

- Basic set V.
- Family $\mathcal{S} \subseteq 2^V$ of subsets of V.
- Aim: compute supremum type statistic

$$D^+ := \sup_{S \in \mathcal{S}} f(S)$$

where

$$f(S) = \langle w, \mathbb{1}_S \rangle$$

is a linear function in the indicator function of S.

General starting point

Exmples:

- $D^+ = \sup_{S \in \mathcal{S}} |P_n(S) - P(S)|$
- $D^+ = \sup_{S \in \mathcal{S}} = |P_n(S) - P'_n(S)|$

... Theory of uniform Glivenko-Cantelli-classes,

Vapnik-Chervonenkis theory

.... Applications:

- Generalizations of Kolmogorov-Smirnov-type statistical tests
- Subgroup discovery:

Subgroup discovery

"In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer,...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting" i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest."[Wrobel, 2001]

Subgroup discovery for supervised classification

- Given a family $\mathcal{S}$ of sets (subgroups).
- Given a point $x \in \mathbb{R}^p$ to classify.
- Define the class labels as the property of interest.
- Find that subgroup $S \in \mathcal{S}$ that is (as large and) as pure as possible w.r.t. the class labels and that contains x.
- Classify x according to the majority class label in the subgroup.

Subgroup discovery for supervised classification

Problems:

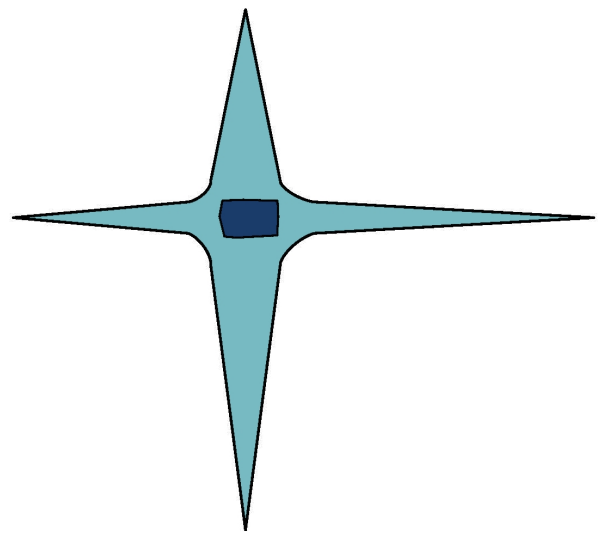
- Computationally very demanding (especially in high dimensions p).
- V.C.-dimension of $\mathcal{S}$ very high for high dimensions p.

Solution: Do simplification of computations and 'regularization' in one step by modifying the problem to 'stylized star-shaped subgroup discovery':

- Classical subgroup discovery: Subgroups $S \in \mathcal{S}$ are described by covariate characteristics, e.g.
  $S = \{x \mid x_1 \in [1, 3] \text{ and } x_5 \geq 10\}$
  $\rightsquigarrow$ high-dimensional hyper-cubes.
- Computing D^+ can be done either by pruning techniques or a **binary** programming formulation. (Note: $\mathcal{S}$ is a closure system (closed under arbitrary intersections) $\rightsquigarrow$ methods of formal concept analysis applicable.)
- Modification: Instead of the closure system $\mathcal{S}$ of all hyper-cubes, look at the 'local ring of star-shaped sets':

