24: Naïve Bayes

Jerry Cain
May 20, 2022

Table of Contents

  2 Intro: Machine Learning
  18 Brute Force Bayes
  29 Naïve Bayes
  38 Netflix and Learning
  56 Spam and Learning
Intro: Machine Learning
Our path from here

- Deep Learning
- Linear Regression
- Naïve Bayes
- Logistic Regression

Parameter Estimation
Our path from here

Deep Learning

- Linear Regression
- Naïve Bayes
- Logistic Regression

Unbiased estimators

- $\bar{X}, S^2$

Maximizing likelihood

- $\theta_{MLE}$

Bayesian estimation

- $\theta_{MAP}$
Machine Learning (formally)

Many different forms of Machine Learning
• We focus on the problem of **prediction** given prior observations.
Machine Learning uses a lot of data.

Task: Identify the chair
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

1. Real World Problem
2. Model the problem
   - Formal Model $\theta$
3. Learning Algorithm
4. Testing Data
5. Prediction Function $\hat{\theta}$
6. Training Data
7. Evaluation score
Supervised learning

Real World Problem

Model the problem

Formal Model $\theta$

Learning Algorithm

Training Data

Prediction Function $\hat{\theta}$

Testing Data

Evaluation score

Modeling

not the focus of this class (take CS228)
Supervised learning

Parameter estimation is a basis for learning from data.

Training Data

Real World Problem

Model the problem

Formal Model $\theta$

Learning Algorithm

Prediction Function $\hat{\theta}$

Evaluation score

Testing Data

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2022
### Model and dataset

Many different forms of machine learning

- We focus on a specific type of problem: **prediction** from observations.

#### Goal

Based on observed $X$, predict some unknown $Y$

#### Features

Vector $X$ of $m$ observations (new term: **feature vector**)

$$X = (X_1, X_2, ..., X_m)$$

#### Output

Variable $Y$ (also called **class label** if discrete)

#### Model

$$\hat{Y} = g(X), \text{ a function on } X$$
Training data

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

<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>...</td>
</tr>
<tr>
<td>Patient 2</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient n</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>
Training data notation

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

\(n\) datapoints, assumed to be iid

\(i\)-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

1. Real World Problem
2. Model the problem
3. Formal Model $\theta$
4. Learning Algorithm
5. Testing Data
6. Prediction Function $\hat{\theta}$
7. Evaluation score
Testing data notation

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

\(n_{test}\) other datapoints, assumed to be iid

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

- Has the same structure as your training data

Testing Goal:

Leveraging 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 \textit{prediction} based on observations.

\begin{itemize}
  \item \textbf{Goal} \quad Based on observed $X$, predict some unknown $Y$
  \item \textbf{Features} \quad Vector $X$ of $m$ observations (new term: feature vector) \quad $X = (X_1, X_2, \ldots, X_m)$
  \item \textbf{Output} \quad Variable $Y$ (also called \textit{class label} if discrete)
\end{itemize}

\begin{itemize}
  \item \textbf{Model} \quad $\hat{Y} = g(X)$, a function on $X$
  \item \textbf{Regression} \quad prediction when $Y$ is continuous
  \item \textbf{Classification} \quad prediction when $Y$ is discrete
\end{itemize}
Regression: Predicting real numbers

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

<table>
<thead>
<tr>
<th>Year</th>
<th>CO2 levels</th>
<th>Sea level</th>
<th>Feature (m)</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>338.8</td>
<td>0</td>
<td>... 1</td>
<td>0.26</td>
</tr>
<tr>
<td>Year 2</td>
<td>340.0</td>
<td>1</td>
<td>... 0</td>
<td>0.32</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Year (n)</td>
<td>340.76</td>
<td>0</td>
<td>... 1</td>
<td>0.14</td>
</tr>
</tbody>
</table>
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>...</td>
<td>...</td>
<td>...</td>
<td>...</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 2022
Brute Force

Bayes
Classification: Having a healthy heart

\[ X = (X_1) \]

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

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

How can we predict whether 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 \( X \)

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

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

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

(Bayes’ Theorem)

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

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

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

Training Goal: Use \( n \) datapoints to learn \( 2 \cdot 2 + 2 = 6 \) parameters.

