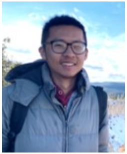
Bandit Multiclass Linear Classification: Efficient Algorithms for the Separable Case

Alina Beygelzimer (Yahoo!)

David Pal (Yahoo!)

Balazs Szorenyi (Yahoo!)

Devanathan Thiruvenkatachari (NYU)



Chen-Yu Wei (USC)

Chicheng Zhang (Microsoft)

Bandit multiclass classification

For $t = 1, 2, \ldots, T$:

Bandit multiclass classification

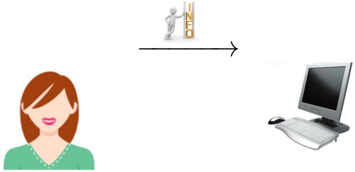
For $t = 1, 2, \ldots, T$:

1. Example (x_t, y_t) is chosen, where

Bandit multiclass classification

For $t = 1, 2, \ldots, T$:

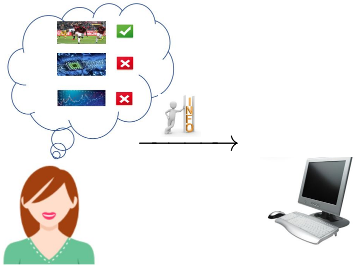
1. Example (x_t, y_t) is chosen, where $x_t \in \mathbb{R}^d$ is the feature (shown to the learner),



Bandit multiclass classification

For $t = 1, 2, \ldots, T$:

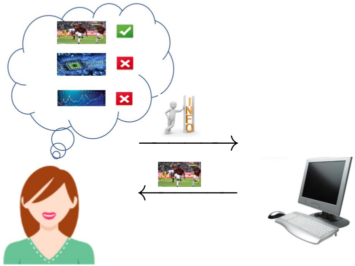
1. Example (x_t, y_t) is chosen, where
   $x_t \in \mathbb{R}^d$ is the feature (shown to the learner),
   $y_t \in [K]$ is the label (hidden).



Bandit multiclass classification

For $t = 1, 2, \ldots, T$:

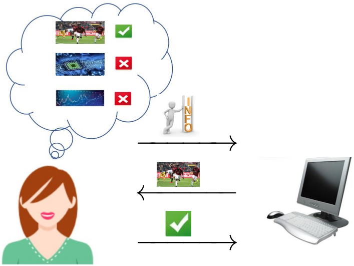
1. Example (x_t, y_t) is chosen, where
   $x_t \in \mathbb{R}^d$ is the feature (shown to the learner),
   $y_t \in [K]$ is the label (hidden).
2. Predict class label $\widehat{y}_t \in [K]$.



Bandit multiclass classification

For $t = 1, 2, \ldots, T$:

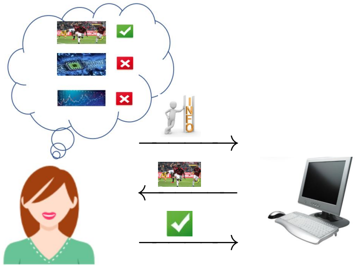
1. Example (x_t, y_t) is chosen, where
   $x_t \in \mathbb{R}^d$ is the feature (shown to the learner),
   $y_t \in [K]$ is the label (hidden).
2. Predict class label $\widehat{y}_t \in [K]$.
3. Observe feedback
   $z_t = \mathbb{1}\left[\widehat{y}_t \neq y_t\right] \in \{0, 1\}$.



Bandit multiclass classification

For $t = 1, 2, \ldots, T$:

1. Example (x_t, y_t) is chosen, where
   $x_t \in \mathbb{R}^d$ is the feature (shown to the learner),
   $y_t \in [K]$ is the label (hidden).
2. Predict class label $\widehat{y}_t \in [K]$.
3. Observe feedback
   $z_t = \mathbb{1}[\widehat{y}_t \neq y_t] \in \{0, 1\}$.



Goal: minimize the total number of mistakes $\sum_{t=1}^{T} z_t$.

Challenge: efficient algorithms in the separable setting

Definition

A dataset is called γ-linearly separable if there exists $w_1, \ldots, w_K$ such that

$$\langle w_y, x \rangle \geq \langle w_{y'}, x \rangle + \gamma, \qquad \forall y' \neq y,$$

for all (x, y) in the dataset. (with the constraint $\sum_{i=1}^{K} \|w_i\|^2 \leq 1$)

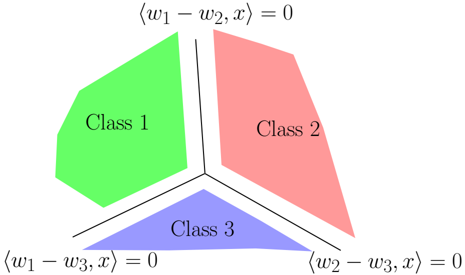
Challenge: efficient algorithms in the separable setting

Definition

A dataset is called γ-linearly separable if there exists $w_1, \ldots, w_K$ such that

$$\langle w_y, x \rangle \geq \langle w_{y'}, x \rangle + \gamma, \qquad \forall y' \neq y,$$

for all (x, y) in the dataset. (with the constraint $\sum_{i=1}^{K} \|w_i\|^2 \leq 1$)



Related work

Algorithm	Mistake Bound	Efficient?

