23: Naïve Bayes

Jerry Cain
May 19, 2021
Intro: Machine Learning
Our path from here
Our path from here

Deep Learning
- Linear Regression
- Naïve Bayes
- Logistic Regression

Unbiased estimators
- $\bar{X}, S^2$
- $\theta_{MLE}$
- $\theta_{MAP}$
Machine Learning (formally)

Many different forms of “Machine Learning”

• We focus on the problem of prediction based on observations.
Machine Learning uses a lot of data.

Task: Identify what a chair is
Data: All the chairs ever

**Supervised learning**: A category of machine learning where you have labeled data for the problem you are solving.
Supervised learning

Real World Problem

Model the problem

Formal Model $\theta$

Learning Algorithm

Testing Data

Prediction Function $\hat{\theta}$

Evaluation score

Training Data
Supervised learning

- Real World Problem
  - Model the problem
    - Formal Model $\theta$
  - Learning Algorithm
    - Prediction Function $\hat{\theta}$
    - Evaluation score
      - Training Data
      - Testing Data
Supervised learning

Real World Problem

Model the problem

Formal Model $\theta$

Training Data

Learning Algorithm

Prediction Function $\hat{\theta}$

Testing Data

Evaluation score
Model and dataset

Many different forms of machine learning
• We focus on the problem of prediction based on observations.

Goal
Based on observed $X$, predict unseen $Y$

• Features
Vector $X$ of $m$ observed variables
$X = (X_1, X_2, \ldots, X_m)$

• Output
Variable $Y$ (also called class label if discrete)

Model
$\hat{Y} = g(X)$, a function of observations $X$
## Training data

\[ X = (X_1, X_2, X_3, ..., X_{300}) \]

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 300</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Patient 2</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Patient n</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Training data notation

\[(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\]

$n$ datapoints, generated i.i.d.

$i$-th datapoint \((x^{(i)}, y^{(i)})\):

- $m$ features: \(x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})\)
- A single output \(y^{(i)}\)
- Independent of all other datapoints

Training Goal: Use these $n$ datapoints to learn a model \(\hat{Y} = g(X)\) that predicts $Y$
Supervised learning

- Real World Problem
  - Model the problem
    - Formal Model $\theta$
      - Learning Algorithm
        - Training Data
          - Testing Data
            - Prediction Function $\hat{\theta}$
              - Evaluation score
Testing data notation

\[ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)}) \]

\[ n_{test} \] other datapoints, generated i.i.d.

i-th datapoint \((x^{(i)}, y^{(i)})\):
- Has the same structure as your training data

Testing Goal: Using the model \( \hat{Y} = g(X) \) that you trained, see how well you can predict \( Y \) on known data
Two tasks we will focus on

Many different forms of “Machine Learning”

• We focus on the problem of prediction based on observations.

Goal
Based on observed $X$, predict unseen $Y$

• Features
  Vector $X$ of $m$ observed variables
  $X = (X_1, X_2, ..., X_m)$

• Output
  Variable $Y$ (also called class label if discrete)

Model
$\hat{Y} = g(X)$, a function of observations $X$

• Regression
  prediction when $Y$ is continuous

• Classification
  prediction when $Y$ is discrete
Regression: Predicting real numbers

Training data: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\)

- **CO2 levels**
  - Year 1: 338.8
  - Year 2: 340.0
  - \ldots
  - Year \(n\): 340.76

- **Sea level**
  - Year 1: 0
  - Year 2: 1
  - \ldots
  - Year \(n\): 0

- **Feature \(m\)**
  - Year 1: 1
  - Year 2: 0
  - \ldots
  - Year \(n\): 1

- **Output**
  - Year 1: 0.26
  - Year 2: 0.32
  - \ldots
  - Year \(n\): 0.14
Classification: Predicting class labels

\[ \mathbf{X} = (X_1, X_2, X_3, \ldots, X_{300}) \]

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 300</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>1</td>
<td>0</td>
<td>... 1</td>
</tr>
<tr>
<td>Patient 2</td>
<td>1</td>
<td>1</td>
<td>... 0</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Patient n</td>
<td>0</td>
<td>0</td>
<td>... 1</td>
</tr>
</tbody>
</table>

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
"Brute Force Bayes"
Classification: Having a healthy heart

$X = (X_1)$

Patient 1  1  0
Patient 2  1  1
Patient $n$  0  1

Single feature: Region of Interest (ROI) is healthy (1) or unhealthy (0)

How can we predict the class label heart is healthy (1) or unhealthy (0)?