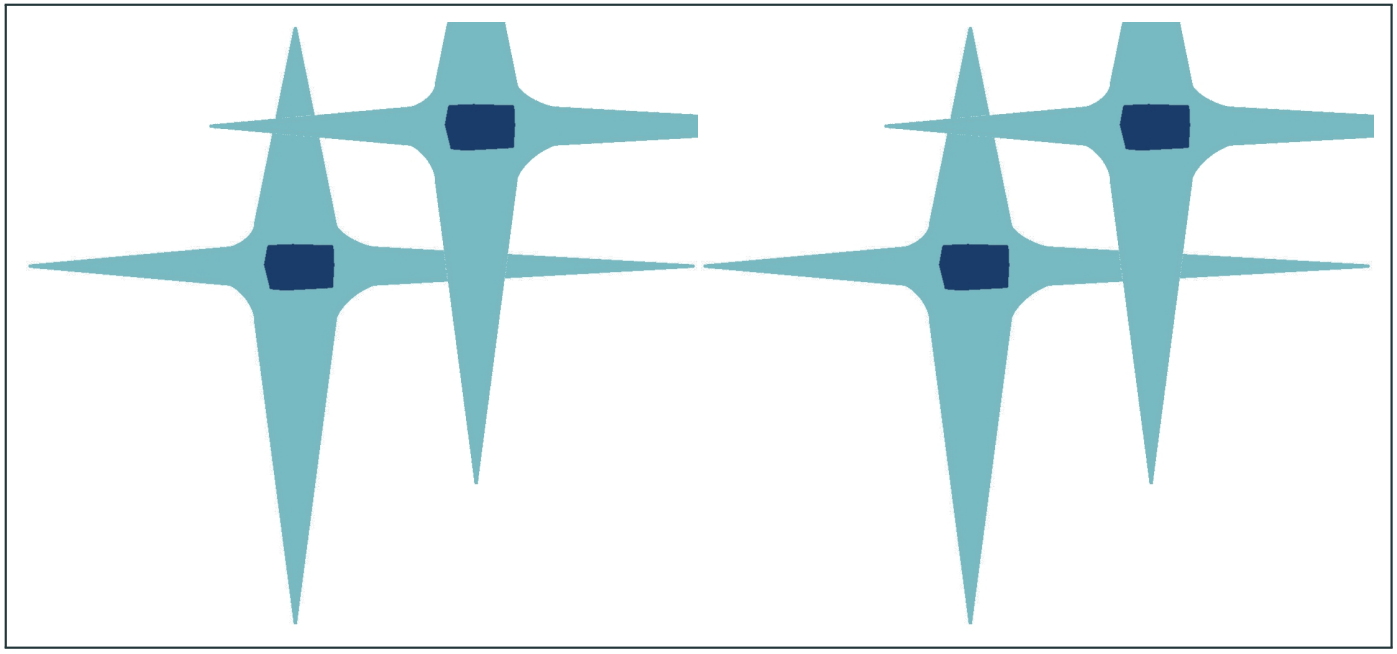
Definition (star-shaped set)

A set S in p-dimensional Euclidean space $\mathbb{R}^p$ is called star-shaped if there exists a center point $c \in S$ such that every other point $p \in S$ is visible from c, i.e., the whole line $\overline{cp}$ lies in S. In this case, any such point c is called a center point of S and the set of all center points is called the kernel of S.



