Machine Learning **Fall 2023**

Lecture 6: Learning Algorithms

Lecturer: Liwei Wang Scribe: Yizao Tang,Tianyi Wu,Zhaoji Zhang,Zirui Zhou,Suhan Ling,Yanzhe Wang

6.1 Recap : VC Theory

VC Theory is data independent & algorithm independent, and only model dependent. $P_D(Y \neq f(X)) \leq P_S(Y \neq f(X)) + O(\sqrt{\frac{d \ln n + ln \frac{1}{\delta}}{n}})$ for $d = \text{VCD}(F)$. When $d > n$, we called the model **overparametrization**, then we can't use VC Theory to analysis generalization. In deep learning, d is often greater than the number of parameters, so overparametrization often occurs in large model. It shows that the algorithm is important as well.

6.2 Linear Classification

Given training data set $(x_i, y_i), \cdots, (x_n, y_n)$, $x_i \in \mathbf{R}^d$, $y_i \in \{\pm 1\}$. Design an efficient algorithm to tell if the training data are linear separable. $\mathcal{F} = \{\text{sgn}(w^T x + b) : w \in \mathbf{R}^d, b \in \mathbf{R}\}$. We can use linear programming to solve it.

When the solution is not unique, we try to find the "best" seperating line. We define the **margin** be the minimal distance between every point to the separating line, and we also want to maximize the margin, called **max margin classifier**. Formally, we want to maximize $\min_i \{y_i(w^T x_i + b)\}$, $\|w\|_2 = 1$. We write it down in the form of programming:

$$\begin{aligned} \max_{w,b,t} \quad & t \\ \text{s.t. } & y_i(w^T x_i + b) \geq t, \forall i \in [n] \\ & \|w\|_2 = 1 \end{aligned}$$

The t here is called a **slack**. $t > 0$ indicates that the data set is separable.

Now we need to introduce the Convex Optimization Problem, which also have known efficient algorithm:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t. } & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

where f is convex function, and g, h is linear function.

If we only concern the $\frac{w}{\|w\|_2}$ and b, we could rewrite it in the form of quadratic programming and transform our problem into Convex Optimization Problem, which is:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \ \forall\, i \in [n]$$

Homework 1: Prove they have the same solution.

6.3 Minmax Theory and Duality

6.3.1 Maxtrix Game, two player, Zero Sum Game, one-shot

Zero Sum Game : the sum of the profit of Alice and Bob is zero.

one-shot : only play once.

for this kind of game, we could write a payoff matrix M

$$\begin{array}{cc} & \begin{array}{ccccc} \text{Bob} \quad 1 & \dots & j & \dots & n \end{array} \\ \text{Alice} & \\ \begin{array}{c} 1 \\ \dots \\ i \\ \dots \\ m \end{array} & \begin{pmatrix} (a_1, b_1) & \dots & (a_1, b_j) & \dots & (a_1, b_n) \\ \dots & \dots & \dots & \dots & \dots \\ (a_i, b_1) & \dots & (a_i, b_j) & \dots & (a_i, b_n) \\ \dots & \dots & \dots & \dots & \dots \\ (a_m, b_1) & \dots & (a_m, b_j) & \dots & (a_m, b_n) \end{pmatrix} \end{array}$$

The Cell (a_i, b_j) describes the reward of Alice and Bob when they choose the action i and j. If we substitute the (a_i, b_j) as $m_{i,j}$, we have

$$\begin{array}{cc} & \begin{array}{ccccc} \text{Bob} \quad 1 & \dots & j & \dots & n \end{array} \\ \text{Alice} & \\ \begin{array}{c} 1 \\ \dots \\ i \\ \dots \\ m \end{array} & \begin{pmatrix} m_{1,1} & \dots & m_{1,j} & \dots & m_{1,n} \\ \dots & \dots & \dots & \dots & \dots \\ m_{i,1} & \dots & m_{i,j} & \dots & m_{i,n} \\ \dots & \dots & \dots & \dots & \dots \\ m_{m,1} & \dots & m_{m,j} & \dots & m_{m,n} \end{pmatrix} \end{array}$$

Let M be the payoff matrix. Then we describe the (Pure Strategy) matrix game:

1. Alice chooses a row i in M
2. Bob chooses a column j after observing Alice's action
3. Alice pays $m_{i,j}$ to Bob

As everyone tries to maximize his profit, Alice actually pays

$$A = \min_i\{\max_j\{m_{i,j}\}\}$$

When Bob moves first, Alice choose a row after Bob's action, Alice actually pays

$$B = \max_j\{\min_i\{m_{i,j}\}\}$$

Then we always have $A \geq B$, the player moves last has the advantage.

Homework 2: Prove that A and B have a certain Ordered Relation that $A \geq B$.