Machine Learning **Spring 2021**

Lecture 4: VC Theory, Practical Learning Algorithms, Game Theory

Lecturer: Liwei Wang *Scribe: Haoyu Li, Ting Lei, Zhenyu Du, Yang Hong, Zhen Wu*

4.1 VC Theory (Cont'd)

4.1.1 VC Theorem

Theorem 4.1 (VC Theorem) *For function space $\mathcal{F}$ with VC-dimension d, with probability at least $1-\delta$ over the random draw of training data,*

$$\mathbb{P}_{\mathcal{D}}(y \neq f(x)) \leq \mathbb{P}_{\mathcal{S}}(y \neq f(x)) + O\left(\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{1}{\delta}}{n}}\right)$$

holds uniformly for all $f \in \mathcal{F}$.

Proof: By Corollary 3.7 in Lecture 3,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbb{P}_{\mathcal{S}}(y \neq f(x)) - \mathbb{P}_{\mathcal{D}}(y \neq f(x))| \geq \epsilon\right) \leq \left(\frac{2en}{d}\right)^d e^{-cn\epsilon^2}$$

Let $\delta = \left(\frac{2en}{d}\right)^d e^{-cn\epsilon^2}$, then

$$\epsilon = \sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{1}{\delta} + 2d}{cn}} = O\left(\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{1}{\delta}}{n}}\right)$$

Therefore, with probability at least $1-\delta$ over the random draw of training data,

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}(y \neq f(x)) &\leq \mathbb{P}_{\mathcal{S}}(y \neq f(x)) + \epsilon \\ &= \mathbb{P}_{\mathcal{S}}(y \neq f(x)) + O\left(\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{1}{\delta}}{n}}\right)\end{aligned}$$

holds for all $f \in \mathcal{F}$. ■

Note that this inequality holds uniformly for all $f \in \mathcal{F}$, which implies that the bound of the deviation of the two probabilities given by this theorem is a 'worst case guarantee', or an algorithm independent bound.

Further, consider ERM(Empirical Risk Minimization) function on training data

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \mathbb{P}_{\mathcal{S}}(y \neq f(x))$$

and the optimal classifier on the distribution

$$f^* := \arg\min_{f \in \mathcal{F}} \mathbb{P}_{\mathcal{D}}(y \neq f(x))$$

Applying the Theorem above, we immediately have a bound of the difference between the error of the two classifiers.

Theorem 4.2

$$\mathbb{P}_{\mathcal{D}}(y \neq \hat{f}(x)) \leq \mathbb{P}_{\mathcal{D}}(y \neq f^*(x)) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right)$$

Proof:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}(y \neq \hat{f}(x)) &\leq \mathbb{P}_{\mathcal{S}}(y \neq \hat{f}(x)) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right) \\
&\leq \mathbb{P}_{\mathcal{S}}(y \neq f^*(x)) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right) \\
&\leq \left(\mathbb{P}_{\mathcal{D}}(y \neq f^*(x)) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right)\right) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right) \\
&= \mathbb{P}_{\mathcal{D}}(y \neq f^*(x)) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right)
\end{aligned}$$

■

We also wonder if $\mathbb{P}_{\mathcal{S}}(y \neq f(x)) + O\left(\sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}\right)$ is a tight upper bound of $\mathbb{P}_{\mathcal{D}}(y \neq f(x))$. The answer to this question is positive, as the equality cases have been found.

4.1.2 VC dimension

To calculate the VC dimensions, we have to find a number d, so that there exists a set of d points which can be shattered by the classifiers, and any set of $d+1$ set can't be shattered. As a example, we introduce the VC dimension of linear classifiers below.

Theorem 4.3 *For* $\mathcal{F} = \left\{\operatorname{sgn}\left(w^T x + b\right) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\right\}, d_{\mathcal{F}} = d+1$. *For* $\mathcal{F}^* = \left\{\operatorname{sgn}\left(w^T x\right) \mid w \in \mathbb{R}^d\right\}, d_{\mathcal{F}^*} = d$.

Proof: Let $x_1 = (1, 0, 0, \ldots, 0), x_2 = (0, 1, 0, \ldots, 0), \ldots, x_d = (0, 0, \ldots, 0, 1), x_{d+1} = (0, 0, \ldots, 0)$. Then we have:

$$N^{\mathcal{F}}(x_1, \ldots, x_{d+1}) = \left|\left\{(\operatorname{sgn}(w_1 + b), \ldots, \operatorname{sgn}(w_d + b), \operatorname{sgn}(b)) \mid w \in R^d, b \in R\right\}\right| = 2^{d+1}$$

Thus $d_{\mathcal{F}} \geq d+1$. Next we show $d_{\mathcal{F}} < d+2$, which indicates that $d_{\mathcal{F}} = d+1$.

