Machine Learning **Autumn 2025**

Lecture 4: VC Theory

Lecturer: Liwei Wang *Scribe: Group 3*

4.1 Term Project 1

4.1.1 Introduction

This project explores the application of machine learning (ML) techniques to improve bounds on chromatic numbers of Euclidean spaces, with a focus on the Hadwiger–Nelson problem and its higher-dimensional analogues.

4.1.2 Hadwiger–Nelson Problem (Chromatic Number of the Plane)

Let $G = (V, E)$ be an infinite graph defined as follows:

$$V = \{x : x \in \mathbb{R}^2\}, \quad E = \{(v, v') : v, v' \in V,\ d(v, v') = 1\}$$

where $d(v, v')$ denotes the Euclidean distance between v and v'.

The **chromatic number of the plane**, denoted $\chi(\mathbb{R}^2)$, is the minimum number of colors required to color all points in $\mathbb{R}^2$ such that no two points at distance 1 have the same color.

State of the Art (SOTA):

$$5 \leq \chi(\mathbb{R}^2) \leq 7$$

4.1.3 Chromatic Number of Space $\chi(\mathbb{R}^3)$

Similarly, define for $\mathbb{R}^3$:

$$V = \{x : x \in \mathbb{R}^3\}, \quad E = \{(v, v') : d(v, v') = 1\}$$

Then $\chi(\mathbb{R}^3)$ is the chromatic number of this graph.

State of the Art (SOTA)

$$6 \leq \chi(\mathbb{R}^3) \leq 15$$

4.1.4 Higher-Dimensional Generalization

The problem can be extended to $\mathbb{R}^d$ for $d \geq 4$, defining $\chi(\mathbb{R}^d)$ analogously.

4.1.5 Asymptotic Behavior

As $d \to \infty$, the chromatic number $\chi(\mathbb{R}^d)$ grows exponentially. Known bounds are of the form:

$$(1.239... + o(1))^d \leq \chi(\mathbb{R}^d) \leq (3 + o(1))^d$$

4.1.6 Machine Learning Objective

Use artificial intelligence (or other computational methods) to improve any of the above bounds — either the lower or upper bounds — for any dimension $d \geq 2$.

4.1.7 Possible ML Approaches

- Graph neural networks for finite subgraph analysis
- Reinforcement learning for coloring strategy discovery
- Supervised learning on known configurations and their chromatic numbers
- Dimensionality reduction and pattern recognition in high-dimensional colorings

4.2 Finite hypothesis space

Let's warm up with the simpler case where the hypothesis space $\mathcal{F}$ is finite, i.e., $|\mathcal{F}| < \infty$. We want to guarantee that the test error (inference error) of our learned classifier, $\mathbb{P}(Y \neq \hat{f}(X))$, is close to its empirical training error on the training set, $\frac{1}{n}\sum_{i=1}^{n} I[y_i \neq \hat{f}(x_i)]$.

By considering the worst case (which also distinguishes VC theory from other theories), we have

$$\begin{aligned} &\mathbb{P}\left(\mathbb{P}(Y \neq \hat{f}(X)) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq \hat{f}(x_i)\right] \geq \epsilon\right) \\ &\leq \mathbb{P}\left(\exists f \in \mathcal{F},\ \mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \geq \epsilon\right) \end{aligned} \tag{4.1}$$

Now, we can apply the union bound (also known as Boole's inequality) over all possible functions in the finite hypothesis space $\mathcal{F}$.

$$\begin{aligned} &\mathbb{P}\left(\exists f \in \mathcal{F},\ \mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \geq \epsilon\right) \\ &\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(\mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \geq \epsilon\right) \end{aligned} \tag{4.2}$$

For any single, fixed function f, since the training error is an empirical average of n i.i.d. Bernoulli random variables and the test error is the expectation, by Hoeffding's inequality, we have

$$\mathbb{P}\left(\mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \geq \epsilon\right) \leq e^{-2n\epsilon^2} \tag{4.3}$$

Combining these results gives us the final uniform convergence bound:

$$\mathbb{P}\left(\exists f \in \mathcal{F},\ \mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \geq \epsilon\right) \leq |\mathcal{F}|e^{-2n\epsilon^2} \tag{4.4}$$

This inequality shows that the generalization error is controlled by $|\mathcal{F}|$, which represents the model complexity. It highlights a fundamental concept: the complexity of the hypothesis space is crucial to generalization.