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

\[
\begin{array}{c|cc}
\text{Patient 1} & \hat{P}(X|Y = 0) & \hat{P}(X|Y = 1) \\
0 & \theta_1 & \theta_3 \\
1 & \theta_2 & \theta_4 \\
\end{array}
\]

\[
\begin{array}{c|c}
\text{Marginal probability table } \hat{P}(Y) & \\
Y = 0 & \theta_5 \\
Y = 1 & \theta_6 \\
\end{array}
\]

\[ X = (X_1) \]

Feature 1

Output

\[
\begin{aligned}
\text{Feature 1} & \quad \text{Output} \\
\text{Patient 1} & \quad 0 \\
\text{Patient 2} & \quad 1 \\
\vdots & \quad \vdots \\
\text{Patient } n & \quad 1
\end{aligned}
\]
Training: Estimate parameters $\hat{P}(X|Y)$

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

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

Use MLE or Laplace (MAP) estimate $\hat{P}(X|Y)$ and $\hat{P}(Y)$ as parameters.

<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>

Patient $n$ 0 1
Training: MLE estimates, $\hat{P}(X|Y)$

$$\hat{P}(X|Y = 0) \begin{array}{c|c|c} X_1 = 0 & 0.4 & 0.0 \\ X_1 = 1 & 0.6 & 1.0 \end{array}$$

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

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

<table>
<thead>
<tr>
<th>X_1 = 0</th>
<th>X_1 = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X</td>
<td>Y = 0)$</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

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, $\hat{P}(X|Y)$

<table>
<thead>
<tr>
<th>Count: $X_1 = 0, Y = 0$:</th>
<th># datpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 = 1, Y = 0$:</td>
<td>4</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>

### MLE

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

Just count!

### MAP

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

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

| (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?
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 \(P(X|Y)\))

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

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

Test Set

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
Brute Force Bayes: $m = 300$ (# features)

$$X = (X_1, X_2, X_3, ..., X_{300})$$

<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>...</td>
<td></td>
<td>:</td>
<td></td>
</tr>
<tr>
<td>Patient $n$</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

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

$$X = (X_1, X_2, X_3, ..., X_{300})$$

<table>
<thead>
<tr>
<th>Count:</th>
<th># datapoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 = 0, X_2 = 0, ..., X_{299} = 0, X_{300} = 0, Y = 0$:</td>
<td>0</td>
</tr>
<tr>
<td>$X_1 = 0, X_2 = 0, ..., X_{299} = 0, X_{300} = 1, Y = 0$:</td>
<td>0</td>
</tr>
<tr>
<td>$X_1 = 0, X_2 = 0, ..., X_{299} = 1, X_{300} = 0, Y = 0$:</td>
<td>1</td>
</tr>
</tbody>
</table>

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

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

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

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

- $\hat{P}(Y = 1 | x)$: 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?

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

A. $2 \cdot 2 + 2 = 6$
B. $2 \cdot 300 + 2 = 602$
C. $2 \cdot 2^{300} + 2 = \text{a lot}$
This approach requires you to learn $O(2^m)$ parameters.

**Brute Force Bayes:** $m = 300$ (# features)

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

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

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

Learn parameters through MLE or MAP

- $\hat{P}(Y = 1 | x)$: 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?

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

A. $2 \cdot 2 + 2 = 6$
B. $2 \cdot 300 + 2 = 602$
C. $2 \cdot 2^{300} + 2 = \text{a lot}$
The problem with our current classifier

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

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

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

(Bayes’ Theorem)

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

\[ \Rightarrow \hat{P}(X_1, X_2, \ldots, X_m \mid Y) \]

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

Estimating this joint conditional distribution is intractable.

What if we could make a simplifying assumption—even if incredibly naïve—to make our parameter estimation effort computationally tractable?
The Naïve Bayes assumption

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

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

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

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

Assumption:

\( X_1, \ldots, X_m \) are all \textit{conditionally independent} given \( Y \). 

