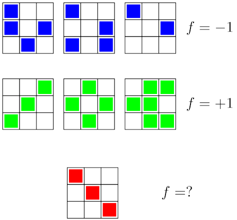
A Simple Visual ML Experiment (2/2)

- **Model 1**
  - $f(\underline{\boldsymbol{x}}) = +1$ when $\underline{\boldsymbol{x}}$ has an axis of symmetry
  - $f(\underline{\boldsymbol{x}}) = -1$ when $\underline{\boldsymbol{x}}$ is not symmetric
  - The test set is symmetrical $\implies f(\underline{\boldsymbol{x}}_0) = +1$
- **Model 2**
  - $f(\underline{\boldsymbol{x}}) = +1$ when the top left square $\underline{\boldsymbol{x}}$ is empty
  - $f(\underline{\boldsymbol{x}}) = -1$ when the top left square $\underline{\boldsymbol{x}}$ is full
  - The test set has top left square full $\implies f(\underline{\boldsymbol{x}}_0) = -1$
- Many functions fit the 6 training examples
  - Some have a value of -1 on the test point, others $+1$
  - Which one is it?
- How can a limited data set reveal enough information to define the entire target function?
  - **Is machine learning possible?**

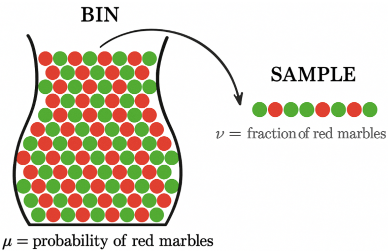


Is Machine Learning Possible?

- The function can assume **any value outside data**
  - E.g., with summer temperature data, the function could assume a different value for winter
- **How to learn an unknown function?**
  - Estimating at unseen points seems impossible in general
  - Requires assumptions or models about behavior
- Difference between:
  - **Possible**
    - No knowledge of the unknown function
    - E.g., could be linear, quadratic, or sine wave outside known data
  - **Probable**
    - Some knowledge of the unknown function from domain knowledge or historical data patterns
    - E.g., if historical weather data forms a sinusoidal pattern, unknown points likely follow that pattern

Supervised Learning: Bin Analogy (1/2)

- Consider a bin with red and green marbles
  - We want to estimate $\Pr(\text{pick a red marble}) = \mu$ where the value of μ is unknown
  - We pick N marbles independently with replacement
  - The fraction of red marbles is ν



- Does ν say anything about μ?
  - **"No"**
    - In strict terms, we don't know anything about the marbles we didn't pick
    - The sample can be mostly green, while the bin is mostly red
    - This is *possible*, but *not probable*
  - **"Yes"**
    - Under certain conditions, the sample frequency is close to the real frequency
- **Possible vs probable**
  - It is **possible** that we don't know anything about the marbles in the bin
  - It is **probable** that we know something
  - Hoeffding inequality makes this intuition formal

Hoeffding Inequality

- Consider a Bernoulli random variable X with probability of success μ
- Estimate the mean μ using N samples with $\nu = \frac{1}{N}\sum_i X_i$
- The **probably approximately correct** (PAC) statement holds:

$$\Pr(|\nu - \mu| > \varepsilon) \leq \frac{2}{e^{2\varepsilon^2 N}}$$

- **Remarks:**
  - Valid for all N and ε, not an asymptotic result
  - Holds only if you sample ν and μ at random and in the same way
  - If N increases, it is exponentially small that ν will deviate from μ by more than ε
  - The bound does not depend on μ
  - Trade-off between N, ε, and the bound:
    - Smaller ε requires larger N for the same probability bound
    - Since $\nu \in [\mu - \varepsilon, \mu + \varepsilon]$, you want small ε with a large probability
  - It is a statement about ν and not μ although you use it to state something about ν (like for a confidence interval)

Supervised Learning: Bin Analogy (2/2)

- Let's connect the bin analogy, Hoeffding inequality, and feasibility of machine learning
  - You know $f(\underline{\boldsymbol{x}})$ at points $\underline{\boldsymbol{x}} \in \mathcal{X}$
  - You choose an hypothesis $h : \mathcal{X} \to \mathcal{Y} = \{0, 1\}$
  - Each point $\underline{\boldsymbol{x}} \in \mathcal{X}$ is a marble
  - You color red if the hypothesis is correct $h(\underline{\boldsymbol{x}}) = f(\underline{\boldsymbol{x}})$, green otherwise
  - The in-sample error $E_{in}(h)$ corresponds to ν
  - The marbles of unknown color corresponds to $E_{out}(h) = \mu$
  - $\underline{\boldsymbol{x}}_1, ..., \underline{\boldsymbol{x}}_n$ are picked randomly and independently from a distribution over $\mathcal{X}$ which is the same as for E_{out}
- Hoeffding inequality holds and bounds the error going from in-sample to out-of-sample

$$\Pr(|E_{in} - E_{out}| > \varepsilon) \leq c$$

  - Generalization over unknown points (i.e., marbles) is possible
  - **Machine learning is possible!**