4.3 Infinite hypothesis space

When hypothesis space is infinite, still consider

$$\mathbb{P}\left(P\left(Y \neq \hat{f}(X)\right) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq \hat{f}(x_i)\right] \geq \epsilon\right). \tag{4.5}$$

In the worst case,

$$\begin{aligned} &\mathbb{P}\left(P\left(Y \neq \hat{f}(X)\right) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq \hat{f}(x_i)\right] \geq \epsilon\right) \\ \leq &\mathbb{P}\left(\sup_{f\in\mathcal{F}} P\left(Y \neq f(X)\right) - \frac{1}{n}\sum_{i=1}^{n} I\left[y_i \neq f(x_i)\right] \geq \epsilon\right). \end{aligned} \tag{4.6}$$

$\mathcal{F}$ is an infinite set. The infinity problem is always tricky. We want to "project" the infinite hypothesis space into finite training data (the amount of distinguishable classifiers f on n training data is limited). However, since (X, Y) and (x_i, y_i) and drawn from the underlying distribution, they are actually "infinite".

In Step 1, we use the double sample trick to replace the gap between the expectation and the empirical average with the gap between the empirical average of two samples so that we can "remove" the "infinity" of (X, Y).

In Step 2, we use symmetrization and rewrite the probability to "remove" the "infinity" of (x_i, y_i).

4.3.1 Step 1 : Double Sample Trick

Let $X_1, \ldots, X_{2n}$ be i.i.d. Bernoulli random variables. Define:

$$\nu_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \nu_2 = \frac{1}{n}\sum_{i=n+1}^{2n} X_i, \quad \text{and } \mathbb{E}[X] = p.$$

Assume $n \geq \frac{\ln 2}{\varepsilon^2}$. Then the following inequalities hold:

$$\frac{1}{2}\mathbb{P}(|\nu_1 - p| \geq 2\varepsilon) \leq \mathbb{P}(|\nu_1 - \nu_2| \geq \varepsilon) \leq 2\mathbb{P}\left(|\nu_1 - p| \geq \frac{\varepsilon}{2}\right). \tag{4.7}$$

(Interpretation: the gap between the expectation and the empirical average is "equivalent" the gap between the empirical average of two samples in some way.)

4.3.1.1 Proof

Part (i): Lower Bound Derivation First, by the Triangle inequality:

$$(|\nu_1 - p| \geq 2\varepsilon) \cap (|\nu_2 - p| \leq \varepsilon) \implies |\nu_1 - \nu_2| \geq \varepsilon.$$

By the *monotonicity of probability* (if $A \implies B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$) and the *independence* of ν_1 and ν_2, we have:

$$\mathbb{P}(|\nu_1 - p| \geq 2\varepsilon) \cdot \mathbb{P}(|\nu_2 - p| \leq \varepsilon) \leq \mathbb{P}(|\nu_1 - \nu_2| \geq \varepsilon).$$

Next, apply the *Chernoff Bound*:

$$\mathbb{P}(|\nu_2 - p| \geq \varepsilon) \leq e^{-2n\varepsilon^2}.$$

Given $n \geq \frac{\ln 2}{\varepsilon^2}$, substituting gives:

$$\mathbb{P}(|\nu_2 - p| \geq \varepsilon) \leq \frac{1}{2},$$

which implies:

$$\mathbb{P}(|\nu_2 - p| \leq \varepsilon) \geq \frac{1}{2}.$$

Substituting back, we obtain the lower bound:

$$\frac{1}{2}\mathbb{P}(|\nu_1 - p| \geq 2\varepsilon) \leq \mathbb{P}(|\nu_1 - \nu_2| \geq \varepsilon).$$

Part (ii): Upper Bound Derivation By Contraposition, we have:

$$|\nu_1 - \nu_2| \geq \varepsilon \implies \left(|\nu_1 - p| \geq \frac{\varepsilon}{2}\right) \vee \left(|\nu_2 - p| \geq \frac{\varepsilon}{2}\right).$$

By the *union bound* (Boole's inequality: $\mathbb{P}(A \vee B) \leq \mathbb{P}(A) + \mathbb{P}(B)$) and the fact that ν_1, ν_2 are identically distributed, we have:

$$\mathbb{P}(|\nu_1 - \nu_2| \geq \varepsilon) \leq 2\mathbb{P}\left(|\nu_1 - p| \geq \frac{\varepsilon}{2}\right).$$

4.3.1.2 Extending to Function Classes $\mathcal{F}$

Consider the inequality:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left[\mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i \neq f(x_i)]\right] \geq \varepsilon\right),$$

where $\mathbb{P}(Y \neq f(X))$ is the *expectation* (test error), and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i \neq f(x_i)]$ is the *empirical average* (training error, denoted ν_1).

