MONASH University

ETC3555

Statistical Machine Learning

The learning problem

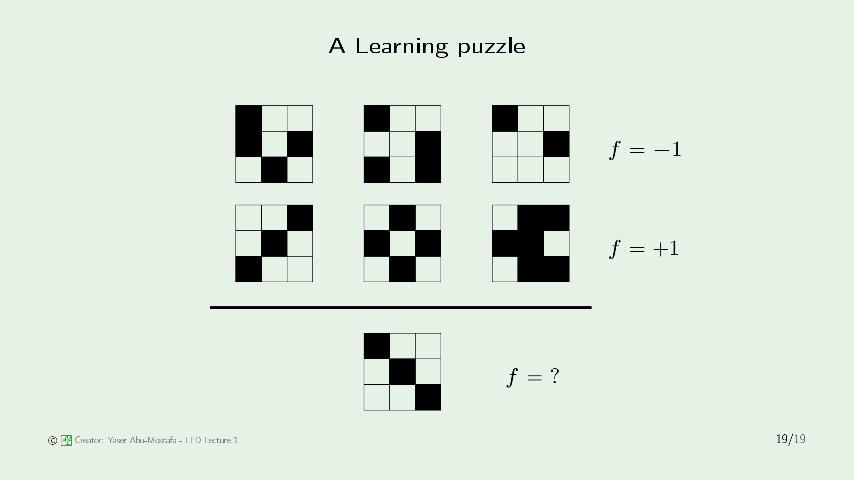
31 July 2018

Outline

Is learning feasible?

The target function is *unknown*. How could a limited data set reveal enough information to pin down the entire target function?

A learning puzzle



Do you obtain -1 or +1?

Is learning feasible?

- More than one function fits the 6 training examples.
  - If the true f is +1 when the pattern is symmetric, then the solution is +1
  - If the true f is -1 when the top left square of the pattern is white, then the solution is -1
- We know the values of f on all the points in the training data $\mathcal{D}$. But since f is an unknown function, f remains unknown outside of $\mathcal{D}$.
- The whole purpose of learning f is to be able to predict the value of f on new points.
- Is learning feasible? Yes, in a *probabilistic sense*.

Outline

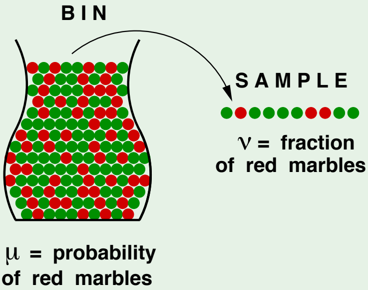
A related experiment

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

  $\mathbb{P}[$ picking a red marble $] = \mu$

  $\mathbb{P}[$ picking a green marble $] = 1 - \mu$

- The value of μ is unknown to us.
- We pick N marbles independently.
- The fraction of red marbles in sample $= \nu$



 3/17

Bin and marbles

Does ν say anything about μ?

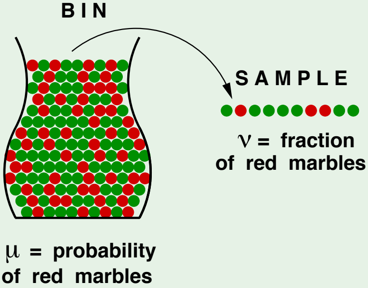
No!

Sample can be mostly green while bin is mostly red.

Yes!

Sample frequency ν is likely close to bin frequency μ.

possible versus probable



4/17

Hoefdding's inequality

What <u>does</u> ν say about μ?

In a big sample (large N), ν is probably close to μ (within ϵ).

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.

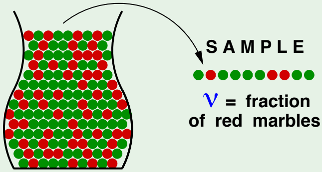
 5/17

P.A.C = Probably Approximately Correct

Hoefdding's inequality

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

- Valid for all N and ϵ
- Bound does not depend on μ
- Tradeoff: N, ϵ, and the bound.
- $\nu \approx \mu \implies \mu \approx \nu$ ☺



6/17

Two rules in probability

Let $\mathcal{B}_1, \mathcal{B}_2$ be any two events. If $\mathcal{B}_1 \implies \mathcal{B}_2$ (i.e. event $\mathcal{B}_1$ implies event $\mathcal{B}_2$), then

$$\mathcal{P}(\mathcal{B}_1) \leq \mathcal{P}(\mathcal{B}_2).$$

Let $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_M$ be any M events, then

$$\mathcal{P}(\mathcal{B}_1 \text{ or } \mathcal{B}_1 \text{ or } \ldots \text{ or } \mathcal{B}_M) \leq \mathcal{P}(\mathcal{B}_1)+\mathcal{P}(\mathcal{B}_2)+\cdots+\mathcal{P}(\mathcal{B}_M).$$

The second rule is known as the *union bound* or *Boole's inequality*.

Exercise

- If $\mu = 0.9$, what is the probability that a sample of 10 marbles will have $\nu \leq 0.1$? [Hint: use a binomial distribution]
- If $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have $\nu \leq 0.1$ and compare the answer to the previous exercise. [Hint: Use one of the previous rule]

Outline

From marbles to learning

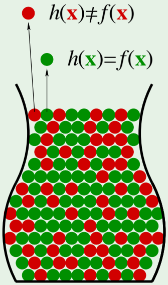
Connection to learning

Bin: The unknown is a number μ

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

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

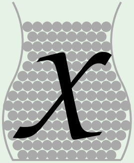
- ● : Hypothesis got it right $h(\mathbf{x}) = f(\mathbf{x})$
- ● : Hypothesis got it wrong $h(\mathbf{x}) \neq f(\mathbf{x})$

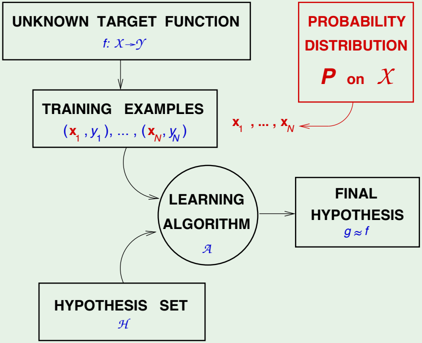


Learning diagram updated

Back to the learning diagram

The bin analogy:





8/17