MLE/MAP + Naïve Bayes
Reminders

• Homework 5: Neural Networks
  – Out: Sun, Feb 27
  – Due: Fri, Mar 18 at 11:59pm
• Homework 6: Learning Theory / Generative Models
  – Out: Fri, Mar. 18
  – Due: Fri, Mar. 25 at 11:59pm
  – IMPORTANT: only 2 grace/late days permitted
• Exam 2 (Thu, Mar 3rd)
• Exam 3 (Tue, May 3rd)
PROBABILITY LEARNING
Previously, we assumed that our output was generated using a deterministic target function:

\[ x^{(i)} \sim p^*(\cdot) \]
\[ y^{(i)} = c^*(x^{(i)}) \]

Our goal was to learn a hypothesis \( h(x) \) that best approximates \( c^*(x) \).

Today, we assume that our output is sampled from a conditional probability distribution:

\[ x^{(i)} \sim p^*(\cdot) \]
\[ y^{(i)} \sim p^*(\cdot | x^{(i)}) \]

Our goal is to learn a probability distribution \( p(y|x) \) that best approximates \( p^*(y|x) \).
MAXIMUM LIKELIHOOD ESTIMATION (MLE)
Likelihood Function

• Given N independent, identically distributed (iid) samples \( D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \) from a random variable \( X \) ...

• The **likelihood** function is
  - **Case 1**: \( X \) is **discrete** with probability mass function (pmf) \( p(x|\theta) \)
    \[
    L(\theta) = p(x^{(1)}|\theta) \cdot p(x^{(2)}|\theta) \cdots p(x^{(N)}|\theta)
    \]
  - **Case 2**: \( X \) is **continuous** with probability density function (pdf) \( f(x|\theta) \)
    \[
    L(\theta) = f(x^{(1)}|\theta) \cdot f(x^{(2)}|\theta) \cdots f(x^{(N)}|\theta)
    \]

• The **log-likelihood** function is
  - **Case 1**: \( X \) is **discrete** with probability mass function (pmf) \( p(x|\theta) \)
    \[
    \ell(\theta) = \log p(x^{(1)}|\theta) + \cdots + \log p(x^{(N)}|\theta)
    \]
  - **Case 2**: \( X \) is **continuous** with probability density function (pdf) \( f(x|\theta) \)
    \[
    \ell(\theta) = \log f(x^{(1)}|\theta) + \cdots + \log f(x^{(N)}|\theta)
    \]
Likelihood Function

• Given N iid samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of random variables $X, Y$

• The conditional likelihood function:
  – Case 1: $Y$ is discrete with pmf $p(y \mid x, \theta)$
    \[
    L(\theta) = p(y^{(1)} \mid x^{(1)}, \theta) \ldots p(y^{(N)} \mid x^{(N)}, \theta)
    \]
  – Case 2: $Y$ is continuous with pdf $f(y \mid x, \theta)$
    \[
    L(\theta) = f(y^{(1)} \mid x^{(1)}, \theta) \ldots f(y^{(N)} \mid x^{(N)}, \theta)
    \]

• The joint likelihood function:
  – Case 1: $X$ and $Y$ are discrete with pmf $p(x,y\mid\theta)$
    \[
    L(\theta) = p(x^{(1)}, y^{(1)}\mid\theta) \ldots p(x^{(N)}, y^{(N)}\mid\theta)
    \]
  – Case 2: $X$ and $Y$ are continuous with pdf $f(x,y\mid\theta)$
    \[
    L(\theta) = f(x^{(1)}, y^{(1)}\mid\theta) \ldots f(x^{(N)}, y^{(N)}\mid\theta)
    \]
Likelihood Function

- Given $N$ iid samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of random variables $X, Y$

