Weighted Distance Nearest Neighbor Condensing

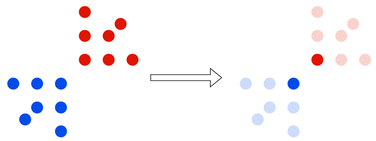
Lee-Ad Gottlieb **Timor Sharabi** **Roi Weiss**

ARIEL UNIVERSITY

ICML International Conference On Machine Learning

Nearest neighbor condensing problem

Find a minimal subset of the sample that is consistent with it, meaning that for every point of the sample, its nearest neighbor in the subset (that is the condensed set) has the same label



Weighted Nearest Neightbor Distance

Introduce the concept of weighted distance nearest neighbor condensing, which involves assigning weights to each point in the condensed set. Then, new points are labeled according to their nearest neighbor based on weighted distance in the condensed set.

$$\tilde{d}(x,x') = \frac{d(x,x')}{w(x)\cdot w(x')},$$

Decision boundary

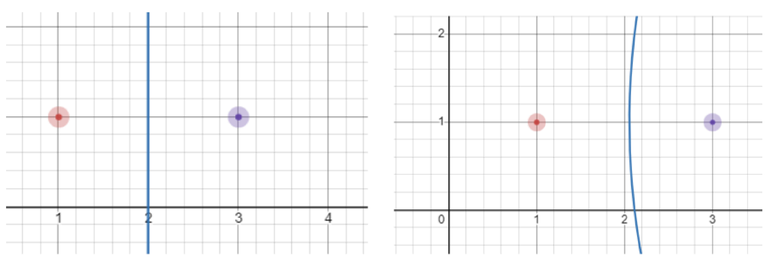
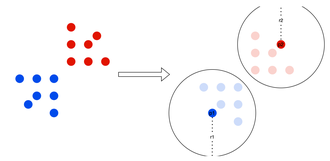


Illustration of decision boundary for equal (left) and unequal weights (right)

Ball Cover Rule

Given a subset of input points S ⊂ X, the goal is to create a subset Ś ⊂ X with the smallest size possible and assign each point xi ∈ Ś a radius ri. Each point xi must have a radius such that for any xi ∈ Ś and xj ∈ S with different labels, ri < d(xi, xj) to avoid conflicting labels within a ball. The decision rule assigns a point x ∈ S the label of the ball's center. A good BC condensing method ensures every x ∈ S \ Ś is in a ball Bi that satisfies l(x) = l(xi).

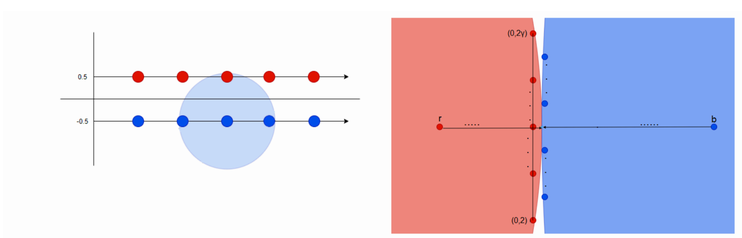


Greedy weighted heuristic

Computing BC for WNN

Input: Point set S
Initialize solution set $T \leftarrow \emptyset$, $S' \leftarrow S$, weight function $w : S \to \{1\}$
while $S' \neq \emptyset$ **do**
  $x \leftarrow \text{argmax}_{x \in S} |B(x, d_{\text{ne}}(x)) \cap S'|$
  $S' \leftarrow S' \setminus B(x, d_{\text{ne}}(x))$
  $T \leftarrow T \cup \{x\}$
  $w(x) \leftarrow d_{\text{ne}}(x)$
end while
return T, w.

Comparison Between NN and BC



A set that admits good condensing under the NN rule, but not under the BC rule. Right: A set that admits good condensing under the BC rule, but not under the NN rule.

The fraction of training samples retained in the condensed subset and the error achieved on the testing samples.

Dataset	Size	Classes	Fraction retained			Test error		
			MSS	RSS	WNN	MSS	RSS	WNN
Magic	19,020	2	0.29	0.37	**0.26**	0.22	0.26	**0.21**
SatImage	6,430	7	0.15	0.19	**0.14**	0.11	0.12	**0.09**
Spambase	4,560	2	**0.27**	0.33	**0.27**	0.21	0.21	**0.18**
Twonorm	7,400	2	0.15	0.16	**0.06**	0.06	0.11	**0.03**
Phoneme	5,404	2	0.19	0.22	**0.16**	0.13	**0.12**	**0.12**
Segment	2,310	7	0.13	0.14	**0.10**	0.07	**0.05**	**0.05**
Shuttle	43,498	7	0.030	0.008	**0.005**	0.004	**0.002**	**0.002**

Number of samples retained in condensed subset

Dataset	Points	MSS	RSS	IP	WNN
Circle	200	52	45	7	12
Banana	200	74	66	32	35
Iris	100	11	9	2	4

MSS - Modified Selective Subset - Ricardo Barandela, Francesc J. Ferri'
RSS - Relaxed Selective Subset - Alejandro Flores-Velazco
IP - Integer Programming - Brute Force
WNN - Weighted Neareset Neighboor

Check Yourself !