$\forall x_1, \ldots, x_{d+2} \in \mathbb{R}^d, (x_1, 1), \ldots, (x_{d+2}, 1)$ are linear dependent. So there exists $c_1, \ldots, c_{d+2}, s.t$

$$c_1 (x_1, 1) + \ldots + c_{d+2} (x_{d+2}, 1) = 0$$

Let $c_{d+2} = 1$, then $\forall w \in \mathbb{R}^d, b \in \mathbb{R}$,

$$w^T x_{d+2} + b = \sum_{i=1}^{d+1} (-c_i) \left(w^T x_i + b\right)$$

Assume that $(\operatorname{sgn}(c_1), \ldots, \operatorname{sgn}(c_{d+1}), 1) \in \{(f(x_1), \ldots, f(x_{d+2}) \mid f \in \mathcal{F}\}$. Then $\exists w \in \mathbb{R}^d, b \in \mathbb{R}$, s.t. $w^T x_i + b$ and c_i have the same sign$(1 \leq i \leq d+1)$, and $\operatorname{sgn}\left(w^T x_{d+2} + b\right) = 1$, which is contradictory to $w^T x_{d+2} + b = \sum_{i=1}^{d+1} (-c_i) \left(w^T x_i + b\right) < 0$.

Thus $(\operatorname{sgn}(c_1), \ldots, \operatorname{sgn}(c_{d+1}), 1) \notin \{(f(x_1), \ldots, f(x_{d+2}) \mid f \in \mathcal{F}\}$, and $N^{\mathcal{F}}(x_1, \ldots, x_{d+2}) < 2^{d+2}$. So $d_{\mathcal{F}} = d+1$. Proof for $\mathcal{F}^*$ is similar. ■

Additionally, for binary classifiers, the VC dimension of Φ is equal to the VC dimension of $\mathcal{F}$.

Theorem 4.4 *$f \in \mathcal{F}$ are binary classifiers, let $\Phi = \{\phi_f(z) = I[y \neq f(x)] \mid f \in \mathcal{F}\}$, then $d_\Phi = d_{\mathcal{F}}$.*

Proof: Let $d_\Phi = d$, by definition we have $N^\Phi(d+1) < 2^{d+1}, N^\Phi(d) = 2^d$. We notice that $f(x) = \phi_f(x, 0)$, so we have:

$$N^{\mathcal{F}}(x_1, \ldots x_{d+1}) = N^\Phi((x_1, 0), \ldots, (x_{d+1}, 0)) \leq N^\Phi(d+1) < 2^{d+1}$$

Then $d_{\mathcal{F}} < d+1$.

Since $d_\Phi = d$, $\exists (x_1, y_1), \ldots, (x_d, y_d)$, s.t. $\forall w \in \{0,1\}^d, \exists f \in \mathcal{F}, (\phi_f(x_1, y_1), \ldots, \phi_f(x_d, y_d)) = w$, which implies that:

$$(f(x_1) + y_1, \ldots, f(x_d) + y_d) \equiv -w (\text{mod} 2) \Rightarrow (f(x_1), \ldots, f(x_d)) \equiv -w + (y_1, \ldots, y_d) (\text{mod} 2)$$

$\forall w \in \{0,1\}^d$, let $w' = -w + (y_1, \ldots, y_d), \exists f \in \mathcal{F}, s.t. (f(x_1), \ldots, f(x_d)) \equiv -w' + (y_1, \ldots, y_d) = w (\text{mod} 2)$. Then $(f(x_1), \ldots, f(x_d)) = w$. Thus $N^{\mathcal{F}}(d) = 2^d$, $d_{\mathcal{F}} = d$. ■

4.2 Practical Learning Algorithms

4.2.1 Linear Classfier

A linear classifier on $\mathbb{R}^d$ is defined as $\mathcal{F} = \{\operatorname{sgn}(w^\top x + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$. Consider how to decide if a dataset is linear separable. That is, for $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{\pm 1\}$, given $(x_i, y_i)_{i=1}^n$, decide if $\exists f \in \mathcal{F}$, $\mathbf{s.t.} \sum_i I[f(x_i) \neq y_i] = 0$. In this specific problem, that's to say to decide if $\exists w \in \mathbb{R}^d, b \in \mathbb{R}, \mathbf{s.t.} y_i(w^\top x_i + b) >= 0 (\forall i \in [n])$. To solve this question efficiently, we can use the following linear programming(LP).

$$\begin{aligned} \max_{w,b,t} \quad & t \\ \mathbf{s.t.} \quad & y_i(w^\top x_i + b) \geq t \quad i \in [n] \end{aligned}$$

When $t > 0$, it's separable, since t is the minimum of all $y_i(w^\top x_i + b)$. For a separable training set, there might be many classifiers, to find the best one, we try to maximize the minimal distance.

