CPSC 440: Machine Learning

Bernoulli Distribution
Winter 2022
Last Time: Binary Density Estimation

• We introduced the problem of **binary density estimation**:
  – Give IID samples for a binary variable, estimate proportion of “1” values.

<table>
<thead>
<tr>
<th>COVID-19?</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

  \[ p(x = 1) = 0.4 \]

• We can then do **inference** with the model:
  – Compute probability that at least one among 10 people has COVID-19.
  – Compute number you would need to recruit to expect to get 50 cases.
Model Definition: Bernoulli Distribution

- Models for binary density estimation need a **parameterization**.
  - A way to go from some “parameters” to the probability ‘p’.

- For binary variables, we usually use the **Bernoulli distribution**:
  - We say that x follows a Bernoulli with parameter $\theta$ if $p(x = 1 \mid \theta) = \theta$.
  - So if $\theta = 0.12$ in the COVID-19 example, we think 12% of population has COVID-19.

- To define a valid probability, we require that $\theta$ is between 0 and 1 (inclusive).
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) $\theta$.
  – “Inference” is the task of using the parameter(s) to infer/predict something.

• In statistics, people often use the reverse terminology:
  – “Inference” is the task of going from data ‘X’ to parameter(s) $\theta$.
  – “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 $\theta$, compute $p(x = 0 \mid \theta)$.

• Recall that probabilities add up to 1 over discrete domains:

\[
p(x=1 \mid \theta) + p(x = 0 \mid \theta) = 1
\]

• Using the “sum to one” property to solve the above inference task:

\[
p(x = 0 \mid \theta) = 1 - p(x=1 \mid \theta) = 1 - \theta
\]

• So for the Bernoulli distribution we have $p(x = 0 \mid \theta) = 1 - \theta$.

• 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, \( p(x = 1 \mid \theta) = \theta \) and \( p(x = 0 \mid \theta) = 1 - \theta \), as:

\[
p(x \mid \theta) = \theta^x (1 - \theta)^{1-x}
\]

• Another notation you might see uses an “indicator function”:

\[
p(x \mid \theta) = \theta \mathbb{I}[x = 1] (1 - \theta) \mathbb{I}[x = 0]
\]

– \( \mathbb{I}[\text{something}] \) is a function that is 1 if “something” is true, and 0 otherwise.
Inference Task: Computing Dataset Probabilities

- **Inference task**: given $\theta$ and IID data, compute $p(x^1, x^2, ..., x^n | \theta)$.
  - Notation warning: in this class I use superscripts for the example number.
    - Different than CPSC 340, where we use subscripts like $x_i$.
  - Why do we care about this quantity?
    - Many ways to estimate $\theta$ require us to compute this “likelihood” of the training data.
      - Such as “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
  \[
p(x^1, x^2, ..., x^n | \theta) = \prod_{i=1}^n p(x^i | \theta)
  \]
  - Technically, this is a “conditional independence” assumption.
    - We will discuss later why the $x^i$ being IID implies this conditional independence holds.
Inference Task: Computing Dataset Probabilities

• Let’s use the independence property to compute \( p(1, 0, 1, 1, 0 \mid \theta) \):

\[
p(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) \\
= p(x_1 \mid \theta) p(x_2 \mid \theta) p(x_3 \mid \theta) p(x_4 \mid \theta) p(x_5 \mid \theta) \\
= \theta^3 (1-\theta)^2
\]

• Abstract ways to write this for a generic dataset of ‘n’ examples:

\[
p(X \mid \theta) = \theta^{\hat{x}} (1-\theta)^n \\
p(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) \\
p(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} I(x_i = 1) (1-\theta)^n
\]
Inference Task: Computing Dataset Probabilities

• So given \( \theta \), we can compute probability of dataset ‘X’ as:
  \[
p(X | \theta) = \theta^n (1-\theta)^{n_0}
\]

• Implementing this in code:

  **First try:**
  ```plaintext
  N = 0
  n_0 = 0
  for i in n:
    if X[i] == 1
      n_1 += 1
    else
      n_0 += 1
  end
  p = (n_1 * \theta^n_1) * (1-\theta)^{n_0}
  ```