The following strategy is not used in practice but helps us understand how to approach classification.
Classification: “Brute Force Bayes”

\[ \hat{Y} = g(X) \]

Our prediction for \( Y \) is a function of \( X \)

\[ = \arg \max_{y={0,1}} P(Y | X) \]

Proposed model: Choose the \( Y \) that is most likely given \( X \)

\[ = \arg \max_{y={0,1}} \frac{P(X|Y)P(Y)}{P(X)} \]

(Bayes’ Theorem)

\[ = \arg \max_{y={0,1}} P(X|Y)P(Y) \]

\[ (1/P(X) \text{ is constant w.r.t. } y) \]

If we estimate \( P(X|Y) \) and \( P(Y) \), we can classify datapoints!
Training: Estimate parameters

\[ X = (X_1) \]

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(X|Y) \hat{P}(Y) \]

| Feature 1 | Output | Conditional probability tables \( \hat{P}(X|Y) \) |
|-----------|--------|----------------------------------|
| Patient 1 | 1      | \( \hat{P}(X|Y = 0) \) | \( \hat{P}(X|Y = 1) \) |
| Patient 2 | 1      | \( \theta_1 \) | \( \theta_3 \) |
| Patient n | 0      | \( \theta_2 \) | \( \theta_4 \) |

<table>
<thead>
<tr>
<th>Marginal probability table ( \hat{P}(Y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
</tr>
<tr>
<td>( Y = 1 )</td>
</tr>
</tbody>
</table>

Training Goal: Use \( n \) datapoints to learn \( 2 \cdot 2 + 2 = 6 \) parameters.
Training: Estimate parameters $\hat{P}(X|Y)$

| $X_1$ | $\hat{P}(X|Y = 0)$ | $\hat{P}(X|Y = 1)$ |
|-------|---------------------|---------------------|
| 0     | $\theta_1$         | $\theta_3$         |
| 1     | $\theta_2$         | $\theta_4$         |

$X|Y = 0$ and $X|Y = 1$ are each multinomials with 2 outcomes!

Use MLE or Laplace (MAP) estimate for parameters $\hat{P}(X|Y)$ and $\hat{P}(Y)$
## Training: MLE estimates, $\hat{P}(X|Y)$

| $X_1$ | $\hat{P}(X|Y = 0)$ | $\hat{P}(X|Y = 1)$ |
|-------|---------------------|---------------------|
| 0     | 0.4                 | 0.0                 |
| 1     | 0.6                 | 1.0                 |

MLE of $\hat{P}(X_1 = x|Y = y) = \frac{\#(X_1 = x, Y = y)}{\#(Y = y)}$

Just count!

<table>
<thead>
<tr>
<th>Patient $n$</th>
<th>$X_1 = 0, Y = 0$:</th>
<th>$X_1 = 1, Y = 0$:</th>
<th>$X_1 = 0, Y = 1$:</th>
<th>$X_1 = 1, Y = 1$:</th>
<th>Total:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>100</td>
<td>110</td>
</tr>
</tbody>
</table>

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Training: Laplace (MAP) estimates, $\hat{P}(X|Y)$

<table>
<thead>
<tr>
<th>Count:</th>
<th># datapoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 = 0, Y = 0$</td>
<td>4</td>
</tr>
<tr>
<td>$X_1 = 1, Y = 0$</td>
<td>6</td>
</tr>
<tr>
<td>$X_1 = 0, Y = 1$</td>
<td>0</td>
</tr>
<tr>
<td>$X_1 = 1, Y = 1$</td>
<td>100</td>
</tr>
<tr>
<td>Total:</td>
<td>110</td>
</tr>
</tbody>
</table>

| $X_1$ | $\hat{P}(X|Y = 0)$ | $\hat{P}(X|Y = 1)$ |
|-------|---------------------|---------------------|
| 0     | 0.4                 | 0.0                 |
| 1     | 0.6                 | 1.0                 |

MLE of $\hat{P}(X_1 = x|Y = y) = \frac{\#(X_1 = x, Y = y)}{\#(Y = y)}$

Just count!

