CSCI 5525 Machine Learning Fall 2019

Lecture 23: Online Learning (Part 1)

Nov 19th 2019

Lecturer: Steven Wu *Scribe: Steven Wu*

We give two online learning algorithms with mistake bounds.

1 Halving Algorithm

Consider the following online prediction problem with N experts:
For rounds $t = 1, \ldots, T$:

- Experts predict $f_{1,t}, \ldots, f_{N,t} \in \{0, 1\}$
- Learner makes prediction $\hat{y}_t$ (based on the experts' predictions)
- Observe the true label y_t and incurs loss $\mathbf{1}[y_t \neq \hat{y}_t]$

Here is the simple *halving algorithm* for solving the online prediction problem:

- Initialization before round 1: $S_1 = \{1, \ldots, N\}$.
- At each round t:
  - Predict $\hat{y}_t$ as the majority vote of S_t
  - Update $S_{t+1} = \{i \in S_t \mid f_{i,t} = y_t\}$, that is the set of experts who still have perfect predictios so far.p

Under the assumption that there is a perfect expert i^* such that $f_{i^*,t} = y_t$, we can show that the halving algorithm makes bounded number of mistakes, regardless of how large T is.

Theorem 1.1. *The number of mistakes of the halving algorithm is bounded by* $\log_2 N$.

2 Perceptron algorithm

Now let's consider an online linear prediction problem. The *perceptron algorithm* initialize $\mathbf{w}_1$ as the all-zero vector in $\mathbb{R}^d$, and proceeds over rounds $t = 1, \ldots, T$:

- observes feature vector $x_t \in \mathbb{R}^d$,
- makes prediction $\hat{y}_t = \mathbf{1}[\boldsymbol{w}_t^\intercal x_t > 0]$,
- observes label y_t

- update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + \mathbf{1}[y_i \mathbf{w}^\intercal x_i \leq 0] y_i x_i$

Thus, in each round t, the perceptron algorithm either makes the correct prediction or moves update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + y_i x_i$.

Under the assumption that there is a perfect linear classifier with a margin, we can show that the perceptron algorithm will make bounded number of mistakes, regardless of how large T is.

Theorem 2.1. *Assume that there exists some $\mathbf{w}^* \in \mathbb{R}^d$ such that for all t,*

$$y_t x_t^\intercal \mathbf{w}^* \geq \gamma,$$

and that $\|x_t\|_2 \leq L$. Then the total number of mistakes made by the algorithm is bounded by

$$\frac{\|\mathbf{w}^*\|_2^2 L^2}{\gamma^2}.$$

Proof idea. Let B be the number of mistakes. To bound B, one can show that $\mathbf{w}_T^\intercal \mathbf{w}^* \geq B\gamma$ and also $\|\mathbf{w}_T\| \leq L\sqrt{B}$. By Cauchy-Schwarz inequality, $\mathbf{w}_T^\intercal \mathbf{w}^* \leq \|\mathbf{w}_T\|_2 \|\mathbf{w}^*\|_2$, and so

$$B\gamma \leq \mathbf{w}_T^\intercal \mathbf{w}^* \leq L\sqrt{B}\|\mathbf{w}^*\|_2$$

which leads to the stated bound. □