- The joint likelihood function:
  - Case 1: $X$ and $Y$ are **discrete** with pmf $p(x, y|\theta)$  
    \[ L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta) \]
  - Case 2: $X$ and $Y$ are **continuous** with pdf $f(x, y|\theta)$  
    \[ L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta) \]
  - Case 3: $Y$ is **discrete** with pmf $p(y|\beta)$ and $X$ is **continuous** with pdf $f(x|y, \alpha)$  
    \[ L(\alpha, \beta) = f(x^{(1)}| y^{(1)}, \alpha) p(y^{(1)}|\beta) \ldots f(x^{(N)}| y^{(N)}, \alpha) p(y^{(N)}|\beta) \]
  - Case 4: $Y$ is **continuous** with pdf $f(y|\beta)$ and $X$ is **discrete** with pmf $p(x|y, \alpha)$  
    \[ L(\alpha, \beta) = p(x^{(1)}| y^{(1)}, \alpha) f(y^{(1)}|\beta) \ldots p(x^{(N)}| y^{(N)}, \alpha) f(y^{(N)}|\beta) \]
Suppose we have data \( \mathcal{D} = \{x^{(i)}\}_{i=1}^N \)

**Principle of Maximum Likelihood Estimation:**
Choose the parameters that maximize the likelihood of the data.

\[
\theta^{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^N p(x^{(i)}|\theta)
\]

Maximum Likelihood Estimate (MLE)
MLE

What does maximizing likelihood accomplish?

• There is only a finite amount of probability mass (i.e. sum-to-one constraint)

• MLE tries to allocate **as much** probability mass **as possible** to the things we have observed...

... **at the expense** of the things we have **not** observed
Recipe for Closed-form MLE

1. Assume data was generated iid from some model, i.e., write the generative story
   \[ x^{(i)} \sim p(x|\theta) \]

2. Write the log-likelihood
   \[ \ell(\theta) = \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta) \]

3. Compute partial derivatives, i.e., the gradient
   \[ \frac{\partial \ell(\theta)}{\partial \theta_1} = \ldots \]
   \[ \ldots \]
   \[ \frac{\partial \ell(\theta)}{\partial \theta_M} = \ldots \]

4. Set derivatives equal to zero and solve for \( \theta \)
   \[ \frac{\partial \ell(\theta)}{\partial \theta_m} = 0 \text{ for all } m \in \{1, \ldots, M\} \]
   \[ \theta^{\text{MLE}} = \text{solution to system of } M \text{ equations and } M \text{ variables} \]

5. Compute the second derivative and check that \( \ell(\theta) \) is concave down at \( \theta^{\text{MLE}} \)
What we earlier called “Closed Form Solution for Linear Regression”

EXAMPLE:
MLE FOR LINEAR REGRESSION
Linear Regression as Function Approximation

1. Assume $\mathcal{D}$ generated as:
   
   \[ x^{(i)} \sim p^*(\cdot) \]
   \[ y^{(i)} = h^*(x^{(i)}) \]

2. Choose hypothesis space, $\mathcal{H}$:
   all linear functions in $M$-dimensional space
   
   \[ \mathcal{H} = \{ h_\theta : h_\theta(x) = \theta^T x, \theta \in \mathbb{R}^M \} \]

3. Choose an objective function:
   mean squared error (MSE)
   
   \[ J(\theta) = \frac{1}{N} \sum_{i=1}^{N} e_i^2 \]
   \[ = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - h_\theta(x^{(i)}))^2 \]
   \[ = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \theta^T x^{(i)})^2 \]

4. Solve the unconstrained optimization problem via favorite method:
   - gradient descent
   - closed form
   - stochastic gradient descent
   - ...
   
   \[ \hat{\theta} = \text{argmin}_{\theta} J(\theta) \]

5. Test time: given a new $x$, make prediction $\hat{y}$
   
   \[ \hat{y} = h_{\hat{\theta}}(x) = \hat{\theta}^T x \]
Optimization Method #2: Closed Form

1. Evaluate
   \[ \theta^{\text{MLE}} = (X^T X)^{-1} X^T y \]

2. Return \( \theta^{\text{MLE}} \)

\[ J(\theta) = J(\theta_1, \theta_2) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \theta^T x^{(i)})^2 \]

### Table:

<table>
<thead>
<tr>
<th>t</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( J(\theta_1, \theta_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.59</td>
<td>0.43</td>
<td>0.2</td>
</tr>
</tbody>
</table>

### Graph:

- \( y = h^*(x) \) (unknown)
- \( h(x; \theta^{\text{MLE}}) \)
MLE for Linear Regression

You’ll work through the view of linear regression as a probabilistic model in the homework!
MLE EXAMPLES
MLE of Exponential Distribution