Laplace of $\hat{P}(X_1 = x|Y = y) = ?$

Just count + add imaginary trials!
Training: Laplace (MAP) estimates, $\hat{P}(X|Y)$

|   | $\hat{P}(X|Y = 0)$ | $\hat{P}(X|Y = 1)$ |
|---|-----------------|------------------|
| $X_1 = 0$ | 0.4             | 0.0              |
| $X_1 = 1$ | 0.6             | 1.0              |

MLE of $\hat{P}(X_1 = x|Y = y) = \frac{\#(X_1 = x,Y = y)}{\#(Y = y)}$

Just count!

|   | $\hat{P}(X|Y = 0)$ | $\hat{P}(X|Y = 1)$ |
|---|-----------------|------------------|
| $X_1 = 0$ | 0.42            | 0.01             |
| $X_1 = 1$ | 0.58            | 0.99             |

Laplace of $\hat{P}(X_1 = x|Y = y) = \frac{\#(X_1 = x,Y = y) + 1}{\#(Y = y) + 2}$

Just count + add imaginary trials!
Testing

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(X|Y)\hat{P}(Y) \]

| (MAP)   | \( \hat{P}(X|Y = 0) \) | \( \hat{P}(X|Y = 1) \) | (MLE) | \( \hat{P}(Y) \) |
|---------|------------------------|------------------------|--------|----------------|
| \( X_1 = 0 \) | 0.42                   | 0.01                   | \( Y = 0 \) | 0.09           |
| \( X_1 = 1 \) | 0.58                   | 0.99                   | \( Y = 1 \) | 0.91           |

New patient has a healthy ROI (\( X_1 = 1 \)). What is your prediction, \( \hat{Y} \)?

\[ \hat{P}(X_1 = 1|Y = 0)\hat{P}(Y = 0) = 0.58 \cdot 0.09 \approx 0.052 \]
\[ \hat{P}(X_1 = 1|Y = 1)\hat{P}(Y = 1) = 0.99 \cdot 0.91 \approx 0.901 \]

A. 0.052 < 0.5  \( \Rightarrow \hat{Y} = 1 \)
B. 0.901 > 0.5  \( \Rightarrow \hat{Y} = 1 \)
C. 0.052 < 0.901  \( \Rightarrow \hat{Y} = 1 \)

Sanity check: Why don’t these sum to 1?
Testing

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(X|Y)\hat{P}(Y) \]

| (MAP) | \( \hat{P}(X|Y = 0) \) | \( \hat{P}(X|Y = 1) \) | (MLE) | \( \hat{P}(Y) \) |
|-------|----------------|----------------|-------|---------------|
| \( X_1 = 0 \) | 0.42 | 0.01 | \( Y = 0 \) | 0.09 |
| \( X_1 = 1 \) | 0.58 | 0.99 | \( Y = 1 \) | 0.91 |

New patient has a healthy ROI (\( X_1 = 1 \)). What is your prediction, \( \hat{Y} \)?

\[
\begin{align*}
\hat{P}(X_1 = 1|Y = 0)\hat{P}(Y = 0) &= 0.58 \cdot 0.09 \approx 0.052 \\
\hat{P}(X_1 = 1|Y = 1)\hat{P}(Y = 1) &= 0.99 \cdot 0.91 \approx 0.901
\end{align*}
\]

A. 0.052 < 0.5 \( \Rightarrow \) \( \hat{Y} = 1 \)
B. 0.901 > 0.5 \( \Rightarrow \) \( \hat{Y} = 1 \)
C. 0.052 < 0.901 \( \Rightarrow \) \( \hat{Y} = 1 \)

Sanity check: Why don’t these sum to 1?
“Brute Force Bayes” classifier

\[
\hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(X|Y)\hat{P}(Y)
\]

(\(\hat{P}(Y)\) is an estimate of \(P(Y)\), 
\(\hat{P}(X|Y)\) is an estimate of \(P(X|Y)\))

**Training**

Estimate these probabilities, i.e., “learn” these parameters using MLE or Laplace (MAP)

\[
\hat{P}(X_1, X_2, ..., X_m|Y = 1) \\
\hat{P}(X_1, X_2, ..., X_m|Y = 0) \\
\hat{P}(Y = 1) \\
\hat{P}(Y = 0)
\]

**Testing**

Given an observation \(X = (X_1, X_2, ..., X_m)\), predict

\[
\hat{Y} = \arg \max_{y=\{0,1\}} \left( \hat{P}(X_1, X_2, ..., X_m|Y)\hat{P}(Y) \right)
\]
Naïve Bayes Classifier
Brute Force Bayes: \( m = 300 \) (# features)

\[
X = (X_1, X_2, X_3, ..., X_{300})
\]