[1] See also [HK11, BOZ17, FKL$^+$18, ..] that have similar guarantees

Related work

Algorithm	Mistake Bound	Efficient?
Minimax algorithm [DH13]	$O(K/\gamma^2)$	No

[1]See also [HK11, BOZ17, FKL$^+$18, ..] that have similar guarantees

Related work

Algorithm	Mistake Bound	Efficient?
Minimax algorithm [DH13]	$O(K/\gamma^2)$	No
Banditron [KSST08] [1]	$O(\sqrt{TK/\gamma^2})$	Yes

[1] See also [HK11, BOZ17, FKL$^+$18, ..] that have similar guarantees

Related work

Algorithm	Mistake Bound	Efficient?
Minimax algorithm [DH13]	$O(K/\gamma^2)$	No
Banditron [KSST08] [1]	$O(\sqrt{TK/\gamma^2})$	Yes
This work	$2^{\widetilde{O}(\min(K \log^2(1/\gamma), \sqrt{1/\gamma} \log K))}$	Yes

[1] See also [HK11, BOZ17, FKL$^+$18, ..] that have similar guarantees

Related work

Algorithm	Mistake Bound	Efficient?
Minimax algorithm [DH13]	$O(K/\gamma^2)$	No
Banditron [KSST08] [1]	$O(\sqrt{TK/\gamma^2})$	Yes
This work	$2^{\widetilde{O}(\min(K \log^2(1/\gamma), \sqrt{1/\gamma} \log K))}$	Yes

Contribution: first efficient algorithm that breaks the $\sqrt{T}$ barrier

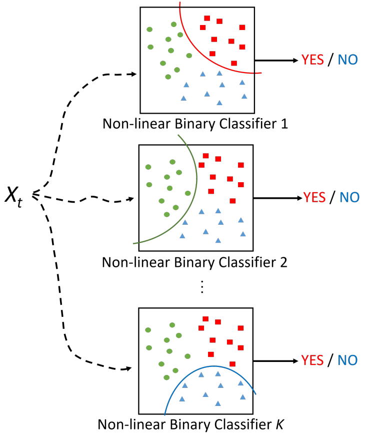
[1] See also [HK11, BOZ17, FKL$^+$18, ..] that have similar guarantees

Algorithm

(One-versus-rest approach)

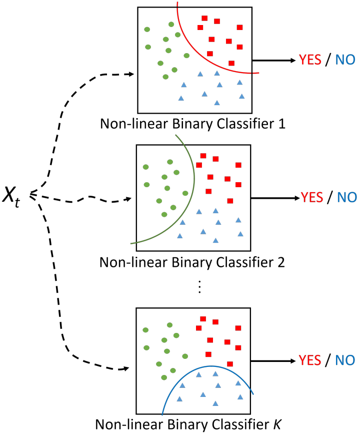
Algorithm

(One-versus-rest approach)



Algorithm

(One-versus-rest approach)

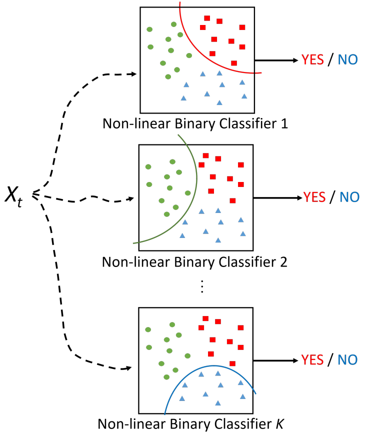


If ≥ 1 of them respond YES:
$\widehat{y}_t \leftarrow$ any one of those YES labels

If all of them respond NO:
$\widehat{y}_t \leftarrow$ uniform from $\{1, \ldots, K\}$

Algorithm

(One-versus-rest approach)



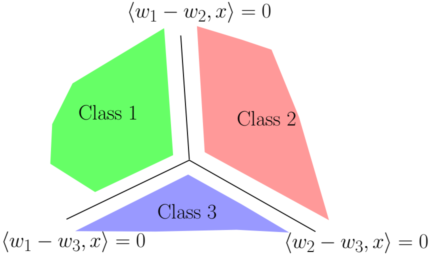
If ≥ 1 of them respond YES:
$\widehat{y}_t \leftarrow$ any one of those YES labels

If all of them respond NO:
$\widehat{y}_t \leftarrow$ uniform from $\{1, \ldots, K\}$

$$\boxed{\mathbb{E}[\#\text{mistakes(alg)}] \leq K \sum_i \#\text{mistakes}(i)}$$

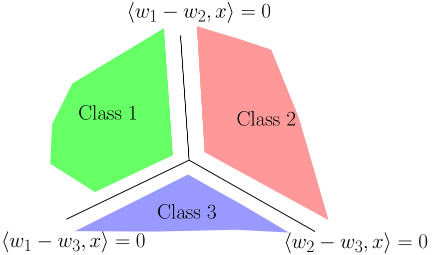
Algorithm

- Each non-linear binary classifier learns the support of class i, which lies in an intersection of $K - 1$ halfspaces with a margin [KS04].



Algorithm

- Each non-linear binary classifier learns the support of class i, which lies in an intersection of $K-1$ halfspaces with a margin [KS04].

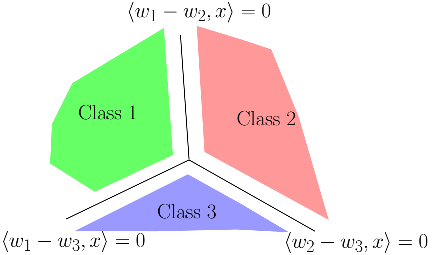


- Choice: **kernel Perceptron** with **rational kernel** [SSSS11]:

$$K(x, x') = \frac{1}{1 - \frac{1}{2}\langle x, x' \rangle}.$$

Algorithm

- Each non-linear binary classifier learns the support of class i, which lies in an intersection of $K-1$ halfspaces with a margin [KS04].



- Choice: **kernel Perceptron** with **rational kernel** [SSSS11]:

$$K(x, x') = \frac{1}{1 - \frac{1}{2}\langle x, x' \rangle}.$$

- Thu. Poster#158