Bandit Multiclass Linear Classification: Efficient Algorithms for the Separable Case

Alina Beygelzimer  Dávid Pál  Balázs Szörényi (Yahoo! Research, New York City)
Devanathan Thiruvenkatachari (New York University)
Chen-Yu Wei (University of Southern California)
Chicheng Zhang (Microsoft Research, New York City)

Abstract
We design efficient algorithms for online bandit $K$-class linear classification when the data is linearly separable by a margin $\gamma$. We consider two notions of linear separability, strong and weak.

1. Under the strong linear separability condition, we design an efficient algorithm that achieves a near-optimal mistake bound $O(1/\gamma^2)$.
2. Under the more challenging weak linear separability condition, we design an efficient algorithm with a mistake bound quasi-polynomial in $K$ for constant $K$. Our key observation is a reduction from the weak linear separability to strong linear separability condition via a specialized nonlinear mapping.

Online Bandit Linear 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 $\hat{y}_t \in [K]$.
3. Observe feedback $y_t = 1$ if $\hat{y}_t \neq y_t \in \{0, 1\}$.

Goal: minimize the total number of mistakes $\sum_{t=1}^T y_t - \hat{y}_t$.

Technical assumption: $|x_t| \leq 1$ for all $i$.

Notions of Linear Separability
Multiclass linear classification: classifier $W = (w_1, w_2, \ldots, w_K) \in \mathbb{R}^{K \times d}$ predicts on $i$ by:
1. Compute $i$-th score $(w_i, x)$ for each label $i$.
2. Predict $\hat{y} = \arg\max_i \langle w_i, x \rangle$.

Weakly linearly separable: there exists $W^*$ with $\|W^*\|_F \leq 1$, and for all $(i, y)$:
$$\langle w_i^*, x \rangle - \gamma \geq \gamma, \quad \forall y \neq y.$$

Strongly linearly separable: there exists $W^*$ with $\|W^*\|_F \leq 1$, and for all $(i, y)$:
$$\langle w_i^*, x \rangle \geq \gamma/2, \quad \forall y \neq y.$$

Algorithm
Key idea:
1. Create $K$ online classification tasks $T_i, i = 1, \ldots, K$, where task $T_i$ is to predict whether examples belong to class $i$.
2. For each task $T_i$, maintain a separate online classification algorithm $A_i$.
3. When predicting, aggregate the predictions from all $A_i$'s after receiving the feedback.

for $t = 1, 2, \ldots, T$: do

Receive example $x_t$.

Query: For $i = 1, \ldots, K$, ask algorithm $A_i$ whether $x_t$ belongs to class $i$.

Predict:
Case 1: if $\geq 1$ of them respond YES: $\hat{y}_t = \text{any one of those YES labels}$
Case 2: if all of them respond NO: $\hat{y}_t = \text{uniform from } \{1, \ldots, K\}$

Receive feedback $y_t = 1$ if $\hat{y}_t \neq y_t$.

Update:
Case 1: if $y_t = 1$, send example $(x_t, \text{NO})$ to $A_y$.
Case 2: if $y_t = 0$, send example $(x_t, \text{YES})$ to $A_y$.

Theorem 1. If for each $i$, $A_i$ makes at most $M_i$ mistakes for task $T_i$, then our proposed algorithm makes at most $K(M_1 + \ldots + M_K)$ in expectation.

Performance Guarantees

<table>
<thead>
<tr>
<th>Setting</th>
<th>Sublearner $A_i$</th>
<th>Sublearner mistake bound</th>
<th>Mistake bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly linearly separable</td>
<td>Perceptron</td>
<td>$O(1/\gamma^2)$</td>
<td>$O(K/\gamma^2)$ (tight)</td>
</tr>
<tr>
<td>Weakly linearly separable</td>
<td>kernel Perceptron with rational kernel $k(x, x') = \ell(x, x')^{-\gamma}$</td>
<td>$\ell$ $\max(K \log(1) \sqrt{\ell(k)}$</td>
<td>$\ell$ $\max(K \log(1) \sqrt{\ell(k)}$</td>
</tr>
</tbody>
</table>

Empirical Evaluation

Experiment 1: strongly separable setting. Our algorithm with linear Perceptron and rational kernel Perceptron performs well and exhibit finite mistake bound experimentally.

Experiment 2: weakly separable setting. Our algorithm with rational kernel Perceptron performs well and exhibit finite mistake bound experimentally. Our algorithm with linear Perceptron has a high number of mistakes, which is within expectation.

Hardness Results
1. Any "ignorant algorithm" will make $\Omega(\min(\sqrt{T}, d^{1/\gamma}))$ mistakes even when the data is strongly linearly separable. An ignorant algorithm does not update itself when it makes a mistake (variants of SOBA [Beygelzimer et al., 2017] and OBAMA [Foster et al., 2018] are of this type).
2. Finding a linear classifier that agrees with a labeled dataset and a complementary labeled dataset is NP-hard (naive algorithm requires $2^{|K|}$ computational complexity). A complementary labeled dataset [Ishida et al., 2017] consists of the following types of examples:

$$(\langle x, \text{NO} \rangle, \text{example } x \text{ does not belong to class } y).$$

Related Work - Weakly Linearly Separable Setting

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mistake bound (in big-O)</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halving [Kakade et al., 2008]</td>
<td>$K^{\gamma} \ln(T)/\gamma$ or $dK^{\gamma} \ln(1/\gamma)$</td>
<td>No</td>
</tr>
<tr>
<td>Minimax algorithm [Littman and Helbertal, 2013]</td>
<td>$\min(K/\gamma^2, dK \ln(1/\gamma))$ (tight)</td>
<td>No</td>
</tr>
<tr>
<td>Banditron [Kakade et al., 2008]</td>
<td>$\min(K/\gamma^2, dK \ln(1/\gamma))$ (tight)</td>
<td>Yes</td>
</tr>
<tr>
<td>SOBA [Beygelzimer et al., 2017]</td>
<td>at least $\sqrt{\gamma T} \ln T$ or $\sqrt{\gamma KT}$</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Observation: finding an intersection of two halfspaces that agrees with a dataset is NP-hard [Blum and Rivest, 1993].