This won’t be too bad, right?
Brute Force Bayes: \( m = 300 \) (# features)

\[
X = (X_1, X_2, X_3, \ldots, X_{300})
\]

<table>
<thead>
<tr>
<th>Count:</th>
<th># datapoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 = 0, X_2 = 0, \ldots, X_{299} = 0, X_{300} = 0, Y = 0: )</td>
<td>0</td>
</tr>
<tr>
<td>( X_1 = 0, X_2 = 0, \ldots, X_{299} = 0, X_{300} = 1, Y = 0: )</td>
<td>0</td>
</tr>
<tr>
<td>( X_1 = 0, X_2 = 0, \ldots, X_{299} = 1, X_{300} = 0, Y = 0: )</td>
<td>1</td>
</tr>
<tr>
<td>( X_1 = 0, X_2 = 0, \ldots, X_{299} = 0, X_{300} = 0, Y = 1: )</td>
<td>2</td>
</tr>
<tr>
<td>( X_1 = 0, X_2 = 0, \ldots, X_{299} = 0, X_{300} = 1, Y = 1: )</td>
<td>1</td>
</tr>
<tr>
<td>( X_1 = 0, X_2 = 0, \ldots, X_{299} = 1, X_{300} = 0, Y = 1: )</td>
<td>1</td>
</tr>
</tbody>
</table>

This won’t be too bad, right?
Brute Force Bayes: \( m = 300 \) (# features)

\[
\hat{Y} = \arg \max_{y \in \{0,1\}} \hat{P}(Y \mid X)
\]

\[
= \arg \max_{y \in \{0,1\}} \frac{\hat{P}(X \mid Y) \hat{P}(Y)}{\hat{P}(X)}
\]

\[
= \arg \max_{y \in \{0,1\}} \hat{P}(X \mid Y) \hat{P}(Y)
\]

- \( \hat{P}(Y = 1 \mid x) \): estimated probability a heart is healthy given \( x \)
- \( X = (X_1, X_2, \ldots, X_{300}) \): whether 300 regions of interest (ROI) are healthy (1) or unhealthy (0)

How many parameters do we have to learn?

\[
\hat{P}(X \mid Y) \quad \hat{P}(Y)
\]

A. \( 2 \cdot 2 + 2 = 6 \)

B. \( 2 \cdot 300 + 2 = 602 \)

C. \( 2 \cdot 2^{300} + 2 = \text{a lot} \)
Brute Force Bayes: $m = 300$ (# features)

\[
\hat{Y} = \arg\max_{y=\{0,1\}} \hat{P}(Y | X)
\]

\[
= \arg\max_{y=\{0,1\}} \frac{\hat{P}(X|Y)\hat{P}(Y)}{\hat{P}(X)}
\]

\[
= \arg\max_{y=\{0,1\}} \hat{P}(X|Y)\hat{P}(Y)
\]

- $\hat{P}(Y = 1 | x)$: estimated probability a heart is healthy given $x$
- $X = (X_1, X_2, ..., X_{300})$: whether 300 regions of interest (ROI) are healthy (1) or unhealthy (0)

How many parameters do we have to learn?

\[
\hat{P}(X|Y), \hat{P}(Y)
\]

A. $2 \cdot 2 + 2 = 6$
B. $2 \cdot 300 + 2 = 602$
C. $2 \cdot 2^{300} + 2 = \text{a lot}$

This approach requires you to learn $O(2^m)$ parameters.
Brute Force Bayes: $m = 300$ (# features)

$\hat{P}(Y = 1 | \boldsymbol{x})$: estimated probability a heart is healthy given $\boldsymbol{x}$

$\boldsymbol{X} = (X_1, X_2, \ldots, X_{300})$: whether 300 regions of interest (ROI) are healthy (1) or unhealthy (0)

How many parameters do we have to learn?

- A. $2 \cdot 2 + 2 = 6$
- B. $2 \cdot 300 + 2 = 602$
- C. $2 \cdot 2^{300} + 2 = \text{a lot}$

This approach requires you to learn $O(2^m)$ parameters.
The problem with our current classifier

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(Y | X) \]

Choose the \( Y \) that is most likely given \( X \)

\[ = \arg \max_{y=\{0,1\}} \frac{\hat{P}(X|Y)\hat{P}(Y)}{\hat{P}(X)} \]  

(Bayes’ Theorem)

\[ = \arg \max_{y=\{0,1\}} \hat{P}(X|Y)\hat{P}(Y) \]

(\( 1/P(X) \) is constant w.r.t. \( y \))

\[ \rightarrow \hat{P}(X_1, X_2, \ldots, X_m | Y) \]

Estimating this joint conditional distribution is intractable.