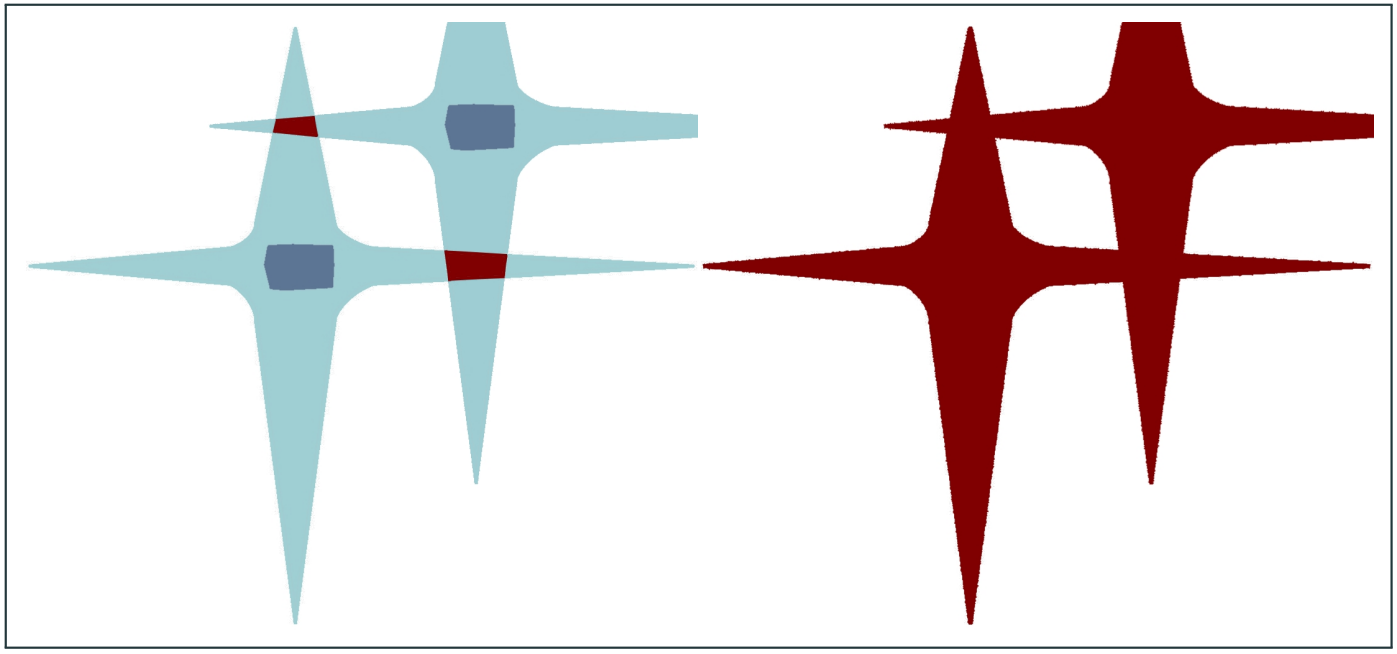
Observations

- The family of all star-shaped sets of $\mathbb{R}^d$ is generally neither closed under intersection, nor closed under union.



