Introduction to Machine Learning

CS4731
Dr. Mihail
Fall 2017
Slide content based on lecture by Dr. Yaser Abu-Mostafa of Caltech.
`http://work.caltech.edu/telecourse.html`

September 5, 2019

Feasibility

Learning is used when:

- We know a pattern exists
- We don't know the mathematical expression that generated the pattern
- We have **finite** data

Supervised Learning

- Unknown function $y = f(x)$
- Data set $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$
- Learning algorithm picks a $g \approx f$ from a hypothesis set $\mathcal{H}$

- Learn an unknown function?

Supervised Learning

- Unknown function $y = f(x)$
- Data set $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$
- Learning algorithm picks a $g \approx f$ from a hypothesis set $\mathcal{H}$

- Learn an unknown function?
- Impossible. Why?

Supervised Learning

- Unknown function $y = f(x)$
- Data set $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$
- Learning algorithm picks a $g \approx f$ from a hypothesis set $\mathcal{H}$

- Learn an unknown function?
- Impossible. Why?
- The function can take on any value outside of the data we have.

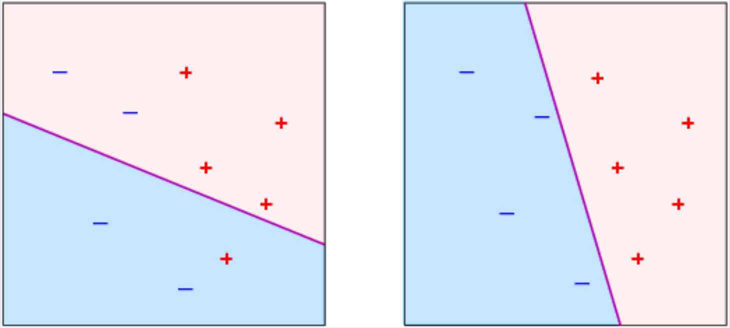
Supervised Learning

- Unknown function $y = f(x)$
- Data set $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$
- Learning algorithm picks a $g \approx f$ from a hypothesis set $\mathcal{H}$

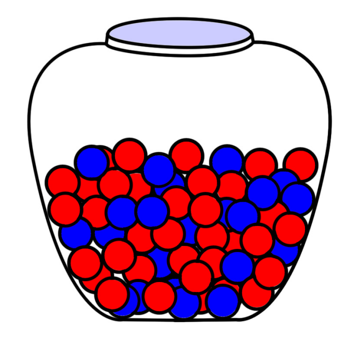
- Learn an unknown function?
- Impossible. Why?
- The function can take on any value outside of the data we have.

What if no pattern exists?

- Learning algorithm will still work, but won't learn anything.
- The algorithm should tell us if/when that is the case.



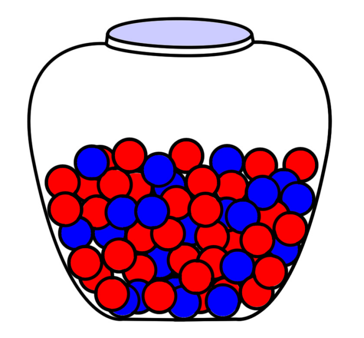
Example



Sample

- There exists a probability μ for picking a red marble: $P(redmarble) = \mu$

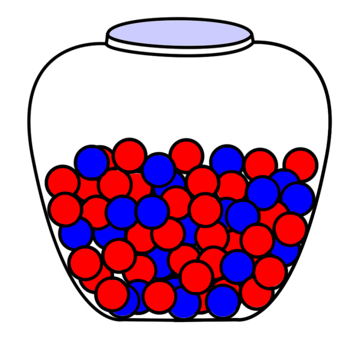
Example



Sample

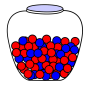
- There exists a probability μ for picking a red marble: $P(redmarble) = \mu$
- What is $P(bluemarble)$

Example



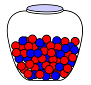
Sample

- There exists a probability μ for picking a red marble: $P(redmarble) = \mu$
- What is $P(bluemarble) = 1 - \mu$



Sample

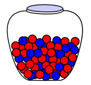
- The value μ is unknown, and we pick N marbles independently with replacement
- The fraction of red marbles is ν



Sample

- The value μ is unknown, and we pick N marbles independently with replacement
- The fraction of red marbles is ν

Does ν say anything about *mu*?

