Label Efficient Learning by Exploiting Multi-class Output Codes

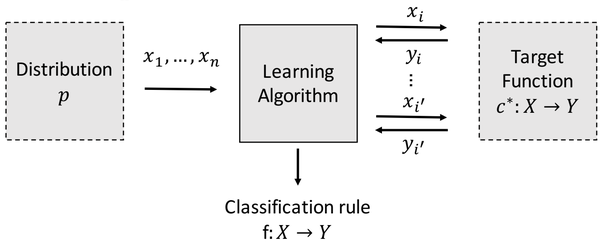
Maria-Florina Balcan, Travis Dick, Yishay Mansour

Overview and Background

Overview

- Data efficient algorithms for classification should minimize the amount of *labeled* data that is required, since most modern classification tasks have an abundance of cheap unlabeled data, but annotating it is relatively expensive.
- This is especially true for problems with many classes, since we require labeled examples for all classes.
- We show that by making the implicit assumptions of output-coding explicit, we can more fully exploit them when learning from limited labeled data.
- Our main assumption is that there exists a low-error (unknown) output-code classifier.

Learning Model



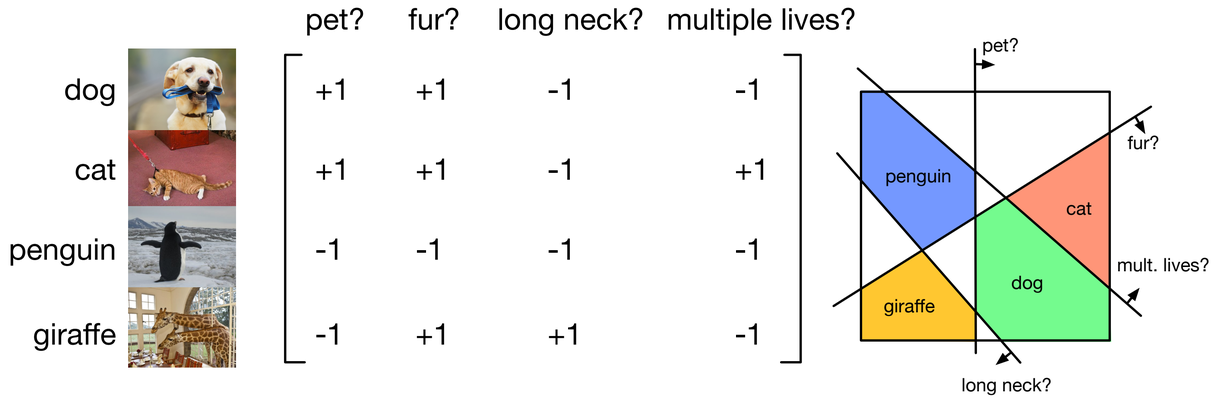
- Given iid sample $S = \{x_1, \dots, x_n\} \sim p$.
- Can query the label $c^*(x)$ for any $x \in S$.
- Goal: minimize $\Pr_{x \sim p}(h(x) \neq c^*(x))$ and number of label queries.

Linear Output-code Classifiers

- Reduction from multi-class to binary learning.
- Design m binary partitions of the L classes (a code matrix $C \in \{\pm 1\}^{L \times m}$).
- Learn a linear classifier $h_i(x) = w_i^\top x - b_i$ for each partition.
- Output

$$f(x) = \operatorname*{argmin}_{1 \leq i \leq L} \mathrm{d}_{\mathrm{Ham}}(\langle h_1(x), \dots, h_m(x) \rangle, C_i),$$

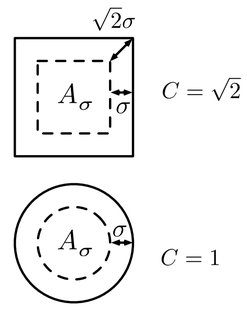
where C_i is the i^{th} row of C.



Error Correcting Output-codes

1. Output-code makes at most β mistakes when predicting codewords.
2. Codewords have Hamming distance at least $2\beta + d + 1$.
3. The data distribution satisfies the following thick level set condition:

   p has thick level sets up to level λ_0 with parameters (C, σ_0) if: $\forall \lambda \leq \lambda_0$, $\forall \sigma \leq \sigma_0$, the set $A_\sigma = \{x : B(x, \sigma) \subset \{p \geq \lambda\}\}$ is nonempty and $\sup_{x \in \{p \geq \lambda\}} d(x, A_\sigma) \leq C\sigma$.