Goal:
- pdf of Exponential(\lambda): f(x) = \lambda e^{-\lambda x}
- Suppose \( X_i \sim \text{Exponential}(\lambda) \) for \( 1 \leq i \leq N \).
- Find MLE for data \( \mathcal{D} = \{x^{(i)}\}_{i=1}^{N} \)

Steps:
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for \( \lambda \).
- Compute second derivative and check that it is concave down at \( \lambda^{\text{MLE}} \).
MLE of Exponential Distribution

- pdf of Exponential($\lambda$): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim \text{Exponential} (\lambda)$ for $1 \leq i \leq N$.
- Find MLE for data $D = \{x^{(i)}\}_{i=1}^N$

- First write down log-likelihood of sample.

\[
\ell(\lambda) = \sum_{i=1}^{N} \log f(x^{(i)})
\]

(1)

\[
= \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)}))
\]

(2)

\[
= \sum_{i=1}^{N} \log(\lambda) + -\lambda x^{(i)}
\]

(3)

\[
= N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)}
\]

(4)
MLE of Exponential Distribution

- pdf of Exponential($\lambda$): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim \text{Exponential}(\lambda)$ for $1 \leq i \leq N$.
- Find MLE for data $D = \{x^{(i)}\}_{i=1}^N$

- Compute first derivative, set to zero, solve for $\lambda$.

\[
\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^N x^{(i)} \tag{1}
\]

\[
= \frac{N}{\lambda} - \sum_{i=1}^N x^{(i)} = 0 \tag{2}
\]

\[
\Rightarrow \lambda^{\text{MLE}} = \frac{N}{\sum_{i=1}^N x^{(i)}} \tag{3}
\]
MLE of Bernoulli

In-Class Exercise
Show that the MLE of parameter $\phi$ for $N$ samples drawn from Bernoulli($\phi$) is:

$$\phi_{MLE} = \frac{\text{Number of } x_i = 1}{N}$$

Steps to answer:
1. Write log-likelihood of sample
2. Compute derivative w.r.t. $\phi$
3. Set derivative to zero and solve for $\phi$
MLE of Bernoulli

**Question:**
Assume we have $N$ iid samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ drawn from a Bernoulli($\phi$).

**Step 1:** What is the log-likelihood of the data $\ell(\phi)$?

Assume $N_1 = \# \text{ of } (x^{(i)} = 1)$

$N_0 = \# \text{ of } (x^{(i)} = 0)$

**Answer:**
A. $l(\phi) = N_1 \log(\phi) + N_0 (1 - \log(\phi))$
B. $l(\phi) = N_1 \log(\phi) + N_0 \log(1-\phi)$
C. $l(\phi) = \log(\phi)^{N_1} + (1 - \log(\phi))^{N_0}$
D. $l(\phi) = \log(\phi)^{N_1} + \log(1-\phi)^{N_0}$
E. $l(\phi) = N_0 \log(\phi) + N_1 (1 - \log(\phi))$
F. $l(\phi) = N_0 \log(\phi) + N_1 \log(1-\phi)$
G. $l(\phi) = \log(\phi)^{N_0} + (1 - \log(\phi))^{N_1}$
H. $l(\phi) = \log(\phi)^{N_0} + \log(1-\phi)^{N_1}$
I. $l(\phi) = N_0 + N_1$
MLE of Bernoulli

Question:
Assume we have N iid samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ drawn from a Bernoulli($\phi$).

Step 2: What is the derivative of the log-likelihood $\partial l(\theta)/\partial \theta$?

Assume $N_1 =$ # of $(x^{(i)} = 1)$
$N_0 =$ # of $(x^{(i)} = 0)$

Answer:
A. $\partial l(\theta)/\partial \theta = \phi^{N_1} - (1 - \phi)^{N_0}$
B. $\partial l(\theta)/\partial \theta = \phi / N_1 - (1 - \phi) / N_0$
C. $\partial l(\theta)/\partial \theta = N_1 / \phi - N_0 / (1 - \phi)$
D. $\partial l(\theta)/\partial \theta = \log(\phi) / N_1 - \log(1 - \phi) / N_0$
E. $\partial l(\theta)/\partial \theta = N_1 / \log(\phi) - N_0 / \log(1 - \phi)$
F. $\partial l(\theta)/\partial \theta = 0$
MLE of Bernoulli

