CMSC 691
Probabilistic and Statistical Models of Learning
Probabilities, Common Distributions, and Maximum Likelihood Estimation
Outline

Basics of Learning

Probability

Maximum Likelihood Estimation
What does it mean to learn?

Chris has just begun taking a machine learning course

Pat, the instructor has to ascertain if Chris has “learned” the topics covered, at the end of the course

What is a “reasonable” exam?

(Bad) Choice 1: History of pottery
Chris’ s performance is not indicative of what was learned in ML

(Bad) Choice 2: Questions answered during lectures
Open book?

A good test should test ability to answer “related” but “new” questions on the exam

Generalization
Model, parameters and hyperparameters

Model: mathematical formulation of system (e.g., classifier)

Parameters: primary “knobs” of the model that are set by a learning algorithm

Hyperparameter: secondary “knobs”
score()
scoring model

\[ \text{score}_{\theta}(\cdot) \]

objective

\[ F(\theta) \]
scoring model

(score, )_\theta(\cdot)

objective

(implicitly) dependent on the observed data X=
Machine Learning Framework: Learning

Instances are typically examined independently.

Gold/correct labels give feedback to the predictor.

Evaluator

$F(\theta)$ objective
give feedback to the predictor

Machine Learning Predictor

Scoring model

$\text{score}_\theta(X)$
How do we optimize? Follow the derivative/gradient of our training score function

Set $t = 0$
Pick a starting value $\theta_t$
Until converged:
1. Get value $y_t = F(\theta_t)$
2. Get derivative $g_t = F'(\theta_t)$
3. Get scaling factor $\rho_t$
4. Set $\theta_{t+1} = \theta_t + \rho_t * g_t$
5. Set $t += 1$
Outline

Basics of Learning

Probability

Maximum Likelihood Estimation
Probability Topics (High-Level)

Basics of Probability: Prereqs

Philosophy of Probability, and Terminology

Useful Quantities and Inequalities
Probability Prerequisites

Basic probability axioms and definitions
Joint probability
Marginal probability
Probabilistic Independence
Definition of conditional probability

Bayes rule
Probability chain rule
Common distributions
Expected Value (of a function) of a Random Variable
(Most) Probability Axioms

\[ p(\text{everything}) = 1 \]

\[ p(\phi) = 0 \]

\[ p(A) \leq p(B), \text{ when } A \subseteq B \]

\[ p(A \cup B) = p(A) + p(B), \text{ when } A \cap B = \phi \]

\[ p(A \cup B) \neq p(A) + p(B) \]

\[ p(A \cup B) = p(A) + p(B) - p(A \cap B) \]
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails

- X is a random variable denoting the possible outcomes
- X=HEADS or X=TAILS
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails
    X is a random variable denoting the possible outcomes
    X=HEADS or X=TAILS

Example #2: Measuring the amount of snow that fell in the last storm
    Y is a random variable denoting the amount snow that fell, in inches
    Y=0, or Y=0.5, or Y=1.0495928591, or Y=10, or ...
Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails
   X is a random variable denoting the possible outcomes
   X=HEADS or X=TAILS

Example #2: Measuring the amount of snow that fell in the last storm
   Y is a random variable denoting the amount snow that fell, in inches
   Y=0, or Y=0.5, or Y=1.0495928591, or Y=10, or ...

DISCRETE random variable

CONTINUOUS random variable
## Random Variables

<table>
<thead>
<tr>
<th>If X is a...</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values $k$ that $X$ can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
</tbody>
</table>
## Random Variables

<table>
<thead>
<tr>
<th>If $X$ is a...</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values $k$ that $X$ can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
<tr>
<td>The function that gives the relative likelihood of a value $p(X=k)$ is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
</tbody>
</table>
# Random Variables

<table>
<thead>
<tr>
<th>If X is a…</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values k that X can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
<tr>
<td>The function that gives the relative likelihood of a value p(X=k) is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
<tr>
<td>The values that PMF/PDF can take are</td>
<td>0 ≤ p(X=k) ≤ 1</td>
<td>p(X=k) ≥ 0</td>
</tr>
</tbody>
</table>
# Random Variables

<table>
<thead>
<tr>
<th>If X is a...</th>
<th>Discrete random variable</th>
<th>Continuous random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The values $k$ that $X$ can take are</td>
<td>Discrete: finite or countably infinite (e.g., integers)</td>
<td>Continuous: uncountably infinite (e.g., real values)</td>
</tr>
<tr>
<td>The function that gives the relative likelihood of a value $p(X=k)$ is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
<tr>
<td>The values that PMF/PDF can take are</td>
<td>$0 \leq p(X=k) \leq 1$</td>
<td>$p(X=k) \geq 0$</td>
</tr>
<tr>
<td>We “add” with</td>
<td>Sums ($\sum$)</td>
<td>Integrals ($\int$)</td>
</tr>
</tbody>
</table>
## Random Variables