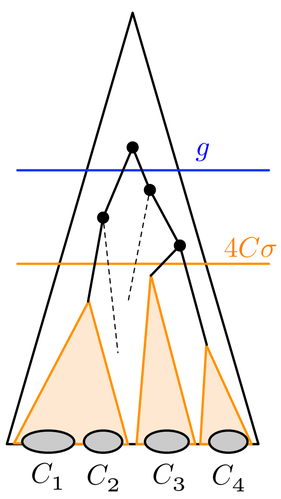


1. Compute the single linkage hierarchical clustering. **2.** Query the labels of m randomly chosen points. **3.** Find the coarsest pruning with each cluster containing exactly one label.

4. Output classifier that assigns point x to the label of the nearest cluster.

Theorem. *Fix $\varepsilon, \delta > 0$, set $\lambda = \varepsilon / \mathrm{Vol}(X)$ and $\sigma = g/4C$. If $\mathrm{int}_\sigma\{p \geq \lambda\}$ has N connected components each with prob. mass $\geq \gamma$, running the above algorithm with $m = O(\frac{2}{\gamma} \ln \frac{N}{\delta})$ and $n = \tilde{O}(\frac{1}{(\varepsilon\sigma)^2})$ will have error $\leq \varepsilon$ w.p. $1 - \delta$.*

Main Ideas:

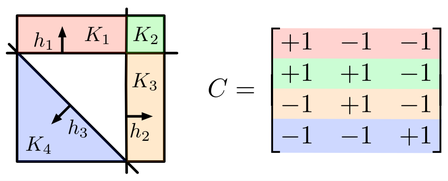


- Separation: if $f(x) \neq f(x')$, $[x, x']$ hits $\geq d + 1$ hyperplanes.
- $\{h_i\}$ in general position: $\exists g > 0$ s.t. if $f(x) \neq f(x')$, $\|x - x'\| \geq g$.
- So our algorithm outputs a coarsification of the dist.-g pruning.
- Let $A_1, \dots, A_N$ be the CCs of $\mathrm{int}_\sigma\{p \geq \lambda\}$ and define $C_i = \{x : d(x, A_i) \leq C\sigma\}$ for $i = 1, \dots, N$.
- Thick level sets guarantee that for our value of n, all samples in C_i will be connected in the dist.-$4C\sigma$ pruning.
- Choose $\sigma = g/(4C)$.
- For our choice of m, w.h.p. every C_i set contains at least one labeled example, so we never make mistakes on C_i sets.
- A new point $x \sim p$ lands in $\bigcup_i C_i$ w.p. $\geq 1 - \varepsilon$.

The Boundary Features Condition

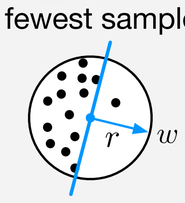
1. Let $K_i = \{x : \langle h_1(x), \dots, h_m(x) \rangle = C_i\}$ for each label $i \in [L]$.
2. *For every column j of C, there is a row i such that negating C_{ij} produces a row not present in C, and the corresponding partition cell is non-empty.*

   (Equivalently: every h_j forms a face of one K_i not shared by another $K_{i'}$).
3. $\exists R > 0$ such that: (i) The "unshared faces" have length at least R and (ii) Non-class cells are at least distance R apart.
4. The data is nearly uniform on the set $K = \bigcup_{i=1}^{L} K_i$ (See one-vs-all for definition).

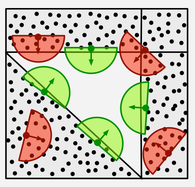


$$C = \begin{bmatrix} +1 & -1 & -1 \\ +1 & +1 & -1 \\ -1 & +1 & -1 \\ -1 & -1 & +1 \end{bmatrix}$$

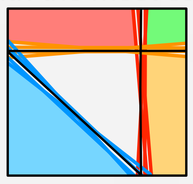
1. For each sample x, find the half-ball of radius r containing the fewest samples



2. If it contains fewer than τn samples, add (x, w) to the set H



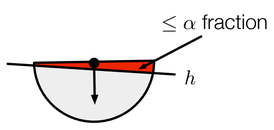
3. The hyperplanes $\{x' : w^\top(x' - x) = 0, (x, w) \in H\}$ partition the input space. Query label of the L cells with the most samples.



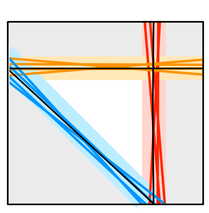
4. If x belongs to a labeled cell, output that label, otherwise guess.