What if we could make a simplifying (but naïve) assumption to make estimation easier?

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
The Naïve Bayes assumption

\[ \hat{Y} = \arg \max_{y \in \{0, 1\}} \hat{P}(Y | X) \]

\[ = \arg \max_{y \in \{0, 1\}} \frac{\hat{P}(X|Y)\hat{P}(Y)}{\hat{P}(X)} \]

Assumption:

\[ X_1, \ldots, X_m \text{ are conditionally independent given } Y. \]

Naïve Bayes Assumption

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Naïve Bayes Classifier

\[
\hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y)
\]

Training

What is the Big-O of # of parameters we need to learn?

A. \(O(m)\)
B. \(O(2^m)\)
C. other
Naïve Bayes Classifier

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y) \]

for \( j = 1, \ldots, m: \)

\[ \hat{P}(X_j = 1|Y = 0), \quad \hat{P}(X_j = 1|Y = 1) \]

\( \hat{P}(Y = 1) \)

Training

Testing

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \left( \log \hat{P}(Y) + \sum_{j=1}^{m} \log \hat{P}(X_j|Y) \right) \) (for numeric stability)
### Classification terminology check

Training data: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\)

<table>
<thead>
<tr>
<th>Movie 1</th>
<th>Movie 2</th>
<th>Movie m</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image]</td>
<td>[Image]</td>
<td>[Image]</td>
<td>[Image]</td>
</tr>
</tbody>
</table>

1. \(x^{(i)}\)  
2. \(y^{(i)}\)  
3. \((x^{(i)}, y^{(i)})\)  
4. \(x_j^{(i)}\)

1: like movie  
0: dislike movie

---

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Classification terminology check

Training data: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\)

A. \(x^{(i)}\)
B. \(y^{(i)}\)
C. \((x^{(i)}, y^{(i)})\)
D. \(x_j\)

1: like movie
0: dislike movie

<table>
<thead>
<tr>
<th>User 1</th>
<th>Movie 1</th>
<th>Movie 2</th>
<th>Movie (m)</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 2</td>
<td>Movie 1</td>
<td>Movie 2</td>
<td>Movie (m)</td>
<td>Output</td>
</tr>
<tr>
<td>User (n)</td>
<td>Movie 1</td>
<td>Movie 2</td>
<td>Movie (m)</td>
<td>Output</td>
</tr>
</tbody>
</table>

1. 1 0 \cdots 1
2. 1
3. 1 1 \cdots 0
4. 0

\(i\): \(i\)-th user
\(j\): movie \(j\)
NETFLIX and Learn
Predicting user TV preferences

Will a user like the Pokémon TV series?

Observe indicator variables $X = (X_1, X_2)$:

- $X_1 = 1$: “likes Star Wars”
- $X_2 = 1$: “likes Harry Potter”

Output $Y$ indicator:

- $Y = 1$: “likes Pokémon”

Predict $\hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(Y | X)$
Predicting user TV preferences

Which probabilities do you need to estimate?
How many are there?

- Brute Force Bayes (strawman, without NB assumption)
- Naïve Bayes

During training, how to estimate the prob
\[ \hat{P}(X_1 = 1, X_2 = 1|Y = 0) \] with MLE? with Laplace?

- Brute Force Bayes
- Naïve Bayes
Predicting user TV preferences

Which probabilities do you need to estimate?
How many are there?

- Brute Force Bayes (strawman, without NB assumption)

- Naïve Bayes

During training, how to estimate the prob
\( \hat{P}(X_1 = 1, X_2 = 1 | Y = 0) \) with MLE? with Laplace?