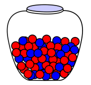


Sample

- The value μ is unknown, and we pick N marbles independently with replacement
- The fraction of red marbles is ν

Does ν say anything about *mu*?

- No!

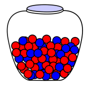


Sample

- The value μ is unknown, and we pick N marbles independently with replacement
- The fraction of red marbles is ν

Does ν say anything about *mu*?

- No!All samples can be blue.
- Yes!



Sample

- The value μ is unknown, and we pick N marbles independently with replacement
- The fraction of red marbles is ν

Does ν say anything about *mu*?

- No!All samples can be blue.
- Yes!Possible vs. probable! Intuition: more samples give you more certainty.

How many samples? Large N

How is μ close to ν?

$$|\mu - \nu| < \epsilon \tag{1}$$

How many samples? Large N

How is μ close to ν?

$$|\mu - \nu| < \epsilon \tag{1}$$

Bad situation

$$P(bad\,event) \leq$$

How many samples? Large N

How is μ close to ν?

$$|\mu - \nu| < \epsilon \tag{2}$$

How many samples? Large N

How is μ close to ν?

$$|\mu - \nu| < \epsilon \tag{2}$$

Bad situation

$$P(|\nu - \mu| > \epsilon) \leq \text{small number}$$

How many samples? Large N

Hoeffding's Inequality

$$P(|\nu - \mu| > \epsilon) \leq 2e^{-2e^2 N}$$

How many samples? Large N

Hoeffding's Inequality

$$P(|\nu - \mu| > \epsilon) \leq 2e^{-2e^2 N}$$

Plain English

The statement that $\nu = \mu$ is probably almost correct.

- Valid for all N and ϵ
- Bound does not depend on μ
- Smaller ϵ, the bigger N we need to be sure ν is close μ

Hoeffding Inequality

It does not apply to multiple hypotheses!

Consider a fair coin. Toss 10 times. What is the probability of getting 10 heads? What is the probability of one person getting 10 heads if 1000 people do it?

- Consider multiple hypotheses, $h_1, h_2, ..., h_M$. ν and μ depend on h.
  - h_1: $\nu = 0.2$
  - h_2: $\nu = 0.4$
  - h_m: $\nu = 0.1$
- ν is "in sample", called $E_{in}(h)$
- μ is "out of sample", called $E_{out}(h)$

Hoeffding Inequality

Single Hypothesis

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2e^2 N}$$

Hoeffding Inequality

Single Hypothesis

$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2e^2N}$

Picking a final hypothesis g

Worst case:

$$
\begin{aligned}
P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \quad & P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \\
& \textbf{or } |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \\
& \textbf{or } |E_{in}(h_3) - E_{out}(h_3)| > \epsilon \\
& \dots \\
& \textbf{or } |E_{in}(h_M) - E_{out}(h_M)| > \epsilon) \\
& \leq \sum_{m=1}^{M} P(|E_{in}(h_m) - E_{out}(h_M)| > \epsilon)
\end{aligned}
$$

Hoeffding Inequality

Single Hypothesis

$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2e^2 N}$

Picking a final hypothesis g

Worst case:

$$
\begin{aligned}
P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \quad & P(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \\
& \textbf{or } |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \\
& \textbf{or } |E_{in}(h_3) - E_{out}(h_3)| > \epsilon \\
& \dots \\
& \textbf{or } |E_{in}(h_M) - E_{out}(h_M)| > \epsilon) \\
& \leq \sum_{m=1}^{M} P(|E_{in}(h_m) - E_{out}(h_M)| > \epsilon)
\end{aligned}
$$

Finally

$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq \sum_{m=1}^{M} 2e^{-2e^2 N} = 2Me^{-2e^2 N}$

Conclusion

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq \sum_{m=1}^{M} 2e^{-2e^2 N} = 2Me^{-2e^2 N}$$

Model Complexity

- Sophisticated models mean high M, the more sophisticated the model, the more likely you will learn sample space and not generalize.
- The difficulty in choosing the right method is based on the above intuition.