Point Estimation

• A random variable $X$ distributed according to $D(\theta)$
• Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from $D$
• Goal: Estimate $\theta$

• Three methods
  – Method of moments (MoM)
  – Maximum Likelihood Estimation (MLE)
  – Bayesian Estimation (Later)
Method of Moments

• Given a coin
  – Comes up heads(1) with probability $p$.
  – Comes up tails(0) with probability $1 - p$.
• Estimate $p$?
• Idea: match observed distribution to true distribution
  – Moments: an elegant way to achieve this.

\[
\hat{p} = \frac{\sum X_i}{n} = \bar{X}_n
\]

Mom estimate
Maximum Likelihood Estimation

• Given a coin
  – Comes up heads(1) with probability $p$.
  – Comes up tails(0) with probability $1 - p$.

• Estimate $p$?

• Idea: find $p$ that is most likely to generate the given data.

$X = \text{Coin}$

$\Theta = p$
Maximum Likelihood Estimation

• A random variable $X$ distributed according to $D(\theta)$
• Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from $D$
• Goal: Estimate $\theta$
• Idea: output $\hat{\theta}$ that is most likely to generate the data.
Maximum Likelihood Estimation

• Given a coin
  – Comes up heads(1) with probability $p$.
  – Comes up tails(0) with probability $1 - p$.

• Estimate $p$?

• Idea: find $p$ that is most likely to generate the given data.

For a given $\hat{p}$

$$L(\hat{p}) = P_{\hat{p}}(x_1, x_2, \ldots, x_n; \hat{p})$$

Likelihood function: probability that we see the given data assuming model is $\hat{p}$
Maximum Likelihood Estimation

For a given $\hat{p}$

$$L(\hat{p}) = P(x_1, x_2, \ldots, x_n \mid \hat{p})$$

Likelihood function

probability that we see the given data assuming model is $\hat{p}$

output $\arg\max \ L(\hat{p})$

MLE estimate
Maximum Likelihood Estimation

• Given a coin
  – Comes up heads(1) with probability $p$.
  – Comes up tails(0) with probability $1 - p$.

• Estimate $p$?

• Idea: find $p$ that is most likely to generate the given data.

\[
L(p) = P(X_1, X_2, \ldots, X_n \mid p) \\
= \prod_{i=1}^{n} P(X_i \mid p) \quad \text{(why?)}
\]

\[
P(X_i \mid p) = \begin{cases} 
  p & \text{if } X_i = 1 \\
  1 - p & \text{if } X_i = 0 
\end{cases}
\]

\[
= p^x (1-p)^{1-x_i}
\]
Maximum Likelihood Estimation

\[ L(\hat{\theta}) = \prod_{i=1}^{n} P(X_i; \hat{\theta}) \]

\[ = \prod_{i=1}^{n} (\hat{\theta})^{x_i} (1-\hat{\theta})^{1-x_i} \]

Typically easier to maximize

\[ \log L(\theta) = \log \text{likelihood of } \theta \]

Output

\[ \text{argmax } \log L(\hat{\theta}) \]
Maximum Likelihood Estimation

\[ L(\hat{\theta}) = \prod_{i=1}^{n} \theta^{x_i}(1-\hat{\theta})^{(1-x_i)} \]

\[ \log L(\theta) = \sum_{i=1}^{n} \log(\theta^{x_i}) + \log(1-\hat{\theta})^{(1-x_i)} \]

\[ = \sum_{i=1}^{n} x_i \log \theta + \sum_{i=1}^{n} (1-x_i) \log(1-\hat{\theta}) \]

\[ = a \log \theta + b \log(1-\hat{\theta}) \]

