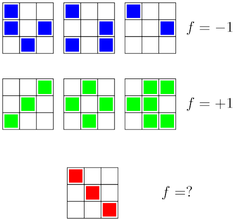
A Simple Visual ML Experiment (2/2)

- **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?**

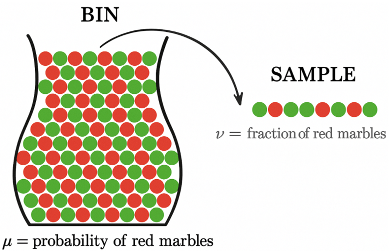


Is Machine Learning Possible?

- 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 Set-Up: 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

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

  - **Problem**: the bound is weak

Validation vs Learning Set-Up: Coin Analogy

- In a **validation set-up**, we 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 we get 10 heads (i.e., the coin looks biased $\nu = 0$)?

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

- In other terms the probability that the out-of-sample performance ($\nu = 0.0$) is very different from the in-sample perf ($\mu = 0.5$) is very low

Validation vs Learning Set-Up: Coin Analogy

- In a **learning set-up**, we have many coins and we need to choose one and determine if it's fair
- If we 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

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

- It is probable, more than 50%

Hoeffding Inequality: Validation vs Learning

- In **validation / testing**
  - We can use Hoeffding to assess how well our g (the chosen hypothesis) approximates f (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 we can learn since it gives a bound for E_{out} to track E_{in}
- In **learning / training**
  - We need to account that our hypothesis is the best of M hypotheses, so the union bound gives:

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

  - The bound for E_{out} from Hoeffding is weak
- Is the bound weak because it needs to be or because the Hoeffding inequality is not good enough?