By Equation 4.7, we have

$$\begin{aligned} &\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left[\mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i \neq f(x_i)]\right] \geq \varepsilon\right) \\ \leq &2\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[y_i \neq f(x_i)] - \frac{1}{n}\sum_{i=n+1}^{2n} \mathbb{I}[y_i \neq f(x_i)]\right] \geq \frac{\varepsilon}{2}\right). \end{aligned} \tag{4.8}$$

4.3.1.3 Why does sup still hold?

Right side (pointwise ⇒ supremum). If $A(f) \le B(f)$ for all $f \in \mathcal{F}$, then by monotonicity

$$\sup_{f\in\mathcal{F}} A(f) \le \sup_{f\in\mathcal{F}} B(f).$$

Equivalently,

$$\{\sup_f A(f) \ge \varepsilon\} \subseteq \{\sup_f B(f) \ge \varepsilon\} \;\Rightarrow\; \mathbb{P}(\sup_f A(f) \ge \varepsilon) \le \mathbb{P}(\sup_f B(f) \ge \varepsilon).$$

Left side (from fixed-f to uniform). Use *symmetrization*, not maximizing sequences. With data $(Z_i)_{i=1}^n$ and an i.i.d. ghost $(Z_i')_{i=1}^n$, for any fixed f,

$$\mathbb{P}\big((P - P_n)\ell_f \ge \varepsilon\big) \;\le\; 2\,\mathbb{P}\big((P_n' - P_n)\ell_f \ge \varepsilon/2\big).$$

Since this holds pointwise for every f, the same event-inclusion gives

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}}(P - P_n)\ell_f \ge \varepsilon\Big) \;\le\; 2\,\mathbb{P}\Big(\sup_{f\in\mathcal{F}}(P_n' - P_n)\ell_f \ge \varepsilon/2\Big).$$

Control the RHS by capacity: union bound if $|\mathcal{F}| < \infty$; otherwise covering numbers/VC dimension, then apply a standard concentration step. Measurability issues (uncountable $\mathcal{F}$) can be handled via outer probability or separability assumptions.

4.3.2 Step 2 : Symmetrization

Let $(X_i, Y_i)_{i=1}^{2n}$ be independent and identically distributed (i.i.d.) samples, and denote

$$Z_i = (X_i, Y_i), \quad i = 1, 2, \ldots, 2n.$$

For any hypothesis $f \in \mathcal{F}$, define

$$\varphi_f(Z_i) = \mathbb{I}[Y_i \ne f(X_i)].$$

We are interested in the probability

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_i)\right| \ge \epsilon'\right) \tag{4.9}$$

where $\epsilon' = \epsilon/2$. Since $Z_1, \ldots, Z_{2n}$ are i.i.d., any permutation of them has the same joint distribution. Introduce a random permutation $\sigma \in S_{2n}$, giving

$$(Z_{\sigma(1)}, Z_{\sigma(2)}, \ldots, Z_{\sigma(2n)}).$$

Hence, the probability above can be equivalently written as

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_i)\right| \ge \epsilon'\right) \\
=&\mathbb{E}_{\{z_1,\cdots,z_{2n}\}}\left\{\mathbb{P}_{\sigma\in S_{2n}}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_{\sigma_i}) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_{\sigma_i})\right| \ge \epsilon'\right)\right\}
\end{aligned} \tag{4.10}$$

where σ denotes a random permutation.

From the Concentration Inequality for Drawing without Replacement in Lecture 3, we have

$$\begin{aligned}
&P_{\sigma\in S_{2n}}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_{\sigma_i})-\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_{\sigma_i})\right)\geq\epsilon'\right)\\
=&P_{\sigma\in S_{2n}}\left(\left(\frac{2}{n}\sum_{i=1}^{n}\varphi_f(Z_{\sigma_i})-\frac{1}{n}\sum_{i=1}^{2n}\varphi_f(Z_{\sigma_i})\right)\geq\epsilon'\right)\\
=&P_{\sigma\in S_{2n}}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_{\sigma_i})-\frac{1}{2n}\sum_{i=1}^{2n}\varphi_f(Z_{\sigma_i})\right)\geq\epsilon'/2\right)\\
\leq&e^{-O(n\epsilon'^2)}
\end{aligned}\tag{4.11}$$

Definition 4.1.

$$N^{\mathcal{F}}(Z_1,\cdots,Z_n)=|\{(\varphi_f(Z_1),\cdots,\varphi_f(Z_n)),\ f\in\mathcal{F}\}|$$

This shows how much distinction can be generated when modeling samples with all functions f in the hypothesis space $\mathcal{F}$ and reflects the complexity of $\mathcal{F}$.