Set \[ \frac{\partial L}{\partial \hat{\theta}} = 0 \]

\[ \frac{d^2 L}{d \hat{\theta}^2} < 0 \]
Maximum Likelihood Estimation

\[ \log L(\hat{p}) = \sum_{i=1}^{n} x_i \log \hat{p} + \sum_{i=1}^{n} (1-x_i) \log (1-\hat{p}) \]

\[ \frac{2}{\hat{p}} \log L(\hat{p}) = \sum_{i=1}^{n} \frac{x_i}{\hat{p}} + \sum_{i=1}^{n} \frac{(1-x_i)}{1-\hat{p}} = 0 \]

\[ \Rightarrow \frac{\sum x_i}{\hat{p}} = \frac{\sum (1-x_i)}{1-\hat{p}} \]

\[ \Rightarrow 1-\hat{p} = \frac{n - \sum x_i}{\sum x_i} \]

\[ \Rightarrow \frac{1}{\hat{p}} - 1 = \frac{n}{\sum x_i} - 1 \]

\[ \hat{p} = \frac{\sum x_i}{n} \]

MLE estimate

Check \[ \frac{\partial^2 L}{\partial \hat{p}^2} < 0 \] (excluded)
Maximum Likelihood Estimation

• How good is the estimate?
• Need a notion of error
  – Mean squared error (MSE)

\[
\text{MSE of our estimator } = \frac{p(1-p)}{n}
\]

\[
\hat{\theta} = \frac{\sum x_i}{n} = \bar{x}_n
\]

MOM, MLE estimate of \( p \)
Maximum Likelihood Estimation

• How good is the estimate?
• Need a notion of error
  – Mean squared error (MSE)
  – How many samples?

\[ n \geq \frac{p(1-p)}{\epsilon} \]
Maximum Likelihood Estimation

• A random variable X distributed according to D(θ)
• Given i.i.d. samples \( X_1, X_2, \ldots, X_n \) from D
• Goal: Estimate \( \theta \)
• Idea: output \( \hat{\theta} \) that is most likely to generate the data.
  – Typically different than MoM
  – Involves maximizing a function
  – A lot of algorithmic tools from optimization
    • Gradient descent, newton methods etc.
  – Optimal among a certain class of estimators!
Maximum Likelihood Estimation

- Asymptotics of MLE
  - Unbiased
  - Optimal among all unbiased estimators
  - Approximately globally optimal
Asymptotics of MLE

• A random variable $X$ distributed according to $D(\theta)$
• Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from $D$
• Goal: Estimate $\theta$
• $\hat{\theta} = MLE$

\[
\sqrt{n}(\hat{\theta} - \theta) \to N(0, \sigma^2)
\]

\[
(\hat{\theta} - \theta) \approx N(0, \frac{\sigma^2}{n})
\]

asymptotically unbiased
Asymptotics of MLE

\( \sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \sigma^2) \)

\( \sigma^2 = \frac{1}{I(\theta)} \)

- Amount of information in the likelihood about \( \theta \).
Asymptotics of MLE

• How to calculate MLE?

\[ \frac{\partial L}{\partial \theta} = 0 \]

• How to check maximizer criteria?

\[ \frac{d^2 L}{d \theta^2} \leq 0 \]

\[ I(\theta) = \frac{1}{n} E \left[ -\frac{d^2 L}{d \theta^2} \right] \]

not a property of estimator.

\[ \text{Expectation over } X_1, \ldots, X_n \text{ assuming data coming from } P \]

• \( I(\theta) = \) how negative the second derivative is.
Maximum Likelihood Estimation

• Given a coin
  – Comes up heads(1) with probability $p$.
  – Comes up tails(0) with probability $1 - p$.
• Estimate $p$?
• Idea: find $p$ that is most likely to generate the given data.

