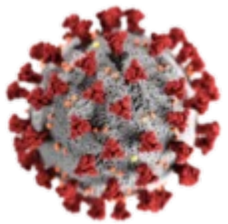
CPSC 440: Machine Learning

Binary Density Estimation

Winter 2023

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 $x_1, x_2, x_3, \ldots, x_n$ from population.
  - Output: a probability model for a random X: here, just $\Pr(X = 1)$.
- Binary density estimation as a picture:

X: a generic sample from the iid population (random variable)

X =

COVID-19?
1
0
0
1
0

density est

$\Pr(X = 1) = 0.4$

X: the $n \times 1$ matrix of our sample data $(x_1, x_2, \ldots, x_n)$

- 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.

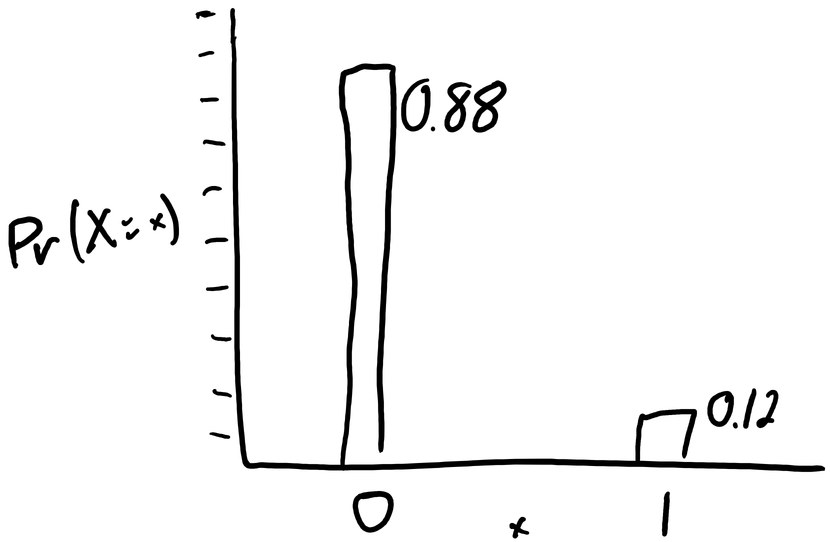
Other Applications of Binary Density Estimation

- 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 $\Pr(X = 1)$, as in COVID-19 example.
- Items 2 and 3 use $\Pr(X = 1)$ to compute some other quantity.
  - In ML, we call all three cases "inference" with the model.
    - Inference is a broad term; it basically means "doing calculations with a model".

Model Definition: Bernoulli Distribution

- Models for binary density estimation need a parameterization.
  - A probability model based on some "parameters."
- For binary variables, we usually use the Bernoulli distribution:
  - We say that X follows a Bernoulli with parameter θ, X ~ Bern(θ), if $\Pr(X = 1 \mid \theta) = \theta$.
  - So if θ = 0.12 in the COVID-19 example, we think 12% of population has COVID-19.



  - To define a valid probability, we require that θ is between 0 and 1 (inclusive).

bonus!

Digression: “Inference” in Statistics vs. ML

- In machine learning, people often use this terminology:
  - “Learning” is the task of going from data **X** to parameter(s) θ.
  - “Inference” is the task of using the parameter(s) to infer/predict something.
- In statistics, people sometimes use a “reverse” terminology:
  - “Inference” is the task of going from data **X** to parameter(s) θ.
  - “Prediction” is the task of using the parameters to infer/predict something.
- This partially reflects historical views of both fields:
  - Statisticians often focused on finding the parameters.
  - ML hackers often focused on making predictions.
- And some people also use “inference” to refer to both tasks!
  - But, this course will use the machine learning terminology.

Inference Task: Computing Probabilities

- Inference task: given θ, compute $\Pr(X = 0 \mid \theta)$.
- We'll also write this as $P(0 \mid \theta)$
  - Be careful you know what we're abbreviating! ("Explicit is better than implicit")
- Recall that probabilities add up to 1:

$$\underbrace{\Pr(x=1 \mid \theta) + \Pr(x=0 \mid \theta)}_{\text{summing over all values of 'x'}} = 1$$

- Using the "sum to one" property to solve the above inference task:

$$\Pr(x=0 \mid \theta) = 1 - \underbrace{\Pr(x=1 \mid \theta)}_{\theta} = 1 - \theta$$

- So for the Bernoulli distribution we have $\Pr(X = 0 \mid \theta) = 1 - \theta$.
  - If $\theta = 0.12$ in the COVID-19 case, we think $1 - 0.12 = 0.88$ does not have disease.