- Brute Force Bayes

- Naïve Bayes

\[ \hat{y} = \arg \max_{y \in \{0,1\}} \hat{p}(X|Y)\hat{p}(Y) \]

Naïve Bayes Assumption

\[ P(X|Y) = \prod_{j=1}^{m} P(X_j|Y) \]
(strawman brute force) Multinomial MLE and MAP

Model: Multinomial, $m$ outcomes: $p_j$ probability of outcome $j$

Observe: $n_j = \#$ of trials with outcome $j$
Total of $\sum_{j=1}^{m} n_j$ trials

-----

**MLE**

\[
\hat{p}_j = \frac{n_j}{\sum_{j=1}^{m} n_j}
\]

**Laplace estimate** (MAP w/Laplace smoothing)

\[
\hat{p}_j = \frac{n_j + 1}{\sum_{j=1}^{m} n_j + m}
\]

----

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

training data

\[
\hat{P}(X_1 = 1 \; X_2 = 1|Y = 0)
\]
(Naïve Bayes) Multinomial MLE and MAP

Model: Multinomial, $m$ outcomes:
$p_j$ probability of outcome $j$

Observe: $n_j$ = # of trials with outcome $j$
Total of $\sum_{j=1}^{m} n_j$ trials

MLE
$\hat{p}_j = \frac{n_j}{\sum_{j=1}^{m} n_j}$

Laplace estimate (MAP w/Laplace smoothing)
$\hat{p}_j = \frac{n_j + 1}{\sum_{j=1}^{m} n_j + m}$

$\hat{P}(X_1 = 1 \ X_2 = 1|Y = 0)$

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

training data

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
and Learn

naively
Ex 1. Naïve Bayes Classifier (MLE)

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y) \]

Training:

\[ \forall i: \hat{P}(X_j = 1|Y = 0), \hat{P}(X_j = 0|Y = 0), \] Use MLE or

\[ \hat{P}(X_j = 1|Y = 1), \hat{P}(X_j = 0|Y = 0), \] Laplace (MAP)

\[ \hat{P}(Y = 1), \hat{P}(Y = 0) \]

Testing:

\[ \hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y) \]
Training: Naïve Bayes for TV shows (MLE)

Observe indicator vars. $X = (X_1, X_2)$:
- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Predict $Y$: “likes Pokémon”

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

Training data counts

1. How many datapoints ($n$) are in our training data?

2. Compute MLE estimates for $\hat{P}(X_1|Y)$:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{P}(X_1 = 0</td>
<td>Y = 0)$</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>$\hat{P}(X_1 = 0</td>
<td>Y = 1)$</td>
</tr>
</tbody>
</table>
Training: Naïve Bayes for TV shows (MLE)

Observe indicator vars. $X = (X_1, X_2)$:

- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Predict $Y$: “likes Pokémon”

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

Training data counts

1. How many datapoints ($n$) are in our training data?

2. Compute MLE estimates for $\hat{P}(X_1 | Y)$:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Training: Naïve Bayes for TV shows (MLE)

Observe indicator vars. $X = (X_1, X_2)$:
- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Predict $Y$: “likes Pokémon”

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>0.23</td>
<td>0.77</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.24</td>
<td>0.76</td>
</tr>
</tbody>
</table>

(from last slide)

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>5/13 ≈ 0.38</td>
<td>8/13 ≈ 0.62</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>7/17 ≈ 0.41</td>
<td>10/17 ≈ 0.59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>13/30 ≈ 0.43</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>17/30 ≈ 0.57</td>
<td></td>
</tr>
</tbody>
</table>
Training: Naïve Bayes for TV shows (MLE)

Observe indicator vars. $X = (X_1, X_2)$:

- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Predict $Y$: “likes Pokémon”

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$Y$</th>
<th></th>
<th>$X_2$</th>
<th>$Y$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0.23</td>
<td>0.77</td>
<td></td>
<td>0.38</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.24</td>
<td>0.76</td>
<td></td>
<td>0.41</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.57</td>
<td></td>
<td>1</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>

Now that we’ve trained and found parameters, it’s time to classify new users!
Ex 1. Naïve Bayes Classifier (MLE)

\[ \hat{Y} = \arg\max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y) \]

Training

\forall i: \hat{P}(X_j = 1|Y = 0), \hat{P}(X_j = 0|Y = 0), \quad \text{Use MLE or Laplace (MAP)}

\hat{P}(Y = 1), \hat{P}(Y = 0)

Testing

\[ \hat{Y} = \arg\max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y) \]
Testing: Naïve Bayes for TV shows (MLE)

Observe indicator vars. $X = (X_1, X_2)$:

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0.23</td>
<td>0.77</td>
</tr>
<tr>
<td>1</td>
<td>0.24</td>
<td>0.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td>1</td>
<td>0.41</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Predict $Y$: “likes Pokémon”

Suppose a new person “likes Star Wars” ($X_1 = 1$) but “dislikes Harry Potter” ($X_2 = 0$).