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

However this is neither a LP nor a QP(quadratic programming), we can't solve it directly with an efficient algorithm. Fortunately, the following QP is equivalent to that.

$$\begin{aligned} \min_{w,b,t} \quad & \tfrac{1}{2}\|w\|^2 \\ \textbf{s.t.} \quad & y_i(w^\top x_i + b) \geq 1 \quad i \in [n] \end{aligned}$$

The equivalence mentioned above could be obvious, i.e, if we divide the former inequation by t, and adjust the objective to $min \quad 1/t$,through some trivial substitution, the former one could have an equivalence relation with the latter one.

4.3 Appendix: Game Theory

In this section, we'll mainly talk about fundamental game theory. Game theory is the study of mathematical models of strategic interaction among rational decision-makers.Two-player Matrix Game is a simple model of game theory thus we address it as an introduction. We only consider zero-sum games, in which each participant's gains or losses are exactly balanced.

In the first step, we consider two players —— Alice and Bob —— choosing pure strategies when playing a zero-sum game. We'll find that the later mover will take advantage, which correspond to our intuition.

However, in the second step, our intuition is not that reliable. The game based on mixed strategy seems to come to an equilibrium point, making the later mover no longer take advantage at all.

Modern game theory began with the idea of mixed-strategy equilibria in two-person zero-sum games and its proof by John von Neumann. Von Neumann's original proof used the Brouwer fixed-point theorem on continuous mappings into compact convex sets, which became a standard method in game theory and mathematical economics.

In the 1950s, John Nash developed a criterion for mutual consistency of players' strategies known as the Nash equilibrium, applicable to a wider variety of games than the criterion proposed by von Neumann. Nash proved that every finite $n-$player, non-zero-sum (not just two-player zero-sum) non-cooperative game has a Nash equilibrium.

At the end of the notes, we'll talk about Sion's minimax theorem, which is a generalization of John von Neumann's minimax theorem.

So here we go. First of all, we define what a Two-player Matrix Game is.

Definition 4.5 (Two-player Matrix Game) *Let Alice and Bob are two players, $M = ((a_{ij}, b_{ij}))_{r\times c}$ is a matrix where every element is a pair of numbers $(a_{ij}, b_{ij}) \in \mathbb{R}^2$. Alice choose a row i while Bob choose a column j. Then the number a_{ij} and b_{ij} is the feedback of Alice and Bob respectively. Notice that M is known to Alice and Bob before making choices.*

As mentioned above, we only talk about Zero-sum Game. Here is its conception.

Definition 4.6 (Zero-sum Game) *The matrix M is defined as above, if $\forall i \in [1..r], j \in [1..c]$, we have $a_{ij} + b_{ij} = 0$, then we call the Two-player Matrix Game is a Zero-sum Game.*

That is to say, the benefit of one side of the opponent participating in the game process must mean the loss of the other side, so the sum of the gain and loss of both sides of the game must be zero. So M can be simplified to $M = (m_{ij})_{r\times c}$, where $m_{ij} \in \mathbb{R}$ indicates Alice should pay m_{ij} to Bob. Naturally, Alice want to minimize m_{ij} while Bob want to maximize it.

In order to maximize their own interests, both sides usually choose certain strategy.

Under each given information, if only one specific strategy can be selected, we call the strategy is a pure strategy. In the Two-player Matrix Game (if Alice go first), that is to say:

step 1: Alice choose a determined row i;

step 2: Bob observed Alice's strategy.According to information about M and Alice's strategy, Bob choose a determined column j;

step 3: Alice pays m_{ij} to Bob.

Then we say Alice and Bob use pure strategy.

If Alice go first, she always assume that if she choose row i, Bob will choose column j_i s.t. $m_{ij_i} = \max\limits_j m_{ij}$. So a natural strategy for Alice is to choose a row i_0 to minimize $\max\limits_j m_{ij}$. In this case, Alice's aim is $\min\limits_i \max\limits_j m_{ij}$.

