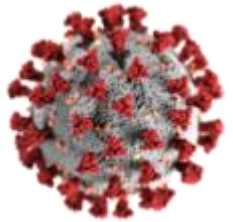
CPSC 440: Machine Learning

Binary Density Estimation

Winter 2022

Motivation: COVID-19 Prevalence



- Want to know prevalence of COVID-19 in a population.
  - For example, what percentage of UBC students have it right now?

- "Brute force" approach:
  - Grab and test every single student, compute proportion that tests positive.

- Statistical approach:
  - Grab an "independent and identically distributed" (IID) sample of students.
  - Estimate the proportion that have it based on the sample.

When I use other people's images, the links are here

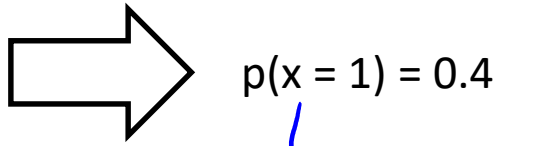
https://en.wikipedia.org/wiki/COVID-19

General Problem: Binary Density Estimation

- This is a special case of binary density estimation:
  - Input: 'n' IID samples of binary values x^1, x^2, x^3,…, x^n from population.
  - Output: model of probability that x=1.
- Binary density estimation as a picture:

X =

COVID-19?
1
0
0
1
0



p(x = 1) = 0.4

I am using 'x' here as a generic IID sample from population

- We'll spend several lectures discussing big concepts in this simple case.
  - And we will slowly build to more-complicated cases.
    - Going beyond binary, more than one variable, conditional versions, deep versions, and so on.

Other Applications of Binary Density Estimation

- Other applications where binary density estimation is useful:
  1. What is the probability that this medical treatment works?
     - Does it work 60% of the time? Does it work 99% of the time?
  2. What is the probability of at least one "success" after 10 tries?
     - For example, if you plant 10 seeds will at least one germinate?
  3. What is the expected number of "tries" before the first success?
     - For example, how many lottery tickets do you expect to buy before you win?

- Item 1 we use the model to compute p(x = 1), as in COVID-19 example.
- Items 2 and 3 use p(x=1) to compute some other quantity.
  - In all 3 cases, in ML we call this "inference" with the model.
    - Inference is a broad term, that basically means "doing calculations with a model".