Review of Lecture 1

- Learning is used when
  - A pattern exists
  - We cannot pin it down mathematically
  - We have data on it

- Focus on supervised learning
  - Unknown target function $y = f(x)$
  - Data set $(x_1, y_1), \cdots, (x_N, y_N)$
  - Learning algorithm picks $g \approx f$ from a hypothesis set $\mathcal{H}$

Example: Perceptron Learning Algorithm

- Learning an unknown function?
  - Impossible 😞. The function can assume any value outside the data we have.
  - So what now?
Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 2: Is Learning Feasible?
Feasibility of learning - Outline

- Probability to the rescue
- Connection to learning
- Connection to real learning
- A dilemma and a solution
A related experiment

- Consider a 'bin' with red and green marbles.

\[ P[\text{picking a red marble}] = \mu \]
\[ P[\text{picking a green marble}] = 1 - \mu \]

- The value of \( \mu \) is unknown to us.

- We pick \( N \) marbles independently.

- The fraction of red marbles in sample = \( \nu \)
Does $\nu$ say anything about $\mu$?

No!
Sample can be mostly green while bin is mostly red.

Yes!
Sample frequency $\nu$ is likely close to bin frequency $\mu$.

possible versus probable
What does $\nu$ say about $\mu$?

In a big sample (large $N$), $\nu$ is probably close to $\mu$ (within $\epsilon$).

Formally,

$$\mathbb{P} \left[ |\nu - \mu| > \epsilon \right] \leq 2e^{-2\epsilon^2 N}$$

This is called **Hoeffding’s Inequality**.

In other words, the statement “$\mu = \nu$” is P.A.C.
\[ \mathbb{P} \left[ |\nu - \mu| > \epsilon \right] \leq 2e^{-2\epsilon^2 N} \]

- Valid for all \( N \) and \( \epsilon \)
- Bound does not depend on \( \mu \)
- Tradeoff: \( N \), \( \epsilon \), and the bound.

\[ \nu \approx \mu \implies \mu \approx \nu \]
Connection to learning

**Bin:** The unknown is a number $\mu$

**Learning:** The unknown is a function $f : \mathcal{X} \rightarrow \mathcal{Y}$

Each marble $\bullet$ is a point $x \in \mathcal{X}$

- : Hypothesis got it right $h(x) = f(x)$
- : Hypothesis got it wrong $h(x) \neq f(x)$
The bin analogy:
Are we done?

Not so fast! \( h \) is fixed.

For this \( h \), \( \nu \) generalizes to \( \mu \).

‘verification’ of \( h \), not learning

No guarantee \( \nu \) will be small.

We need to choose from multiple \( h \)’s.
Multiple bins

Generalizing the bin model to more than one hypothesis:

\[ h_1 \quad h_2 \quad h_M \]

\[ \mu_1 \quad \mu_2 \quad \mu_M \]

\[ \nu_1 \quad \nu_2 \quad \nu_M \]
Notation for learning

Both $\mu$ and $\nu$ depend on which hypothesis $h$

$\nu$ is 'in sample' denoted by $E_{\text{in}}(h)$

$\mu$ is 'out of sample' denoted by $E_{\text{out}}(h)$

The Hoeffding inequality becomes:

$$\mathbb{P} \left[ |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon \right] \leq 2e^{-2\epsilon^2N}$$
Notation with multiple bins

$h_1$ \hspace{2cm} $h_2$ \hspace{2cm} $h_M$

$E_{\text{out}}(h_1)$ \hspace{2cm} $E_{\text{out}}(h_2)$ \hspace{2cm}$E_{\text{out}}(h_M)$

$E_{\text{in}}(h_1)$ \hspace{2cm} $E_{\text{in}}(h_2)$ \hspace{2cm} $E_{\text{in}}(h_M)$
Are we done already? 😊

Not so fast!! Hoeffding doesn’t apply to multiple bins.

What?
Coin analogy

**Question:** If you toss a fair coin 10 times, what is the probability that you will get 10 heads?

**Answer:** $\approx 0.1\%$

**Question:** If you toss 1000 fair coins 10 times each, what is the probability that some coin will get 10 heads?

**Answer:** $\approx 63\%$
From coins to learning

BINGO?
A simple solution

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \mathbb{P}[|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon$$

or $$|E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon$$

or $$|E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon$$

or $$|E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon$$

$$\leq \sum_{m=1}^{M} \mathbb{P}[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon]$$
The final verdict

\[ \mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \sum_{m=1}^{M} \mathbb{P}[|E_{in}(h_m) - E_{out}(h_m)| > \epsilon] \]

\[ \leq \sum_{m=1}^{M} 2e^{-2\epsilon^2 N} \]

\[ \mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N} \]