Bernoulli Distribution Notation

- We can write both cases, $\Pr(X = 1 \mid \theta) = \theta$ and $\Pr(X = 0 \mid \theta)=1 - \theta$, as

$$P(x \mid \theta) = \underbrace{\theta^{x}}_{\text{If } x=0\text{, this is } \theta^0=1\text{, so becomes irrelevant}} \underbrace{(1-\theta)^{1-x}}_{\text{If } x=1\text{, becomes } (1-\theta)^0=1}$$

x is either 0 or 1

- Another notation you might see uses an "indicator function":

$$P(x \mid \theta) = \theta^{\mathbb{1}(x=1)} (1-\theta)^{\mathbb{1}(x=0)}$$

– $\mathbb{1}$(condition) is a function that is 1 if "something" is true, and 0 otherwise.

Inference Task: Computing Dataset Probabilities

- Inference task: given θ and IID data, compute $P(x_1, x_2, \ldots, x_n \mid \theta)$.
  - This is called the "likelihood": $\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \theta)$
    - Many ways to estimate θ require us to compute this, e.g. "maximum likelihood estimation".
    - We may want to compute this on validation/test data to compare models.

- Assuming "independence of IID data given parameters", we have

$$p(x^1, x^2, \ldots, x^n \mid \theta) = \prod_{i=1}^{n} p(x^i \mid \theta)$$

  - Technically, this is a "conditional independence" assumption.
    - We will discuss later why the x_i being IID implies this conditional independence holds.

bonus!

Inference Task: Computing Dataset Probabilities

- Let's use the independence property to compute P(1, 0, 1, 1, 0 | θ):

$$P(x^1, x^2, \ldots, x^n \mid \theta) = \prod_{i=1}^{n} P(x^i \mid \theta)$$

$$= P(x^1 \mid \theta) P(x^2 \mid \theta) P(x^3 \mid \theta) P(x^4 \mid \theta) P(x^5 \mid \theta)$$

$$= \theta \quad (1-\theta) \quad \theta \quad \theta \quad (1-\theta)$$

$$= \theta^3 (1-\theta)^2$$

- Abstract ways to write this for a generic dataset of *n* examples:

$$P(\mathbf{X} \mid \theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{\sum_{i=1}^{n} (1-x_i)}$$

use '$\mathbf{X}$' for the whole dataset

n_1: "number of 1 values"

$$P(\mathbf{X} \mid \theta) = \theta^{n_1} (1-\theta)^{n_0}$$

$$p(x_1, x_2, \ldots, x_n \mid \theta) = \theta^{\sum_{i=1}^{n} \mathbb{1}(x_i=1)} (1-\theta)^{\sum_{i=1}^{n} \mathbb{1}(x_i=0)}$$

with indicator functions

Inference Task: Computing Dataset Probabilities

- So given θ, we can compute probability of dataset **X** as: $P(X \mid \theta) = \theta^{n_1}(1-\theta)^{n_0}$
- Implementing this in code:

First try:

```
n1=0
n0=0
for i in 1:n
    if X[i] == 1
        n1 += 1
    else
        n0 += 1
    end
end
p=(theta**n1)*(1-theta)**n0
```

Nicer version:

```
n1 = sum(X)
n0 = n - n1
logp = n1*log(theta)+n0*log(1-theta)
```

- Computational complexity: $O(n)$.
  – You do a simple addition for each of the n elements, then do some simple operations to get final value.
- Notice that the "nicer version" returns the logarithm, log(P(X | θ)).
  – If n is large and/or θ is close to 0 or 1, the probability will be very small.
    - Calculation might underflow and return 0 due to truncation in floating point arithmetic.
  – With logarithm, you will still be able to compare different θ values.