Whiteboard

– Example: MLE of Bernoulli
MAP ESTIMATION
MLE vs. MAP

Suppose we have data \( \mathcal{D} = \{x^{(i)}\}_{i=1}^{N} \)

**Principle of Maximum Likelihood Estimation:** Choose the parameters that maximize the likelihood of the data.

\[
\theta^{\text{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(x^{(i)}|\theta)
\]

Maximum Likelihood Estimate (MLE)

**Principle of Maximum a posteriori (MAP) Estimation:** Choose the parameters that maximize the posterior of the parameters given the data.

\[
\theta^{\text{MLE}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} f(\theta) \prod_{i=1}^{N} p(x^{(i)}|\theta)
\]

Maximum a posteriori (MAP) estimate
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$.

**Principle of Maximum Likelihood Estimation (MLE):**
Choose the parameters that maximize the likelihood of the data.

$$\theta^{\text{MLE}} = \arg\max_\theta p(\mathcal{D}|\theta)$$

**Maximum Likelihood Estimate (MLE)**

**Principle of Maximum a posteriori (MAP) Estimation:**
Choose the parameters that maximize the posterior of the parameters given the data.

$$\theta^{\text{MAP}} = \arg\max_\theta p(\theta|\mathcal{D}) = \arg\max_\theta f(\theta) \prod_{i=1}^N p(x^{(i)}|\theta)$$

**Maximum a posteriori (MAP) estimate**

**Important!**
Usually the parameters are **continuous**, so the prior is a probability **density** function.

"Important!" box with the following text:
"Usually the parameters are **continuous**, so the prior is a probability **density** function."
Learning from Data (Bayesian)

Whiteboard
  – *maximum a posteriori* (MAP) estimation
Recipe for Closed-form MLE

1. Assume data was generated iid from some model, i.e., write the generative story
   \[ x^{(i)} \sim p(x|\theta) \]

2. Write the log-likelihood
   \[ \ell(\theta) = \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta) \]

3. Compute partial derivatives, i.e., the gradient
   \[ \frac{\partial \ell(\theta)}{\partial \theta_1} = \ldots \]
   \[ \ldots \]
   \[ \frac{\partial \ell(\theta)}{\partial \theta_M} = \ldots \]

4. Set derivatives equal to zero and solve for \( \theta \)
   \[ \frac{\partial \ell(\theta)}{\partial \theta_m} = 0 \text{ for all } m \in \{1, \ldots, M\} \]
   \[ \theta^{\text{MLE}} = \text{solution to system of } M \text{ equations and } M \text{ variables} \]

5. Compute the second derivative and check that \( \ell(\theta) \) is concave down at \( \theta^{\text{MLE}} \)
Recipe for Closed-form MAP

1. Assume data was generated iid from some model, i.e., write the generative story
   \( \theta \sim p(\theta) \) and then for all \( i \): \( x^{(i)} \sim p(x|\theta) \)

2. Write the log posterior
   \( \ell_{\text{MAP}}(\theta) = \log p(\theta) + \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta) \)

3. Compute partial derivatives, i.e., the gradient
   \[ \frac{\partial \ell_{\text{MAP}}(\theta)}{\partial \theta_1} = \ldots \]
   \[ \ldots \]
   \[ \frac{\partial \ell_{\text{MAP}}(\theta)}{\partial \theta_M} = \ldots \]

4. Set derivatives to equal zero and solve for \( \theta \)
   \[ \frac{\partial \ell_{\text{MAP}}(\theta)}{\partial \theta_m} = 0 \text{ for all } m \in \{1, \ldots, M\} \]
   \( \theta^{\text{MAP}} = \text{solution to system of } M \text{ equations and } M \text{ variables} \)

5. Compute the second derivative and check that \( \ell(\theta) \) is concave down at \( \theta^{\text{MAP}} \)
MAP of Beta-Bernoulli Model

Whiteboard

– Example: MAP of Beta-Bernoulli Model
Takeaways

• One view of what ML is trying to accomplish is function approximation
• The principle of maximum likelihood estimation provides an alternate view of learning

• Synthetic data can help debug ML algorithms
• Probability distributions can be used to model real data that occurs in the world
  (don’t worry we’ll make our distributions more interesting soon!)
Learning Objectives

**MLE / MAP**

*You should be able to…*

1. Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence

2. Describe common probability distributions such as the Beta, Dirichlet, Multinomial, Categorical, Gaussian, Exponential, etc.

