Nearly optimal classification for semimetrics

Lee-Ad Gottlieb, Aryeh Kontorovich, Pinhas Nisnevitch

Ariel University, Ariel, Israel Ben-Gurion University, Beer Sheva, Israel Tel-Aviv University, Tel-Aviv, Israel

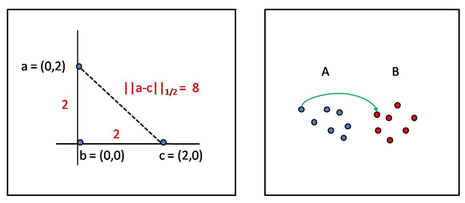
1. Semimetric spaces

Definition: A semimetric is an abstract point set equipped with a distance function that is

- Non-negative
- Symmetric
- But may not obey the triangle inequality

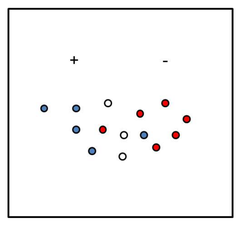
Some well-known semimetrics include

- Shannon-Jensen divergence (equivalent to ℓ_2^2)
- Fractional Minkowski norms (ℓ_p for $p < 1$)
- Hausdorff distance



2. Classification for semimetrics

Classic problem: Classify unlabeled data based on labeled database.



Possible solution? Support Vector Machines, kernels, etc.

- No!
- Points are not necessarily vectors
- Imposing a kernel on semimetrics can cause large inter-point distortion.

3. Our approach

We achieve near-optimal bounds

- Using sample compression
- Compression bounds depend on the density dimension of the space

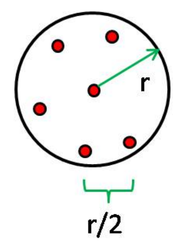
What do we mean by near-optimal?

- statistical
- computational

4. Preliminaries: Density dimension

Definition Density dimension dens:

- Semimetric M has density constant c for min c such that
- Every r radius ball in M
- Contains at most c points at mutual inter-point distance $\frac{r}{2}$
- Density dimension $\text{dens} = \log_2 c$



History

- First defined by Gottlieb and Krauthgamer (2013) for metrics
- We find that it controls learning for semimetrics

5. Application to nets

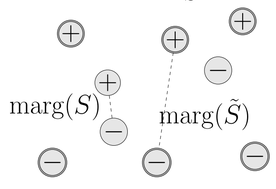
Low density dimension of semimetric M implies a small γ-net for M. And a γ-net is a consistent classifier for M.

Defintion γ-net

- Subset $S' \subset S$
- Packing: minimum inter-point distance in S' is γ
- Covering: Every point $v \in S$ satisfies $\rho(v, S') < \gamma$

Definition margin, sub-sample

- margin of $\tilde{S} \subset S$ is $\text{marg}(\tilde{S}) = \rho(\tilde{S}_+, \tilde{S}_-)$,
- sub-sample $\tilde{S} \subset S$ induces the 1-NN classifier $h_{\tilde{S}}(x) = \text{sign}(\rho(x, \tilde{S}_-) - \rho(x, \tilde{S}_+))$



Theorem. Any semimetric set $S \subset M$ admits a consistent subset of size

$$\lceil \text{radius}(S)/\text{marg}(S) \rceil^{\text{dens}(M)}.$$

This subset is exactly a $\text{marg}(S)$-net of S. It is NP-hard to find a smaller consistent subset (Gottlieb et al., 2014b).

6. Net construction algorithm

A greedy $O(n^2)$ time algorithm for constructing a net of point set S.

Brute-force net

Require: S

$p \leftarrow$ arbitrary point of S
$S' \leftarrow \{p\}$
for all $q \in S$ **do**
  if $\rho(q, S') > \text{marg}(S)$ **then**
    $S' = S' \cup \{q\}$
  end if
end for

A faster $O(n \log(\text{radius}(S)/\text{marg}(S)))$ time algorithm for constructing a consistent subset with the same size guarantee.

Fast-compression

Require: S

$p \leftarrow$ arbitrary point of S
$S' \leftarrow \{p\}$
for all $q \in S$ **do**
  if $\rho(q, S') > \frac{\text{radius}(S)}{2}$ **then**
    $S' = S' \cup \{q\}$
  end if
end for
for all $p \in S'$ **do**
  for all $q \in S$ **do**
    if $\rho(q, S') \leq \frac{\text{radius}(S)}{2}$ **then**
      $S_p = S_p \cup \{q\}$
    end if
  end for
  $S' = S' \cup$ **Fast-compression**(S_p)
end for

7. Generalization bounds

- Sample compression scheme is formalized in Graepel et al. (2005); Littlestone and Warmuth (1986); Devroye et al. (1996)
- learning algorithm maps sample S of size n to hypothesis h_S
- d-sample compression scheme if sub-sample of size d suffices to produce consistent hypothesis
- (d, ε)-compression scheme if a sub-sample of size d yields hypothesis that disagrees with the labels of at most εn of the n sample points
- sample S is (d, ε)-compressible if some the algorithm succeeds in finding an d, ε-compression scheme for S
- Theorem (fast rate, related result in Shalev-Shwartz and Ben-David (2014)):
  Fix a distribution over $\mathcal{X} \times \{-1, 1\}$, an $n \in \mathbb{N}$ and $0 < \delta < 1$. With probability at least $1 - \delta$ over the random sample S of size n: If S is (d, ε)-compressible, then

$$\text{err}(h_S) = O\left(\varepsilon + \frac{d \log n + \log(1/\delta)}{n} + \sqrt{\frac{\varepsilon d \log n + \log(1/\delta)}{n}}\right)$$

- to optimize generalization bound over ε, d is NP-hard
- for $k \in \mathbb{N}$ and $\gamma > 0$, sample S is (k, γ)-separable if has sub-sample $S' \subset S$ s.t. $|S \setminus S'| \leq k$ and $\text{marg}(S') > \gamma$
- Theorem (margin-based generalization)
  If S is (k, γ)-separable then it is $\left(\lceil \text{radius}(S)/\gamma \rceil^{\text{dens}(S)}, k/|S|\right)$-compressible.
- we can efficiently optimize resulting bound over k, γ

8. Sample complexity lower bounds

Even under margin assumptions, a sample of size exponential in dens will be required for some distributions.

Theorem. There are universal constants $c, \delta > 0$ such that for every semimetric space $(\mathcal{X}, \rho)$ with $\text{dens}(\mathcal{X}) > 6$ and any learning algorithm mapping samples S of size n to hypotheses $h_n : \mathcal{X} \to \{-1, 1\}$, there is a distribution $\mathbb{P}$ over $\mathcal{X}$ and a target concept $f : \mathcal{X} \to \{-1, 1\}$, such that $\text{err}(f) = 0$ yet

$$\mathbb{P}\left(\text{err}(h_n) \geq \frac{c \lceil \text{radius}(S)/\text{marg}(S) \rceil^{\text{dens}(\mathcal{X})}}{n}\right) \geq 1 - \delta.$$

References

L. Devroye, L. Györfi, G. Lugosi. *A probabilistic theory of pattern recognition*, 1996.

L. Gottlieb and R. Krauthgamer. Proximity algorithms for nearly doubling spaces. *SIAM J. Disc. Math.*, 27(4):1759–1769, 2013.

L. Gottlieb, A. Kontorovich, P. Nisnevitch. Near-optimal sample compression for nearest neighbors. NIPS 2014.

T. Graepel, R. Herbrich, J. Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

S. Shalev-Shwartz, S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*, 2014.

N. Littlestone M. Warmuth. Relating data compression and learnability, (unpublished) 1986.