Inference Task: Finding the mode ("decoding")

- Inference task: given θ, find x that maximizes $P(x \mid \theta)$.
  - "What's most likely to happen?" It's finding the mode; also called decoding

- For Bernoulli models:
  - If $\theta < 0.5$, the mode is x= 0.
    - If $\theta = 0.12$, it is more likely that a random person **does not** have COVID-19.
  - If $\theta > 0.5$, the mode is x = 1.
    - If $\theta = 0.6$, it is more likely that a random person **does** have COVID-19.
  - If $\theta = 0.5$, both x=1 and x=0 are both valid decodings.

- Decoding is not very exciting for Bernoulli models.
  - It is more-difficult for more-complicated models, and it will be important later.
  - In supervised learning, you sometimes want to make predictions using the mode.

Inference Task: Most Likely Dataset

- Inference task: given θ, find **X** that maximizes $P(x_1, x_2, \ldots, x_n \mid \theta)$.
  - "What set of training examples are we most likely to observe"?

- Recall that we showed: $P(X \mid \theta) = \theta^{n_1}(1-\theta)^{n_0}$

- If $\theta < 0.5$, then the decoding is $x_1=0$, $x_2=0$, $x_3=0$, $x_4=0$, $x_5=0$, $x_6=0$,…
  - We maximize $P(\mathbf{X} \mid \theta)$ by making n_0 as big as possible and n_1 as small as possible.
  - In the "most likely" set of sample with θ=0.12, nobody has COVID-19!

- The dataset mode usually does not represent "typical" behavior.
  - For example, if θ=0.12 we should expect 12% of samples to be 1, not 0%!
  - Decoding has the "highest" probability, but that probability might be really low.
    - There are many datasets with 1 values, but each has a lower probability than "all zeros".

Inference Task: Sampling

- Inference task: given θ,
  generate samples of X distributed according to $p(X \mid \theta)$.
  - This is called sampling from the distribution.

- Sampling is the "opposite" of density estimation:

$p(x = 1) = 0.4$ Sample

COVID-19?
1
0
0
1
0

- You are given the model, and your job is to generate IID examples.
  - Often write code to generate one IID sample, then call it many times.

bonus!

Digression: Motivation for Sampling

- Sampling is not especially interesting for Bernoulli distributions.
  - Because knowing θ tells you everything about the distribution.

- But sampling will let us do neat things in more-complicated density models:
  - thispersondoesnotexist.com, DALLE, ChatGPT, …

  - Sampling often gives indications about whether the model is reasonable.
    - If samples look nothing like the data, then model is not very good.

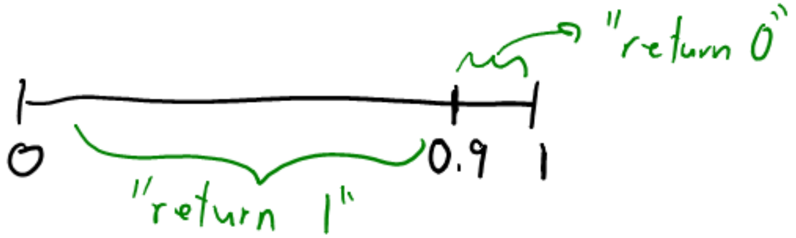
https://thispersondoesnotexist.com/

Inference Task: Sampling

- Basic ingredient of all sampling methods:
  - We assume we can sample uniformly on the interval between 0 and 1.
  - In practice, we use a "pseudo-random" number generator.
    - `rng = np.random.default_rng(); rng.random()`
    - We won't talk about how this works

- Consider sampling from a Bernoulli with θ = 0.9.
  - 90% of the time our sampler should produce a 1.
  - 10% of the time our sampler should produce a 0.

- How to generate a 1 in 90% of samples based on uniform sampling?
  1. Generate a uniform sample (between 0 and 1).
  2. If the sample is less than 0.9, return 1.
     - Otherwise, return 0.



Inference Task: Sampling

- Sampling from a Bernoulli with generic θ value:
  - Generate a sample uniformly on the interval between 0 and 1.
  - If the sample is less than θ, return 1.
    - Otherwise, return 0.
- In code:

```
u = rng.random()
if u <= theta
    x = 1
else
    x = 0
```

```
X = 1 if rng.random() <= theta else 0
```

```
X = (rng.random(t) <= theta).astype(int)
```

- Cost is $O(1)$, assuming that random number generator costs $O(1)$.
  - To generate t samples, call the function t times. Cost in this case is $O(t)$.