Definition 4.2.

$$N^{\mathcal{F}}(n)=\max_{Z_1,\cdots,Z_n}N^{\mathcal{F}}(Z_1,\cdots,Z_n)$$

then, we have:

$$\begin{aligned}
&\mathbb{P}_{\sigma\in S_n}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_i)\right|\geq\epsilon'\right)\\
\leq&N^{\mathcal{F}}(Z_1,\cdots,Z_{2n})\cdot\mathbb{P}_{\sigma\in S_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_i)\right|\geq\epsilon'\right)\\
\leq&N^{\mathcal{F}}(2n)\cdot\mathbb{P}_{\sigma\in S_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_i)\right|\geq\epsilon'\right)\\
\leq&N^{\mathcal{F}}(2n)e^{-O(n\epsilon'^2)}
\end{aligned}\tag{4.12}$$

therefore,

$$\mathbb{E}_{\{z_i,\cdots,z_{2n}\}}\left\{\mathbb{P}_{\sigma\in S_n}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi_f(Z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_f(Z_i)\right|\geq\epsilon\right)\right\}\leq N^{\mathcal{F}}(2n)e^{-O(n\epsilon'^2)}\tag{4.13}$$

4.3.3 Step 3 : Characterize $N^{\mathcal{F}}(n)$

Apparently, $N^{\mathcal{F}}(n)$ reflects the complexity of hypothesis space $\mathcal{F}$. Now we need to characterize $N^{\mathcal{F}}(n)$ in a growth manner, i.e., how the amount of distinguishable classifiers increases as the amount of input data increases. We call $N^{\mathcal{F}}(n)$ as "growth function", which measures how many distinct classifications a hypothesis class $\mathcal{F}$ can realize on n data points.

Figure 4.1 depicts a general behavior of growth function $N^{\mathcal{F}}(n)$. For smaller n, the hypothesis class can typically realize all possible 2^n classifications, so the growth function increases exponentially (its logarithm

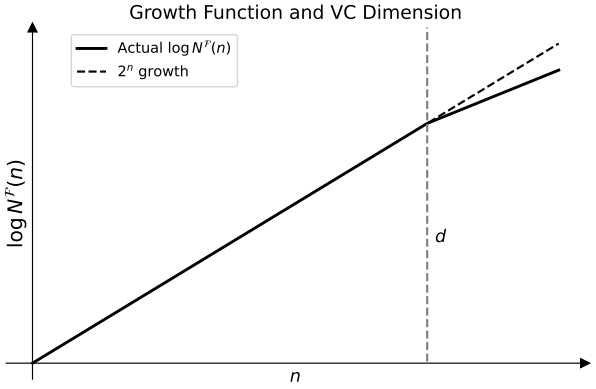

Figure 4.1: Growth Function and VC Dimension

grows linearly with n). However, as n goes beyond a certain critical point d, the hypothesis class can no longer shatter every possible configuration of points. As a result, the growth function starts to grow more slowly and deviates from the ideal exponential trend. d is defined as "VC dimension".

4.3.3.1 The "All Zeros" Special Case

Let $\mathcal{F}$ have VC dimension d. For every set of $d+1$ points $\{z_{i_1}, \ldots, z_{i_{d+1}}\}$, we start from a special case where we fix the all-zeros pattern $(0, \ldots, 0) \in \{0,1\}^{d+1}$ as unrealizable.

Consider any labeling $s \in \{0,1\}^n$. If s contains more than d zeros, then by the pigeonhole principle, there exists some subset of $d+1$ indices where s is identically zero. Since the all-zeros pattern is unrealizable on every $d+1$-point subset, s cannot be realized by $\mathcal{F}$.

Thus, the number of realizable labelings is bounded by the number of strings with at most d zeros:

$$N^{\mathcal{F}}(n) \leq \sum_{k=0}^{d} \binom{n}{k}$$

For fixed d and $n \to \infty$, the dominant term is $\binom{n}{d} = O(n^d)$, hence:

$$N^{\mathcal{F}}(n) = O(n^d) \quad \text{for } n > d$$

4.3.3.2 General Cases

In more general cases, the unrealizable string $s \in \{0,1\}^{d+1}$, and it can be different for different $\{z_{i_k}\}_{k=1}^{d+1}$. Since the overlap between unrealizable strings is smaller (compared to the "all zero" case), the total amount of unrealizable strings is greater than the 'all zero' case. So $N^{\mathcal{F}}(n) = O(n^d), n > d$ still holds.