Naïve Bayes Assumption
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

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

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

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

Training

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

Testing

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

Use MLE or Laplace (MAP)
NETFLIX and Learn
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. \(x^{(i)}\)

1: like movie
0: dislike movie

User 1
1. 1 0 ... 1
2. 1

User 2
3. 1 1 ... 0
4. 0

... :: ::

User n
0 4. 0 ... 1
Classification terminology check

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

Movie 1  Movie 2  ...  Movie \(m\)  Output

User 1 1. 1 0 ... 1 2. 1
User 2 3. 1 1 ... 0 0
... ... ... ... ... ...
User \(n\) 4. 0 0 ... 1 1

1: like movie
0: dislike movie

\(i\): \(i\)-th user
\(j\): movie \(j\)
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 $\hat{Y} = \arg \max_{y=\{0,1\}} \hat{P}(Y \mid X)$
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
(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)

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

\[ \hat{P}(X_1 = 1 \text{ and } X_2 = 1 | Y = 0) \]

<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
# (Naïve Bayes) Multinomial MLE and MAP

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

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

<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

- $\hat{p}(X_1 = 1 \text{ and } X_2 = 1 | Y = 0)$

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

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

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

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

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>( X_2 )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>3</td>
<td>( 0 )</td>
<td>5</td>
</tr>
<tr>
<td>( 1 )</td>
<td>4</td>
<td>( 1 )</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>( Y )</th>
<th>( X_2 )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( \hat{P}(X_1 = 0</td>
<td>Y = 0) )</td>
<td>( \hat{P}(X_1 = 1</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( \hat{P}(X_1 = 0</td>
<td>Y = 1) )</td>
<td>( \hat{P}(X_1 = 1</td>
</tr>
</tbody>
</table>

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

2. Compute MLE estimates for \( \hat{P}(X_1|Y) \):
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>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>$Y$</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>5</td>
<td>8</td>
</tr>
<tr>
<td>$Y$</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

### Training data counts

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3 10</td>
<td>0 5</td>
</tr>
<tr>
<td>1</td>
<td>4 13</td>
<td>1 8</td>
</tr>
</tbody>
</table>

---

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

2. Compute MLE estimates for $\hat{P}(X_1 | Y)$:
### 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>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td></td>
<td></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>

(from last slide)

<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>5/13</td>
<td>8/13</td>
</tr>
<tr>
<td>1</td>
<td>7/17</td>
<td>10/17</td>
</tr>
</tbody>
</table>

Training data counts:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

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

$\frac{5}{13} \approx 0.38$, $\frac{8}{13} \approx 0.62$

$\frac{7}{17} \approx 0.41$, $\frac{10}{17} \approx 0.59$

$\frac{13}{30} \approx 0.43$, $\frac{17}{30} \approx 0.57$
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>
</table>
| $Y$   | ![Table](image)

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
</table>
| $Y$   | ![Table](image)

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

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

Training

\forall i: \hat{P}(X_j = 1|Y = 0), \hat{P}(X_j = 0|Y = 0), \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

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

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

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

Predict $Y$: “likes Pokémon”

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

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

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

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

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**

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

**Testing**

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

Use MLE or Laplace (MAP) (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 ) ( Y )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

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

Training data counts

\[ \hat{P}(X_j = x | Y = y) : \]

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

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

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>1</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>0</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

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

\[
\hat{y} = \arg\max_{y\in\{0,1\}} \left( \prod_{i=1}^m \hat{P}(X_i | Y) \right) \hat{P}(Y)
\]
and Learn naively
What is Bayes doing in my mail server?

Let’s get Bayesian on your spam:

Content analysis details: (49.5 hits, 7.0 required)

0.9 RCVD_IN_PBL  RBL: Received via a relay in Spamhaus PBL [93.40.189.29 listed in zen.spamhaus.org]
1.5 URIBL_WS_SURBL Contains an URL listed in the WS SURBL blocklist [URIs: recragas.cn]
5.0 URIBL_JP_SURBL Contains an URL listed in the JP SURBL blocklist [URIs: recragas.cn]
5.0 URIBL_OB_SURBL Contains an URL listed in the OB SURBL blocklist [URIs: recragas.cn]
5.0 URIBL_SC_SURBL Contains an URL listed in the SC SURBL blocklist [URIs: recragas.cn]
2.0 URIBL_BLACK Contains an URL listed in the URIBL blacklist [URIs: recragas.cn]
8.0 BAYES_99 BODY: Bayesian spam probability is 99 to 100% [score: 1.0000]
Ex 3. Naïve Bayes Classifier \((m, n \text{ large})\)

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

- **Training**
  \[
  \forall i: \hat{P}(X_i|Y) = 0, \text{ Use MLE or Laplace (MAP)}
  \]

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

What changes are necessary?
Email classification

Goal
Based on email content $X$, predict if email is spam or not.