Validation vs Learning: Bin Analogy

- You have learned that for a given h, in-sample performance $E_{in}(h) = \nu$ needs to be close to out-of-sample performance $E_{out}(h) = \mu$
  - This is the **validation setup**, after you have already learned a model
- In a **learning setup** you have h to choose from M hypotheses
  - You need a bound on the out-of-sample performance of the chosen hypothesis $h \in \mathcal{H}$, regardless of which hypothesis you choose
  - You need a Hoeffding counterpart for the case of choosing from multiple hypotheses

$$
\begin{aligned}
\forall g \in \mathcal{H} = \{h_1, \dots, h_M\} \ & \Pr(|E_{in}(g) - E_{out}(g)| > \varepsilon) \\
&\leq \Pr\left(\bigcup_{i=1}^{M} (|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon)\right) \\
&\leq \sum_{i=1}^{M} \Pr(|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon) && \text{(by the union bound)} \\
&\leq 2M \exp(-2\varepsilon^2 N) && \text{(by Hoeffding)}
\end{aligned}
$$

- **Problem**: the bound is weak

Validation vs Learning: Coin Analogy

- In a **validation set-up**, you have a coin and want to determine if it is fair
- Assume the coin is unbiased: $\mu = 0.5$
- Toss the coin 10 times
- How likely is that you get 10 heads (i.e., the coin looks biased $\nu = 0$)?

$$\Pr(\text{coin shows } \nu = 0) = 1/2^{10} = 1/1024 \approx 0.1\%$$

- **Conclusion**: the probability that the out-of-sample performance ($\nu = 0.0$) is completely different from the in-sample perf ($\mu = 0.5$) is very low

Validation vs Learning: Coin Analogy

- In a **learning set-up**, you have many coins and you need to choose one and determine if it's fair
- If you have 1000 fair coins, how likely is it that at least one appears totally biased using 10 experiments?
  - I.e., out-of-sample performance is completely different from in-sample performance

$$\begin{aligned}
\Pr(\text{at least one coin has } \nu = 0) &= 1 - \Pr(\text{all coins have } \nu \neq 0) \\
&= 1 - \left(\Pr(\text{a coin has } \nu \neq 0)\right)^{10} \\
&= 1 - \left(1 - \Pr(\text{a coin has } \nu = 0)\right)^{10} \\
&= 1 - \left(1 - 1/2^{10}\right)^{1000} \\
&\approx 0.63\%
\end{aligned}$$

- **Conclusion**: It is probable, more than 50%

Validation vs Learning: Hoeffding Inequality

- In **validation / testing**
  - Use Hoeffding to assess how well our g (the *chosen hypothesis*) approximates f (the *unknown hypothesis*):

$$\Pr(|E_{in} - E_{out}| > \varepsilon) \leq 2\exp(-2\varepsilon^2 N)$$

    where:

$$E_{in}(g) = \frac{1}{N}\sum_{i} e(g(\underline{\boldsymbol{x}}_i), f(\underline{\boldsymbol{x}}_i))$$

$$E_{out}(g) = \mathbb{E}_{\underline{\boldsymbol{x}}}[e(g(\underline{\boldsymbol{x}}), f(\underline{\boldsymbol{x}}))]$$

  - Since the hypothesis g is final and fixed, Hoeffding inequality guarantees that you can learn since it gives a bound for E_{out} to track E_{in}
- In **learning**
  - Need to account that our hypothesis is the best of M hypotheses, so:

$$\Pr(|E_{in} - E_{out}| > \varepsilon) \leq 2M\exp(-2\varepsilon^2 N)$$

  - The bound for E_{out} from Hoeffding is weak
- **Questions**:
  - Is the bound weak because it needs to be?
  - Is it possible to replace it with a stricter bound?