\[
\frac{dL}{dp} = \frac{\sum X_i}{p} - \frac{\sum (1-X_i)}{1-p}
\]

\[
\frac{d^2L}{dp^2} = -\frac{\sum X_i}{p^2} - \frac{\sum (1-X_i)}{(1-p)^2} < 0
\]

\[
\text{Expectation over } X_1 \ldots X_n \text{ assuming data coming from } D(\hat{p})
\]

\[
\text{I}(\hat{p}) = \frac{1}{n} E \left[ \frac{d^2L}{dp^2} \right] = \frac{1}{n} E \left[ \frac{\sum X_i + \sum (1-X_i)}{p^2} \right]
\]

\[
= \frac{1}{n} \left[ \frac{n\hat{p}}{p^2} + \frac{n(1-\hat{p})}{(1-p)^2} \right] = \frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}
\]

\[
= \frac{1}{\hat{p}(1-\hat{p})}
\]
Maximum Likelihood Estimation

• A random variable $X$ distributed according to $D(\theta)$
• Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from $D$
• Goal: Estimate $\theta$
• Idea: output $\hat{\theta}$ that is most likely to generate the data.
• Asymptotics of MLE
  – Unbiased
  – Optimal among all unbiased estimators
  – Approximately optimal

\[
\hat{\theta} - \theta \sim N\left(0, \frac{1}{nI(\theta)} \right)
\]
Point Estimation

• A random variable \( X \) distributed according to \( D(\theta) \)
• Given i.i.d. samples \( X_1, X_2, \ldots, X_n \) from \( D \)
• Goal: Estimate \( \theta \)

• Three methods
  – Method of moments (MoM)
  – Maximum Likelihood Estimation (MLE)
  – Bayesian Estimation (Later)
Point Estimation

Whatever. Should I use MLE or not?

Yes and No
Maximum Likelihood Estimation

• Very general
  – First thing to try

• Strong theory
  – Only kicks in for large n

• Often best estimators are not MLE
  – Especially true in high dimensional settings.
  – Often best estimators are a “tweak” of the MLE.
ML is about tradeoffs

- MLE
- Naïve Bayes
- Logistic Regression
- SVMs
- Decision Trees
- Deep Nets

ML algorithm

ML problem

Time
Space
Data
Parallelization
Interpretability
Robustness

....

....
ML is about tradeoffs

• Useful to think about methods in terms of tradeoffs
  – When is a method optimal?
  – What key assumptions are needed for the method?
  – How robust is the method to its assumptions?
  – Which resources does the method use efficiently?
Prediction

• Random variables $X,Y$ distributed according to $D(X,Y)$
• Given i.i.d. samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from $D$
• Goal: Predict $P(Y|X = x)$ häufiger als Estimation, oft schwieriger zu beschreiben $D$
  – $X=$ images, $Y=$ is there a person?
  – $X=$ speech, $Y=$ male/female?
  – $X=$ movie review, $Y=$ positive/negative?
Your grade in 536

- You want to estimate your grade
- You have access to last year’s roster and grades
Your grade in 536

- Flip coin(p)
  - Heads – A grade
  - Tails – F grade

- Flip coin(q)
  - Heads – A grade
  - Tails – F grade

\[ P(Y = A) \] \text{RANDOM} \cong \text{MLE/ } \text{Mom estimate} \]
### Your grade in 536

<table>
<thead>
<tr>
<th>HW1</th>
<th>HW2</th>
<th>HW3</th>
<th>HW4</th>
<th>HW5</th>
<th>HW6</th>
<th>HW7</th>
<th>Midterm</th>
<th>Final</th>
<th>Coding skill</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
</tr>
</tbody>
</table>

\[ X_1 = A \quad X_2 = A \]

\[ P(Y = A | X_1 = A, X_2 = A, \ldots, X_{10} = F) \]

\[ \text{need} \quad n \geq \frac{P(1-P)}{\epsilon^2} \]

\[ \geq 25 \text{ samples} \quad \text{Per estimate} \]

- You want to estimate your grade
- You have access to last year’s roster and grades
## Your grade in 536

<table>
<thead>
<tr>
<th>HW1</th>
<th>HW2</th>
<th>HW3</th>
<th>HW4</th>
<th>HW5</th>
<th>HW6</th>
<th>HW7</th>
<th>Midterm</th>
<th>Final</th>
<th>Coding skill</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
</tr>
</tbody>
</table>