3. State the principle of maximum likelihood estimation and explain what it tries to accomplish

4. State the principle of maximum a posteriori estimation and explain why we use it

5. Derive the MLE or MAP parameters of a simple model in closed form
NAÏVE BAYES
Naïve Bayes

• Why are we talking about Naïve Bayes?
  – It’s **just another decision function** that fits into our “big picture” recipe from last time
  – But it’s our first **example of a Bayesian Network** and provides a **clearer** picture of **probabilistic learning**
  – Just like the other Bayes Nets we’ll see, it **admits a closed form solution** for MLE and MAP
  – So learning is **extremely efficient** (just counting)
Today’s Goal: To define a generative model of emails of two different classes (e.g. real vs. fake news)

The Economist

Soybean Prices Surge as South American Outlook Deteriorates
Drought is pushing prices up, with shortfalls in production expected to boost demand for U.S. beans

By Kirk Maltais
Feb. 12, 2022 7:00 am ET

U.S. soybean prices have surged in recent months amid shrinking forecasts for South American crops.

Prices for soybeans—the base ingredient in many food products, poultry and livestock feed and renewable fuel, among other things—are edging back toward highs reached last year, which hadn’t previously been seen in a decade.

The Onion

Watchdog Warns Nearly Every Food Brand In U.S. Owned By Handful Of Companies, Which In Turn Are Controlled By Newman’s Own

WASHINGTON—Calling for a full-scale Federal Trade Commission investigation into the sauce and salad dressing brand, the American Antitrust Institute issued a report Thursday warning that nearly every food brand in the United States was owned by a handful of companies, which in turn were controlled by Newman’s Own. “Kellogg’s, General Mills, PepsiCo, Kraft Heinz—all of these companies are just Newman’s Own by another name,” said Diana L.
Fake News Detector

We can pretend the natural process generating these vectors is stochastic...
Naive Bayes: Model

Whiteboard

– Generating synthetic "labeled documents"
– Definition of model
– Naive Bayes assumption
– Counting # of parameters with / without NB assumption
Model 1: Bernoulli Naïve Bayes

If HEADS, flip each red coin

If TAILS, flip each blue coin

We can generate data in this fashion. Though in practice we never would since our data is given.

Instead, this provides an explanation of how the data was generated (albeit a terrible one).

Each red coin corresponds to an $x_m$
What’s wrong with the Naïve Bayes Assumption?

The features might not be independent!!

• Example 1:
  – If a document contains the word “Donald”, it’s extremely likely to contain the word “Trump”
  – These are not independent!

• Example 2:
  – If the petal width is very high, the petal length is also likely to be very high
Naïve Bayes: Learning from Data

Whiteboard
– Data likelihood
– MLE for Naive Bayes
– Example: MLE for Naïve Bayes with Two Features
– MAP for Naive Bayes
Recipe for Closed-form MLE

1. Assume data was generated iid from some model, i.e., write the generative story
   \( x^{(i)} \sim p(x|\theta) \)

2. Write the log-likelihood
   \( \ell(\theta) = \log p(x^{(1)}|\theta) + \ldots + \log p(x^{(N)}|\theta) \)

3. Compute partial derivatives, i.e., the gradient
   \( \frac{\partial \ell(\theta)}{\partial \theta_1} = \ldots \)
   \( \ldots \)
   \( \frac{\partial \ell(\theta)}{\partial \theta_M} = \ldots \)

4. Set derivatives equal to zero and solve for \( \theta \)
   \( \frac{\partial \ell(\theta)}{\partial \theta_m} = 0 \) for all \( m \in \{1, \ldots, M\} \)
   \( \theta^{\text{MLE}} = \) solution to system of \( M \) equations and \( M \) variables

5. Compute the second derivative and check that \( \ell(\theta) \) is concave down at \( \theta^{\text{MLE}} \)