  **Nicest version:**
  ```plaintext
  n_1 = \sum(X)
  n_0 = n - n_1
  \log p = n_1 * \log(\theta) + n_0 * \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 logarithm, \( \log(p(X | \theta)) \).
  – If ‘n’ is large and/or \( \theta \) 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 \( \theta \) values.
Inference Task: Decoding

• **Inference task**: given $\theta$, find ‘x’ that maximizes $p(x \mid \theta)$.  
  – This is called **decoding**: “what is most likely happen?”

• For Bernoulli models:
  – If $\theta < 0.5$, the decoding 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 decoding 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 often want to **make predictions using the decoding.**
Inference Task: Decoding Dataset

• **Inference task**: given \( \theta \), find ‘x’ that maximizes \( p(x^1, x^2, ..., x^n \mid \theta) \).
  – “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 \( p(X \mid \theta) \) by making \( n_0 \) as big as possible and \( n_1 \) as small as possible.
  – In the “most likely” set of sample with \( \theta = 0.12 \), nobody has COVID-19!

• The **decoding often does not represent “typical” behavior**.
  – For example, if \( \theta = 0.12 \) 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 \( \theta \), generate samples of ‘x’ distributed according to \( p(x \mid \theta) \).
  – This is called *sampling* from the distribution.

• I think of sampling as the “opposite” of density estimation:

  - Often write code to generate one IID sample, then call it many times.

<table>
<thead>
<tr>
<th>COVID-19?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

  \[ p(x = 1) = 0.4 \]
Digression: Motivation for Sampling

• Sampling is not very interesting for Bernoulli distributions.
  – Because knowing $\theta$ tells you everything about the distribution.

• But sampling will let us do neat things in more-complicated density models:
  – “This person does not exist”.
  – Sampling often gives indications about whether model is reasonable.
    • If samples look nothing like the data, then model is not very good.
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.
    • Like Julia’s `rand` function (we won’t discuss how these work, Google it if you want to sleep).

• Consider sampling from a Bernoulli with $\theta = 0.9$.
  – 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 $\theta$ value:
  – Generate a sample uniformly on the interval between 0 and 1.
  – If the sample is less than $\theta$, return 1.
    • Otherwise, return 0.

• In code:
  Nice version: $u = \text{rand}(1)$
  if $u \leq \theta$
    $x = 1$
  else
    $x = 0$

  Slick but less interpretable: $x = \text{rand}(1) \leq \theta$

• 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 $\theta$:

```
   COVID-19?   X =
               1
               0
               0
               0
               1
               0
```

  $\theta = 0.4$

• 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:
    \[ p(X \mid \theta) = \theta^1(1-\theta)^0 \]
    • The probability of seeing the data ‘X’ if our Bernoulli parameter is \( \theta \).
  • Here is a plot of the likelihood if our IID data is \( x^1=1, \ x^2=1, \ x^3=0 \).

  • The likelihood of \( p(1, 1, 0 \mid \theta = 0.5) = (1/2)(1/2)(1/2) = 0.125 \).
  • If \( \theta = 0.75 \), then \( p(1, 1, 0 \mid \theta = 0.75) = (3/4)(3/4)(1/4) \approx 0.14 \) (dataset is more likely for \( \theta = 0.75 \) than 0.5).
  • If \( \theta = 0 \) (“always 0”), then \( p(1, 1, 0 \mid \theta = 0) = 0 \) (dataset is not possible for \( \theta = 0 \)).
    • Data has probability 0 if \( \theta=0 \) or \( \theta=1 \) (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 ‘\( \theta \)’ (area under the curve is not 1).
Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation (MLE):
  - Choose the parameters that have the highest likelihood, \( p(X \mid \theta) \).
  - “Find the parameter(s) \( \theta \) 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 \( \theta = 2/3 \).

- The MLE for general Bernoulli is \( \theta = \frac{n_1}{n_1 + n_0} \).
  - “If you flip a coin 50 times and it lands heads 23 times, your guess for 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:

\[ \hat{\theta} \in \text{argmax}_\theta \{ \theta^n (1-\theta)^{n_0} \} \]

Our estimate of \( \theta \) based on data

"argmax" means "find the values that achieve the maximum"

Likelihood for data with counts \( n \) and \( n_0 \)

"argmax" returns a set containing all the values \( \theta \) that give maximum value

There be more than one element in argmax. We say you “pick one in reset”
Digression: Maximizing the Log-Likelihood

• Instead of finding an element maximizing the likelihood:

\[ \hat{\theta} \in \arg \max_{\theta} \{ p(X \mid \theta) \} \]

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

\[ \hat{\theta} \in \arg \max_{\theta} \{ \log(p(X \mid \theta)) \} \]

- 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:

\[ \hat{\theta} \in \arg\max_{\theta} \left\{ \theta^n (1-\theta)^n \right\} \]

• MLE for Bernoulli by maximizing the log-likelihood:

\[ \hat{\theta} \in \arg\max_{\theta} \left\{ \log \left( \theta^n (1-\theta)^n \right) \right\} \]

\[ \equiv \arg\max_{\theta} \left\{ \sum \log(\theta) + \sum \log(1-\theta) \right\} \]

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

using \( \log(\alpha^t) = t \log(\alpha) \)
Derivation MLE for Bernoulli

• From the last slide we want to find:

\[ \hat{\theta} \in \arg\max_{\theta} \left\{ n_1 \log(\theta) + n_0 \log(1 - \theta) \right\} \]

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

\[ 0 = \frac{n_1}{\theta} - \frac{n_0}{1 - \theta} \]

  \[
  \text{derivative of } n_1 \log \theta \text{ for } \theta > 0 \quad \text{derivative of } n_0 \log(1 - \theta) \text{ for } \theta > 0
  \]

  – 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 $p(x = 1)$ given IID samples $x^1, x^2, \ldots, x^n$.

• Bernoulli distribution:
  – Probability distribution over a binary variable.
  – Parameterized by a number $\theta$ such that $p(x=1 \mid \theta) = \theta$.

• 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 $\theta = (\text{number of 1s})/(\text{number of examples})$.

• Next time: more boring definitions.