In the same way, if Bob go first, Bob will choose a column j_0 to maximize $\min\limits_i m_{ij}$. So Alice's aim is $\max\limits_j \min\limits_i m_{ij}$.

We may ask: if the later mover will take advantage? i.e. $\forall$ given M, if we have $\min\limits_i \max\limits_j m_{ij} \geqslant \max\limits_j \min\limits_i m_{ij}$?

Intuitively, that's true. Because when Bob go first, as long as Alice always choose row i_0 (no matter what Bob choose), her loss cannot beyond $\min\limits_i \max\limits_j m_{ij}$. In mathematical language, that is to say $\min\limits_i \max\limits_j m_{ij} = \max\limits_j m_{i_0 j} \geqslant m_{i_0 j_0} \geqslant \min\limits_i m_{ij_0} = \max\limits_j \min\limits_i m_{ij}$. And we know the inequality sign can be obtained for some M, e.g $M := \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$.

So much for pure strategy. Here we come to mixed strategies where the players can choose a probability distribution on the pure strategies.

When Alice chooses the distribution vector $\vec{p} = (p_1, p_2, ..., p_r) \in [0,1]^r$ and Bob chooses $\vec{q} = (q_1, q_2, ..., q_c) \in [0,1]^c$ where $||\vec{p}|| = ||\vec{q}|| = 1$. The expectation of the value (In other words, Alice's loss or Bob's reward) is $\vec{p}^T M \vec{q}$. Obviously Alice's aim is to choose mixed strategy $\vec{p}_0$ and get $\min\limits_{\vec{p}} \max\limits_{\vec{q}} \vec{p}^T M \vec{q}$ if she go first; Bob will choose mixed strategy $\vec{q}_0$ and the expectation of the value will be $\max\limits_{\vec{q}} \min\limits_{\vec{p}} \vec{p}^T M \vec{q}$ if he go first. Similarly, we have $\min\limits_{\vec{p}} \max\limits_{\vec{q}} \vec{p}^T M \vec{q} = \max\limits_{\vec{q}} \vec{p}_0^T M \vec{q} \geqslant \vec{p}_0^T M \vec{q}_0 \geqslant \min\limits_{\vec{p}} \vec{p} M \vec{q}_0 = \max\limits_{\vec{q}} \min\limits_{\vec{p}} \vec{p}^T M \vec{q}$. It seems that the later mover will take advantage. However, the theorem given by John von Neuman conflict to our intuition, it tells us that there is no difference between go first and later.

Theorem 4.7 $\min\limits_{\vec{p}} \max\limits_{\vec{q}} \vec{p}^T M \vec{q} = \max\limits_{\vec{q}} \min\limits_{\vec{p}} \vec{p}^T M \vec{q}$.

The theorem can be proved via Brouwer's fixed point theorem or Farkas' lemma. In this course, however,

we'll present a proof with the knowledge of machine learning.

Theorem 4.8 (Sion's Minimax Theorem) *Suppose $f(x, y)$ is continuous, $\forall x \in \mathcal{X}, f(x, y)$ is concave, $\forall y \in \mathcal{Y}, f(x, y)$ is convex. Then*

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y)$$

Furthermore, if optimal value of the left-hand side is achieved at (x_1, y_1), and right-hand side is achieved at (x_2, y_2), then consider $(x^*, y^*) := (x_1, y_2)$, we have $f(x_1, y_1) \leq f(x^*, y^*) \leq f(x_2, y_2)$, which means $f(x_1, y_1) = f(x^*, y^*) = f(x_2, y_2)$, implying the optimal value of both sides can be achieved at the same point (x^*, y^*). If the condition changes to "strictly concave" and "strictly convex", then the optimal point is unique.

From the optimality, it's clear to see that $\forall x, f(x, y^*) \geq f(x^*, y^*)$, $\forall y, f(x^*, y) \leq f(x^*, y^*)$. So (x^*, y^*) is a saddle point of this function.

Imagine two players A and B play a game on this function, A wants to minimize the left-hand side and B wants to maximize the other side, (x^*, y^*) is the strategy of players, then this strategy is a equilibrium of the game, which means both A and B can't gain a better utility if one changes its strategy while the other stays still.

John Nash proved that for a finite game with at least 2 players, there exists a mixed Nash equilibrium. However, computing the equilibrium in general case isn't easy. It's proved to be a $\mathcal{PPAD}$-complete question.

References

[1] Vapnik, Vladimir (2000). The nature of statistical learning theory. Springer.

[2] Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M. K. (1989). "Learnability and the Vapnik–Chervonenkis dimension"

[3] https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_dimension