Features
Consider a lexicon $m$ words (for English: $m \approx 100,000$).

$X = (X_1, X_2, ..., X_m)$, $m$ indicator variables

$X_j = 1$ if word $j$ appeared in document

Output
$Y = 1$ if email is spam

Note: $m$ is huge. Make Naïve Bayes assumption: $P(X|\text{spam}) = \prod_{j=1}^{m} P(X_j|\text{spam})$

Appearances of words in email are conditionally independent given the email is spam or not.
Training: Naïve Bayes Email classification

Train set: 

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

\[ x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}) \] for each word, whether it appears in email \( i \)

\[ y^{(i)} = 1 \text{ if spam, 0 if not spam} \]

Note: \( m \) is huge.

Which estimator should we use for \( \hat{P}(X_j|Y) \)?

A. MLE
B. Laplace estimate (MAP)
C. Other MAP estimate
D. Both A and B
Training: Naïve Bayes Email classification

Train set

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

\[ x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}) \] for each word, whether it appears in email \( i \)

\[ y^{(i)} = 1 \text{ if spam, 0 if not spam} \]

Note: \( m \) is huge.

Which estimator should we use for \( \hat{P}(X_j | Y) \)?

A. MLE
B. Laplace estimate (MAP)
C. Other MAP estimate
D. Both A and B

Many words are likely to not appear at all in the training set!
Ex 3. Naïve Bayes Classifier \((m, n \text{ large})\)

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

**Training**

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

**Testing**

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

Laplace (MAP) estimates avoid estimating 0 probabilities for events that don’t occur in your training data.

Use MLE or **Laplace (MAP)**
Testing: Naïve Bayes Email classification

For a new email:

- Generate $X = (X_1, X_2, ..., X_m)$
- Classify as spam or not using Naïve Bayes assumption

Note: $m$ is huge.

Suppose train set size $n$ also huge (many labeled emails).
Can we still use the below prediction?

$$\hat{Y} = \arg \max_{y=\{0,1\}} \left( \prod_{j=1}^{m} \hat{P}(X_j | Y) \right) \hat{P}(Y)$$
Testing: Naïve Bayes Email classification

For a new email:
- Generate $X = (X_1, X_2, \ldots, X_m)$
- Classify as spam or not using Naïve Bayes assumption

Note: $m$ is huge.

Suppose train set size $n$ also huge (many labeled emails).
Can we still use the below prediction?

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

Will probably lead to underflow!
Ex 3. Naïve Bayes Classifier ($m, n$ large)

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

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

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

Use sums of log-probabilities for numerical stability.

\[
\hat{Y} = \arg \max_{y=\{0,1\}} \left( \log \hat{P}(Y) + \sum_{j=1}^{m} \log \hat{P}(X_j|Y) \right)
\]
How well does Naïve Bayes perform?

After training, you can test with another set of data, called the test set.

- Test set also has known values for $Y$ so we can see how often we were right/wrong in our predictions $\hat{Y}$.

Typical workflow:
- Have a dataset of 1789 emails (1578 spam, 211 ham)
- Train set: First 1538 emails (by time)
- Test set: Next 251 messages

Evaluation criteria on test set:

- **Precision** = \[
\frac{\# \text{ correctly predicted class } Y}{\# \text{ predicted class } Y}
\]
- **Recall** = \[
\frac{\# \text{ correctly predicted class } Y}{\# \text{ real class } Y \text{ messages}}
\]

<table>
<thead>
<tr>
<th></th>
<th>Spam Prec.</th>
<th>Spam Recall</th>
<th>Non-spam Prec.</th>
<th>Non-spam Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words only</td>
<td>97.1%</td>
<td>94.3%</td>
<td>87.7%</td>
<td>93.4%</td>
</tr>
<tr>
<td>Words + addtl features</td>
<td>100%</td>
<td>98.3%</td>
<td>96.2%</td>
<td>100%</td>
</tr>
</tbody>
</table>
What are precision and recall?

Accuracy (# correct)/(# total) sometimes just doesn’t cut it.

<table>
<thead>
<tr>
<th>Precision:</th>
<th>Of the emails you predicted as spam, how many are <em>truly</em> spam?</th>
<th>Measure of false positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall:</td>
<td>Of the emails that are truly spam, how many did you predict?</td>
<td>Measure of false negatives</td>
</tr>
</tbody>
</table>