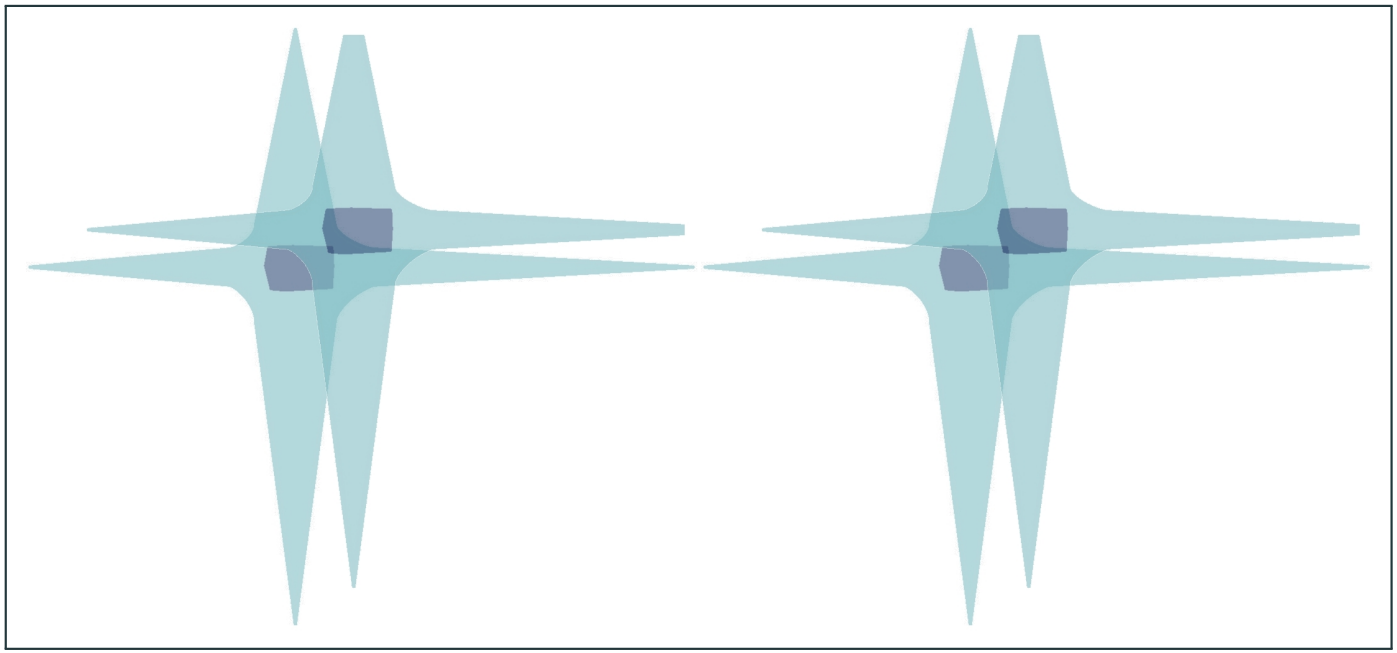
Observations

- The family of all star-shaped sets of $\mathbb{R}^d$ is generally neither closed under intersection, nor closed under union.