Next Topic: Maximum Likelihood Estimation

MLE: Binary Density Estimation

- We have discussed how to use a Bernoulli model ("inference").
- Now we will consider how to train a Bernoulli model ("learning").
  - Goal is to go from samples to an estimate of parameter θ:

X =

COVID-19?
1
0
0
1
0

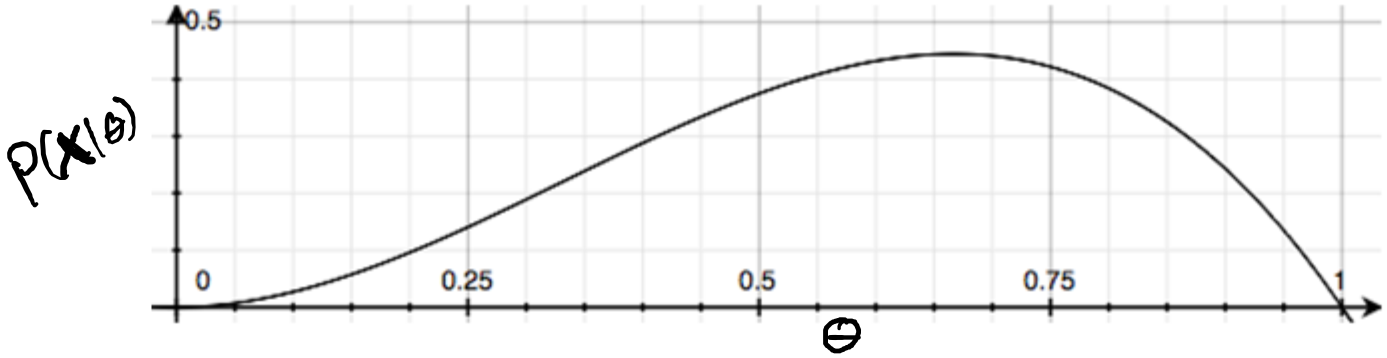
MLE

$\theta = 0.4$

- Classic way to find parameters (used in the picture above):
  - Maximum likelihood estimation (MLE).

The Likelihood Function

- The likelihood function is the probability of the data given parameters.
  - In the Bernoulli model, we showed earlier that our likelihood is: $P(X|\theta)=\theta^{n_1}(1-\theta)^{n_0}$
    - The probability of seeing the data **X** if our Bernoulli parameter is θ.
- Here is a plot of the likelihood if our IID data is x_1=1, x_2=1, x_3=0.

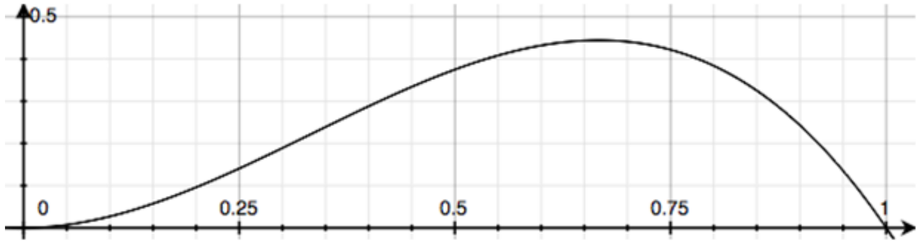


  - For θ = 0.5, the likelihood is P(1, 1, 0 | θ = 0.5) = (1/2)(1/2)(1/2) = 0.125.
  - If θ = 0.75, then P(1, 1, 0 | θ = 0.75) = (3/4)(3/4)(1/4) ≈ 0.14 (dataset is more likely for $\theta = 0.75$ than 0.5).
  - If θ = 0 ("always 0"), then P(1, 1, 0 | θ = 0) = 0 (dataset is not possible for $\theta = 0$).
    - Data has probability 0 if θ=0 or θ=1 (since we have a 1 and a 0 in the data).
  - Data doesn't have highest probability at 0.5 (because we have more 1s than 0s).
  - Note that this is a probability distribution over **X**, not θ (area under the curve is not 1).

Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation (MLE):
  - Choose the parameters that have the highest likelihood, P(**X** | θ).
    - "Find the parameter(s) θ under which the data **X** was most likely to be seen."

- The likelihood from the previous slide with x_1=1, x_2=1, x_3=0:



  - In this example, MLE is θ = 2/3.