Intuition Why Bound for Hoeffding Is Weak

- The Hoeffding inequality and the union bound applied to training set

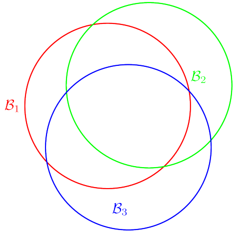
$$\Pr(|E_{in} - E_{out}| > \varepsilon) \leq 2M \exp(-2\varepsilon^2 N)$$

  is **artificially** too loose

- M was coming from the bad event:

$$\begin{aligned}\mathcal{B}_i &= \textit{"hypothesis } h_i \textit{ does not generalize out-of-sample"} \\ &= \text{"}|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon\text{"}\end{aligned}$$

- Since $g \in \{h_1, h_2, \cdots, h_M\}$ then $\Pr(\mathcal{B}) \leq \Pr(\bigcup_i \mathcal{B}_i) \leq \sum_i \Pr(\mathcal{B}_i)$
- The union bound assumes the events are disjoint, leading to a conservative estimate if events overlap
- **In reality**, bad events are extremely overlapping because bad hypotheses are extremely similar



Training vs Testing: College Course Analogy

- In machine learning there is always a training / learning phase and a validation / testing phase
- This set-up is very similar to studying and exams in a college course
- Before the final exam, students receive practice problems and solutions
  - These problems won't appear on the exam
  - Studying the problems improves performance
  - Serve as a "training set" in learning
- Why not give out exam problems to improve performance?
  - Doing well in the exam isn't the goal
  - The goal is to learn the course material
- The final exam isn't strictly necessary
  - Gauges how well you've learned
  - Motivates you to study
  - Knowing exam problems in advance wouldn't gauge learning effectively

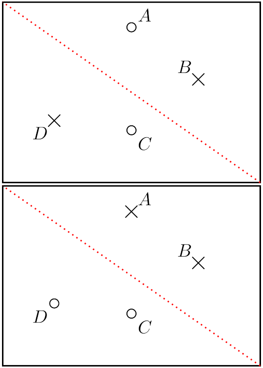
- ***Growth Function***

Dichotomy: Definition

- **Problem**: classify N (fixed) points $\underline{\boldsymbol{x}}_1, \dots, \underline{\boldsymbol{x}}_N$ with an hypothesis set $\mathcal{H}$ of multi-class classifiers
- Consider an assignment D of the points to certain class $\underline{\boldsymbol{d}}_1, \dots, \underline{\boldsymbol{d}}_N$
- D is a **dichotomy** for hypothesis set $\mathcal{H} \iff$ there exists $h \in \mathcal{H}$ that gets the desired classification D
- **Example**
    - 4 points in a plane A, B, C, D
    - Binary classification
    - $\mathcal{H} = \{$ bidimensional perceptrons $\}$
    - Moving the separating hyperplane, you get different classifications for the points (i.e., dichotomies)

	D1	D2	D3	D4	D...
A	o	x	...		
B	x	x			
C	o	o			
D	x	o	...		

    - There are at most 2^N dichotomies
    - Certain classifications are not possible (e.g., XOR assignment)



Dichotomies vs Hypotheses

- An **hypothesis** classifies each point of $\mathcal{X}$: $\mathcal{X} \to \{-1, +1\}$
- A **dichotomy** classifies each point of a fixed set: $\{\underline{\boldsymbol{x}}_1, ..., \underline{\boldsymbol{x}}_N\} \to \{-1, +1\}$
  - Dichotomies are "mini-hypotheses", i.e., hypotheses restricted to given points
  - A dichotomy depends on:
    - The number of points N
    - Hypothesis set $\mathcal{H}$ (i.e., the possible models)
    - Where the points are placed
    - How the points are assigned
- The **number of different dichotomies** is indicated by $|\mathcal{H}(\underline{\boldsymbol{x}}_1, ..., \underline{\boldsymbol{x}}_N)|$
  - The number of dichotomies is always finite, since $|\mathcal{H}(\underline{\boldsymbol{x}}_1, ..., \underline{\boldsymbol{x}}_N)| \leq N^K$
  - The number of hypotheses is usually infinite, i.e., $|\mathcal{H}| = \infty$
- The "complexity" of $\mathcal{H}$ is related to the number of hypothesis
- From the training set point of view what matters are dichotomies and not hypotheses
  - Many (infinite) hypotheses can correspond to the same dichotomy

Growth Function

- The **growth function** counts the maximum number of possible dichotomies on N points for a hypothesis set $\mathcal{H}$:

$$m_{\mathcal{H}}(N) = \max_{\underline{\boldsymbol{x}}_1,\cdots,\underline{\boldsymbol{x}}_N \in \mathcal{X}} |\mathcal{H}(\underline{\boldsymbol{x}}_1,\cdots,\underline{\boldsymbol{x}}_N)|$$