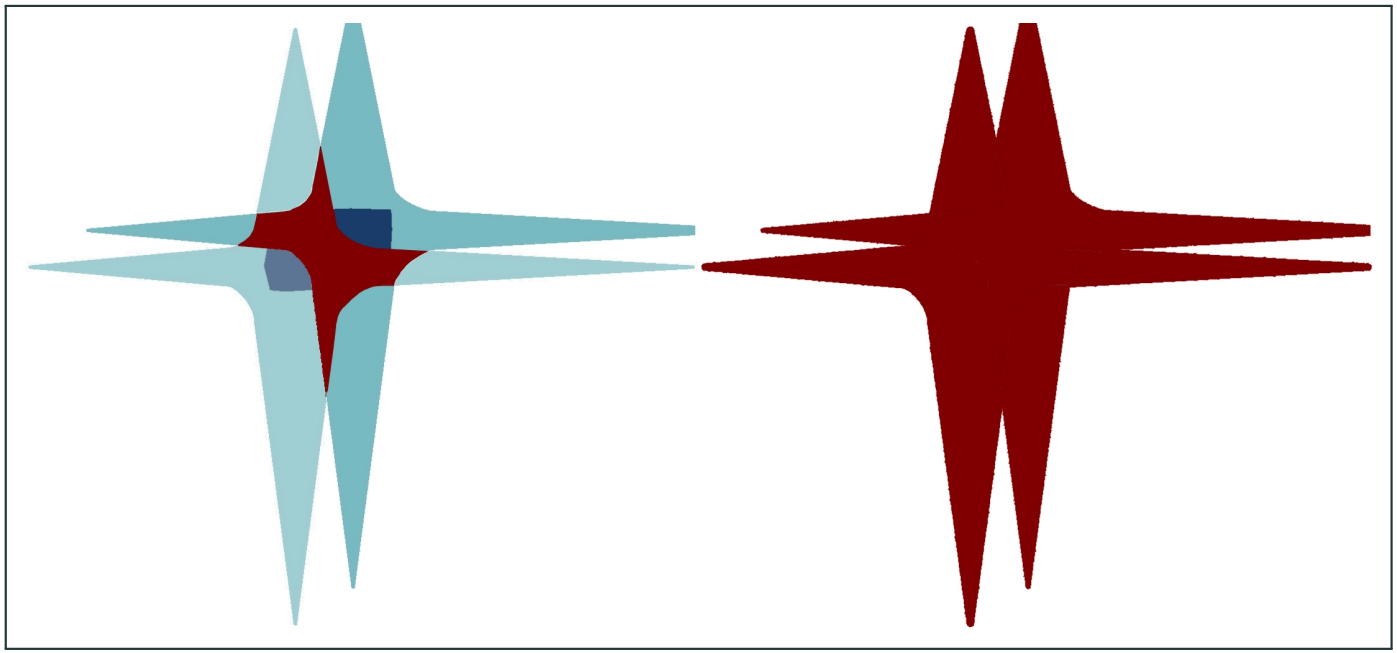
Observations

- But the family of all star-shaped sets is a 'local ring of sets': An arbitrary intersection/union of star-shaped sets with overlapping kernels is again a star-shaped set.



Observations

- But the family of all star-shaped sets is a 'local ring of sets': An arbitrary intersection/union of star-shaped sets with overlapping kernels is again a star-shaped set.



Computation of D^+

- Local ring property makes $\mathcal{S}$ easy to describe locally.
- $\rightsquigarrow$ Computation of D^+ easier by quantifying over all possible center points:
- Instead of solving one **binary** program, one only needs to solve n ordinary linear programs!

'Regularization'

- Concept of star-shaped sets uses notion of q lying between p and r, which can be formalized with a ternary betweenness-relation $B \subseteq V^3$.
- This notion is too 'narrow' (especially in high dimensions).
  $\rightsquigarrow$ Stylized notion of betweenness:

"q lies between p and r." $\rightsquigarrow$ "q lies approximately between p and r."

- Classical notion of betweenness in $\mathbb{R}^p$:

$$p, q, r \in B \iff q = \lambda p + (1 - \lambda) r \text{ for some } \lambda \in [0, 1].$$

- Stylized notion of betweenness (one possibility): $(p, q, r) \in B \iff$ the angle (p, q, r) is approximately π, say $\in [\pi - \delta, \pi + \delta]$.
  (There are many other possibilities.)

Statistical analysis

- Local analysis: Given center point c, V.C.-dimension of set of all star-shaped sets with center point c is the width of the binary relation $B(c, \cdot, \cdot)$.
- Stylization parameter δ controls the width of $B(c, \cdot, \cdot)$.
- 'Local' V.C.-dimension of $\mathcal{S}$ is thus controlled with δ.
- Global analysis: Maximally n center points: growth function is controlled.
- 'Uniform' control of the V.C.-dimension possible.
- $\rightsquigarrow$ Variation of 'local' V.C.-dimension is low.
- $\rightsquigarrow$ V.C.-entropy less dependent on P?
- How does this V.C.-analysis driven regularization compare to more classical regularization where some notion of 'smoothness of functions' is used (e.g., total variation for ternary relations)

Behavior (gene expression dataset, $n = 80,\ p = 23271$)

N	cwsb	linear svm	radial basis svm	k-nearest neighbors
10	0.21	0.27	0.46	0.45
25	0.12	0.11	0.49	0.43
50	0.08	0.05	0.21	0.40
75	0.07	0.04	0.11	0.37

Behavior (gene expression dataset showing anti-learning behavior, $n = 16,\ p = 10944$)

anti-learning:

$$\boxed{\text{training accuracy}} \geq \boxed{\text{random guessing accuracy}} \geq \boxed{\text{off-training accuracy}}$$

cwsb	linear svm	radial basis svm	k-nearest neighbors
0.33	0.69	0.72	0.44

Behavior (synthetic model of perfect antilearning plus noise, $p = 300$, cf., Kowalczyk [2007])

N	cwsb	linear svm	radial basis svm	k-nearest neighbors
10	0.01	0.64	0.65	0.99
50	0.01	0.47	0.44	0.78

References

A. Kowalczyk. Classification of anti-learnable biological and synthetic data. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, pages 176–187, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74976-9.

S. Wrobel. Inductive logic programming for knowledge discovery in databases. In *Relational data mining*, pages 74–101. Springer, 2001.