- The MLE for general Bernoulli is $\theta = n_1/(n_1 + n_0)$.
  - "If you flip a coin 50 times and it lands heads 23 times, I'll guess that prob('head') is 23/50."

Derivation of MLE for Bernoulli

- Let's derive the MLE for Bernoulli.
  - This will seem overly-complicated for such a simple result.
  - But the same steps can be used in more-complicated situations.

- MLE "finds the argument" maximizing the likelihood function:

$$\hat{\theta} \in \underbrace{\underset{\theta}{\operatorname{argmax}}}_{\text{"argmax" means "find the values that achieve the maximum"}} \{ \underbrace{\theta^{n_1}(1-\theta)^{n_0}}_{\text{likelihood for data with counts } n_1 \text{ and } n_0.} \}$$

Our estimate of θ based on data

There be more than one element in argmax. We say you "pick one in the set"

"argmax" returns a set containing all the values θ that give maximum value.

Digression: Maximizing the Log-Likelihood

- Instead of finding an element maximizing the likelihood:

$$\hat{\Theta} \in \underset{\Theta}{\operatorname{argmax}} \{ p(X \mid \Theta) \}$$

- We usually find an element maximizing the log of the likelihood:

$$\hat{\Theta} \in \underset{\Theta}{\operatorname{argmax}} \{ \log(p(X \mid \Theta)) \}$$

  – People often say "log-likelihood" as a short version of "log of the likelihood".

- Both approaches give the same solution.
  – Because logarithm is "strictly monotonic" over positive values.
    - If $\alpha > \beta$, then $\log(\alpha) > \log(\beta)$.
    - See notes on course webpage about "Max and Argmax" for details.
  – And logarithm is nicer numerically since likelihood is usually really close to 0.

Derivation MLE for Bernoulli

- MLE for Bernoulli by maximizing the likelihood:

$$\hat{\Theta} \in \underset{\Theta}{\operatorname{argmax}} \left\{ \Theta^{n_1} (1-\Theta)^{n_0} \right\}$$

- MLE for Bernoulli by maximizing the log-likelihood:

$$\hat{\Theta} \in \underset{\Theta}{\operatorname{argmax}} \left\{ \log\left(\Theta^{n_1} (1-\Theta)^{n_0} \right) \right\}$$

"the sets are equivalent"

$$\equiv \underset{\Theta}{\operatorname{argmax}} \left\{ \log(\Theta^{n_1}) + \log((1-\Theta)^{n_0}) \right\}$$

using $\log(\alpha\beta) = \log(\alpha) + \log(\beta)$

$$\equiv \underset{\Theta}{\operatorname{argmax}} \left\{ n_1 \log(\Theta) + n_0 \log(1-\Theta) \right\}$$

using $\log(\alpha^\beta) = \beta \log(\alpha)$

Derivation MLE for Bernoulli

- From the last slide we want to find:

$$\hat{\theta} \in \underset{\theta}{\operatorname{argmax}} \{ n_1 \log(\theta) + n_0 \log(1-\theta) \}$$

- Recall that a maximum must have derivative equal to zero.
  - Equating the derivative of the log-likelihood with zero:

$$0 = \underbrace{\frac{n_1}{\theta}}_{\substack{\text{derivative of} \\ n_1 \log\theta \text{ for } \theta > 0}} - \underbrace{\frac{n_0}{1-\theta}}_{\substack{\text{derivative of} \\ n_0 \log(1-\theta) \text{ for } 1-\theta > 0}}$$

  - Using HS math: $0 = n_1(1-\theta) - n_0\theta \Rightarrow (n_1+n_0)\theta = n_1 \Rightarrow \theta = \frac{n_1}{n_1+n_0} = \frac{n_1}{n}$

Since $n_1 + n_0 = n$

Summary

- Binary density estimation:
  - Modeling $\Pr(X = 1)$ given IID samples $x_1, x_2, \ldots, x_n$.
- Bernoulli distribution:
  - Probability distribution over a binary variable.
  - Parameterized by a number θ such that $\Pr(X=1 \mid \theta) = \theta$.
- Inference:
  - Computing a quantity based on a model.
  - Examples include computing probabilities, decoding, and sampling.
- Maximum likelihood estimation (MLE):
  - Estimate parameters by maximizing probability of data given parameters.
  - For Bernoulli, sets θ = (number of 1s)/(number of examples).

- Next time: more boring definitions.