- **Why growth function?**
  - The dichotomies depend on point distribution and assignment
  - The growth function considers the maximum by placing points in the most "favorable way" for the hypothesis set
- To compute $m_{\mathcal{H}}(N)$ by **brute force**:
  - Consider all possible placements of N points $\underline{\boldsymbol{x}}_1, ..., \underline{\boldsymbol{x}}_N$
  - Consider all possible assignments of the points to the classes
  - Consider all possible hypotheses $h \in \mathcal{H}$
  - Compute the corresponding dichotomy for h on $\underline{\boldsymbol{x}}_1, ..., \underline{\boldsymbol{x}}_N$
  - Count the number of different dichotomies

What Can Vary in a Dichotomy

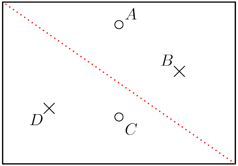
- Given:
  - An hypothesis set $\mathcal{H}$ (e.g., bidimensional perceptrons)
  - N (fixed) points $\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_N$
  - An assignment D of the points to certain class $\underline{\boldsymbol{d}}_1, \ldots, \underline{\boldsymbol{d}}_N$
- D is a **dichotomy** for hypothesis set $\mathcal{H} \iff$ there exists $h \in \mathcal{H}$ that gets the desired classification D
- There are various quantities in the definition of dichotomy
  - The hypothesis set $\mathcal{H}$
    - It is fixed
  - The number of dimensions of the input space
    - It is fixed through the hypothesis set $\mathcal{H}$
  - The number of points N
    - Input to the growth function $m_{\mathcal{H}}(N)$
  - How the points are assigned to the classes $\underline{\boldsymbol{d}}_1, \ldots, \underline{\boldsymbol{d}}_N$
    - It is a free parameter, removed by how each hypothesis in $\mathcal{H}$ "splits" the space
  - Where the points are positioned $\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_N$
    - It is a free parameter, removed by the growth function through max

Growth Function Is Increasing

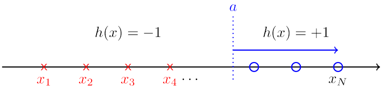
- $m_{\mathcal{H}}(N)$ increases (although not monotonically) with N
- E.g.,
  - The number of dichotomies on $N = 3$ points $m_{\mathcal{H}}(3)$ is smaller or equal than the number of dichotomies on $N = 4$ points
  - In fact we can ignore a new point and get the same classification
- $m_{\mathcal{H}}(N)$ increases with the complexity of $\mathcal{H}$
- $m_{\mathcal{H}}(N)$ increases with the number of dimensions in the input space (i.e., feature space)

Growth Function: Examples

- Consider the growth function $m_{\mathcal{H}}$ for different hypothesis sets $\mathcal{H}$
- **Perceptron on a plane**
  - $m_{\mathcal{H}}(3) = 8$
  - $m_{\mathcal{H}}(4) = 14$ (2 XOR classifications not possible)

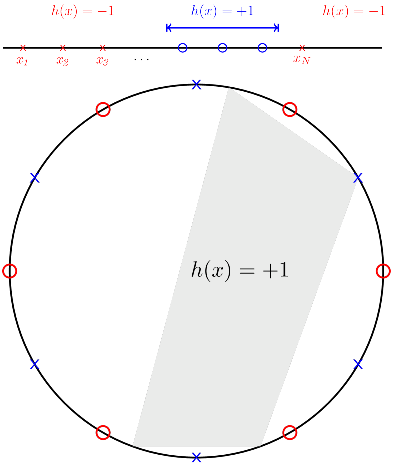


- **Positive rays** $\text{sign}(x - a)$ on $\mathbb{R}$
  - $m_{\mathcal{H}}(N) = N + 1$
  - Origin of rays a can be placed in $N + 1$ intervals



Growth Function: Examples

- Positive intervals on $\mathbb{R}$ $x \in [a, b]$
  - $m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 \sim \mathrm{N}^2$
  - Pick 2 distinct intervals out of $N + 1$, and there is a dichotomy with 2 points in the same interval
- Convex sets on a plane
  - $m_{\mathcal{H}}(N) = 2^N$
  - Place points in a circle and can classify N points in any way



Break Point of an Hypothesis Set

- Given an hypothesis set $\mathcal{H}$
- A hypothesis set $\mathcal{H}$ shatters N points $\iff m_{\mathcal{H}}(N) = 2^N$
  - There is a position of N points that we can classify in any way using $h \in \mathcal{H}$
  - It does not mean all sets of N points can be classified in any way
- k is a break point for $\mathcal{H} \iff m_{\mathcal{H}}(k) < 2^k$
  - I.e., no data set of size k can be shattered by $\mathcal{H}$
- E.g.,
  - For 2D perceptron: a break point is 4
  - For positive rays: a break point is 2
  - For positive intervals: a break point is 3
  - For convex set on a plane: there is no break point