### Naïve Bayes Classification

- You want to estimate your grade
- You have access to last year’s roster and grades
Naïve Bayes Classification

• Random variables $X_1, X_2, \ldots, X_n, Y$ distributed according to $D$

• Given i.i.d. samples from $D$

• Goal: Predict $P(Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$

• Step 1: The Bayes Part

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad \text{Bayes Rule}
\]

\[
P(Y|X = x_1, -X_n = x_n) = \frac{P(Y)P(X_1 = x_1, -X_n = x_n | Y)}{P(X_1 = x_1, -X_n = x_n)}
\]
Naïve Bayes Classification

• Random variables $X_1, X_2, \ldots, X_n, Y$ distributed according to D

• Given i.i.d. samples from D

• Goal: Predict $P(Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$

• Step 2: The Naïve Part

$$P(Y|X_1 = x_1, \ldots, X_n = x_n) = \frac{P(Y) P(X_1 = x_1, \ldots, X_n = x_n | Y)}{P(X_1 = x_1, \ldots, X_n = x_n)}$$

Given $Y$, $X_i$'s are independent.

$$= \frac{P(Y) \prod_{i=1}^{n} P(X_i = x_i | Y)}{P(X_1 = x_1, \ldots, X_n = x_n)}$$
Naïve Bayes Classification

- Random variables $X_1, X_2, \ldots, X_n, Y$ distributed according to $D$
- Given i.i.d. samples from $D$
- Goal: Predict $P(Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$

\[
P(Y = y|X_1 = x_1, \ldots, X_n = x_n) \propto P(Y = y)P(X_1 = x_1, \ldots, X_n = x_n|Y = y)
\]

\[
P(Y = y|X_1 = x_1, \ldots, X_n = x_n) \propto P(Y = y) \prod_{i} P(X_i = x_i|Y = y)
\]
Your grade in 536

<table>
<thead>
<tr>
<th>HW1</th>
<th>HW2</th>
<th>HW3</th>
<th>HW4</th>
<th>HW5</th>
<th>HW6</th>
<th>HW7</th>
<th>Midterm</th>
<th>Final</th>
<th>Coding skill</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
</tr>
</tbody>
</table>

\[
P(Y = A | HW_1, \ldots, \text{Coding}) \propto \prod_{i=1}^{l_0} P(HW_i | Y = A)
\]

\[
P(Y = A) \prod_{i=1}^{l_0} P(X_i | Y)
\]

• What if \(X_i\)'s are continuous?
Naïve Bayes Classification

- Random variables $X_1, X_2, ... X_n, Y$ distributed according to $D$
- Given i.i.d. samples from $D$
- Goal: Predict $P(Y|X_1 = x_1, X_2 = x_2, ... X_n = x_n)$

- Algorithm: Given data
  - Estimate $P(Y = 0), P(Y = 1)$
  - Estimate
    - $P(X_i = 1|Y = 0), P(X_i = 0|Y = 0)$
    - $P(X_i = 1|Y = 1), P(X_i = 0|Y = 1)$
### Your grade in 536

<table>
<thead>
<tr>
<th>HW1</th>
<th>HW2</th>
<th>HW3</th>
<th>HW4</th>
<th>HW5</th>
<th>HW6</th>
<th>HW7</th>
<th>Midterm</th>
<th>Final</th>
<th>Coding skill</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
<td>A/F</td>
</tr>
</tbody>
</table>

**Naive Bayes Rule**

\[
P(Y=A|X_1, X_2, \ldots, X_6) \propto P(Y=A) \cdot P(X_i=A|Y=A) \cdot \ldots \cdot P(X_{i6}=F|Y=A)
\]
Naïve Bayes Classification

• Computationally efficient
• Easy to parallelize
• Need only one pass over the data
• Correctness relies on **Independence**
  – Typically too strong

• Often works even if the assumption is not true.

• Insensitive to irrelevant features.