Theorem. *For any $\varepsilon, \delta > 0$, running the above algorithm with $r = R/2$, $\tau = \frac{1}{4}\alpha c_{lb} r^d v_d$, $n = \tilde{O}(\frac{m^2 c_{ub}^2 D^d}{\varepsilon^4 R^d})$, where D is the diameter of $\mathcal{X}$, will have error $\leq \varepsilon$ w.p. $1 - \delta$.*

Main Ideas:



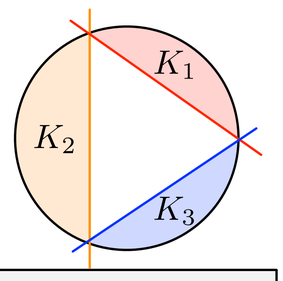
- A half-ball $B^{1/2}$ α-approximates a hyperplane h if most of the ball's volume is separated from its center by h.
- W.h.p., every half-ball that passes step 2 is an α-approximation to one of $h_1, \dots, h_m$, and every hyperplane is α-approximated.
- All α-approximations to h_i agree except on a margin of size $O(\sqrt{\alpha} D)$.
- Each margin has $\leq c_{\mathrm{ub}} \cdot \mathrm{Vol}(\mathrm{margin} \cap \mathcal{X})$ prob. mass.



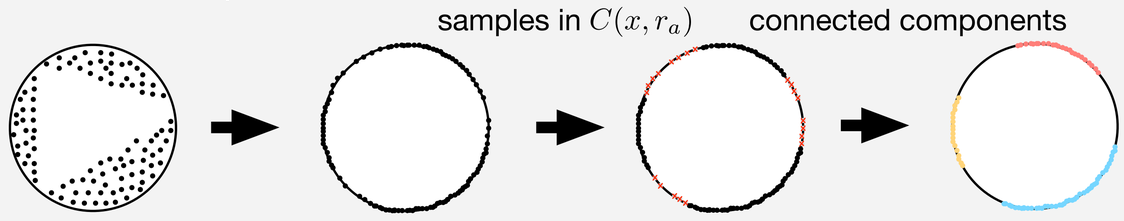
One-vs-all on the Unit Ball

1. Output code is one-vs-all (C is the identity).
2. $K_i = \{x : h_i(x) > 0\}$ is the set of points belonging to class i.
3. The data is nearly uniform on the set $K = \bigcup_{i=1}^{L} K_i$. That is, the density p is supported on K and satisfies

$$0 < c_{lb} \leq p(x) \leq c_{ub} \qquad \forall x \in K.$$



1. Project onto the unit sphere **2.** Discard samples with fewer than τn samples in $C(x, r_a)$ **3.** Connect points closer than r_c and find connected components

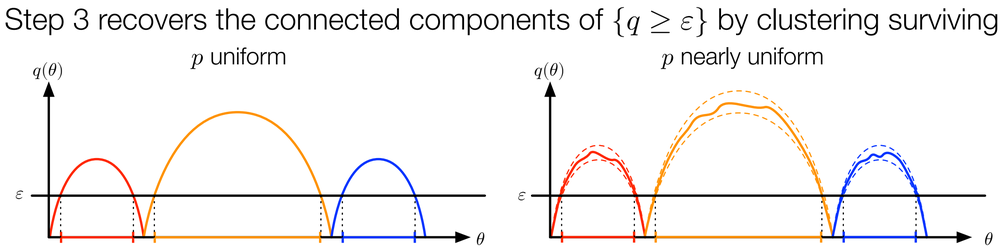


4. Classify x by projecting onto the sphere and outputting label of nearest cluster.

Theorem. *For any $\varepsilon, \delta > 0$, running the above algorithm with $r_c = \Omega(\varepsilon c_{\mathrm{lb}} / (c_{\mathrm{ub}}^2 b_{\min}))$, $r_a = r_c/2$, $\tau = \frac{c_{\mathrm{lb}}}{2c_{\mathrm{ub}}} V^d(r_a)\varepsilon$, and $n = \tilde{O}((c_{\mathrm{ub}}^4 d / (\varepsilon^2 c_{\mathrm{lb}}^2 b_{\min}^2))^d)$ will have error $\leq \varepsilon$ w.p. $\geq 1 - \delta$.*

Main Ideas:

- After projecting onto the sphere, the projected density q is no longer nearly uniform.
- We learn by estimating the connected components of $\{q \geq \varepsilon\}$.
- Step 2 discards low density points and keeps high-density points.
- Step 3 recovers the connected components of $\{q \geq \varepsilon\}$ by clustering surviving points.



Extension to the Agnostic Setting

- Data generated by distribution P over $\mathcal{X} \times \mathcal{Y}$, $\exists f^*$ with $\Pr_{(x,y) \sim P}(f^*(x) \neq y) \leq \eta$.
- Our algorithms have two steps: first, partition the unlabeled data into groups, and second, query the labels of the large groups.
- In the agnostic setting, we simply query multiple labels per group to recover the label assignment given by f^*, which gives error $\leq \eta + \varepsilon$.