Will they like Pokémon? Need to predict $Y$:

$$\hat{Y} = \arg \max_{y \in \{0, 1\}} \hat{P}(X|Y)\hat{P}(Y) = \arg \max_{y \in \{0, 1\}} \hat{P}(X_1|Y)\hat{P}(X_2|Y)\hat{P}(Y)$$

If $Y = 0$:

$$\hat{P}(X_1 = 1|Y = 0)\hat{P}(X_2 = 0|Y = 0)\hat{P}(Y = 0) = 0.77 \cdot 0.38 \cdot 0.43 = 0.126$$

If $Y = 1$:

$$\hat{P}(X_1 = 1|Y = 1)\hat{P}(X_2 = 0|Y = 1)\hat{P}(Y = 1) = 0.76 \cdot 0.41 \cdot 0.57 = 0.178$$

Since term is greatest when $Y = 1$, predict $\hat{Y} = 1$
**Ex 2. Naïve Bayes Classifier (MAP)**

\[
\hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y)
\]

**Training**

∀ \(i\):
\[
\hat{P}(X_j = 1|Y = 0), \hat{P}(X_j = 0|Y = 0), \quad \hat{P}(X_j = 1|Y = 1), \hat{P}(X_j = 0|Y = 0), \\
\hat{P}(Y = 1), \hat{P}(Y = 0)
\]

Use MLE or Laplace (MAP)

**Testing**

\[
\hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j|Y) \right) \hat{P}(Y)
\]

(note the same as before)
Training: Naïve Bayes for TV shows (MAP)

Observe indicator vars. \( X = (X_1, X_2) \):
- \( X_1 \): “likes Star Wars”
- \( X_2 \): “likes Harry Potter”

Predict \( Y \): “likes Pokémon”

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X_2 )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
\hat{P}(X_j = x \mid Y = y) = \frac{\#(X_j = x, Y = y) + 1}{\#(Y = y) + 2}.
\]

What are our MAP estimates using Laplace smoothing for \( \hat{P}(X_j \mid Y) \)?

A. \[
\frac{\#(X_j = x, Y = y)}{\#(Y = y)}
\]

B. \[
\frac{\#(X_j = x, Y = y) + 1}{\#(Y = y) + 2}
\]

C. \[
\frac{\#(X_j = x, Y = y) + 1}{\#(Y = y) + 4}
\]

D. other

Training data counts

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Training: Naïve Bayes for TV shows (MAP)

Observe indicator vars. \( X = (X_1, X_2) \):
- \( X_1 \): “likes Star Wars”
- \( X_2 \): “likes Harry Potter”

Predict \( Y \): “likes Pokémon”

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.27</td>
<td>0.73</td>
</tr>
<tr>
<td>1</td>
<td>0.26</td>
<td>0.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X_2 )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td>1</td>
<td>0.42</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Training data counts

\[
\hat{Y} = \arg \max_{y \in \{0,1\}} \left( \prod_{i=1}^{m} \hat{P}(X_i | Y) \right) \hat{P}(Y)
\]

In practice:
- We use Laplace for \( \hat{P}(X_j | Y) \) in case some events \( X_j = x_j \) don’t appear
- We don’t use Laplace for \( \hat{P}(Y) \), because all class labels should appear reasonably often
Naïve Bayes Model is a Bayesian Network

Naïve Bayes Assumption

\[ P(X|Y) = \prod_{j=1}^{m} P(X_j|Y) \]

Which Bayesian Network encodes this conditional independence?

A. \[ Y \]
   \[ X_1 \quad X_2 \quad \ldots \quad X_n \]
   \( X_i \) are conditionally independent given \( Y \)

B. \[ X_1 \quad X_2 \quad \ldots \quad X_n \]
   \[ Y \]

\( n \), and Jerry Cain, CS109, Spring 2021
Naïve Bayes Model is a Bayesian Network

\[
P(X|Y) = \prod_{i=1}^{m} P(X_i|Y) \quad \Rightarrow \quad P(X, Y) = P(Y) \prod_{j=1}^{m} P(X_j|Y)
\]

Which Bayesian Network encodes this conditional independence?

A. $Y \rightarrow X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n$

B. $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_n \rightarrow Y$

$X_i$ are conditionally independent given parent $Y$