<table>
<thead>
<tr>
<th>The values k that X can take are</th>
<th>Discrete: finite or countably infinite (e.g., integers)</th>
<th>Continuous: uncountably infinite (e.g., real values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The function that gives the relative likelihood of a value p(X=k) is a</td>
<td>probability mass function (PMF)</td>
<td>probability density function (PDF)</td>
</tr>
<tr>
<td>The values that PMF/PDF can take are</td>
<td>0 ≤ p(X=k) ≤ 1</td>
<td>p(X=k) ≥ 0</td>
</tr>
<tr>
<td>We “add” with</td>
<td>Sums ((\sum))</td>
<td>Integrals ((\int))</td>
</tr>
<tr>
<td>Our PMF/PDF satisfies p(everything)=1 by</td>
<td>(\sum_k p(X = k) = 1)</td>
<td>(\int p(x)dx = 1)</td>
</tr>
</tbody>
</table>
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Marginal probability

Probabilistic Independence

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Joint Probability

Probability that multiple things “happen together”
Joint Probability

Probability that multiple things “happen together”

\[ p(x,y), \ p(x,y,z), \ p(x,y,w,z) \]

Symmetric: \( p(x,y) = p(y,x) \)
Joint Probability

Probability that multiple things “happen together”

\[ p(x,y), \quad p(x,y,z), \quad p(x,y,w,z) \]

Symmetric: \( p(x,y) = p(y,x) \)

Form a table based of outcomes: sum across cells = 1

<table>
<thead>
<tr>
<th>( p(x,y) )</th>
<th>( Y=0 )</th>
<th>( Y=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=“cat”</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>X=“dog”</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>X=“bird”</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>X=“human”</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Marginal probability

Probabilistic Independence

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Marginal(ized) Probability: The Discrete Case

Consider the mutually exclusive ways that different values of $x$ could occur with $y$

Q: How do write this in terms of joint probabilities?
Marginal(ized) Probability: The Discrete Case

Consider the mutually exclusive ways that different values of $x$ could occur with $y$

$$p(y) = \sum_{x} p(x, y)$$
Marginal(ized) Probability: The Discrete Case

Consider the mutually exclusive ways that different values of x could occur with y.

\[ p(y) = \sum_x p(x, y) \]

Q: What is \( p(y=1) \)?

<table>
<thead>
<tr>
<th>p(x,y)</th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=&quot;cat&quot;</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>X=&quot;dog&quot;</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>X=&quot;bird&quot;</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>X=&quot;human&quot;</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>
Marginal(ized) Probability: The Discrete Case

Consider the **mutually exclusive** ways that different values of $x$ could occur with $y$

$$p(y) = \sum_x p(x, y)$$

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>$y=0$</th>
<th>$y=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X=\text{“cat”}$</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>$X=\text{“dog”}$</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>$X=\text{“bird”}$</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>$X=\text{“human”}$</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>

Q: What is $p(y=1)$?

A: 0.56
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Marginal probability

Probabilistic Independence

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) \times p(y)$

Generalizable to > 2 random variables

Q: Are the results of flipping the same coin twice in succession independent?
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) \times p(y)$

Generalizable to $> 2$ random variables

Q: Are the results of flipping the same coin twice in succession independent?

A: Yes (assuming no weird effects)
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x) \times p(y)$

Generalizable to > 2 random variables

Q: Are $X$ and $Y$ independent?

<table>
<thead>
<tr>
<th>$p(x,y)$</th>
<th>$Y=0$</th>
<th>$Y=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X=“cat”$</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>$X=“dog”$</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>$X=“bird”$</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>$X=“human”$</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events.

Formally: $p(x,y) = p(x) \times p(y)$

Generalizable to > 2 random variables.

Q: Are X and Y independent?

<table>
<thead>
<tr>
<th>p(x,y)</th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=&quot;cat&quot;</td>
<td>.04</td>
<td>.32</td>
</tr>
<tr>
<td>X=&quot;dog&quot;</td>
<td>.2</td>
<td>.04</td>
</tr>
<tr>
<td>X=&quot;bird&quot;</td>
<td>.1</td>
<td>.1</td>
</tr>
<tr>
<td>X=&quot;human&quot;</td>
<td>.1</td>
<td>.1</td>
</tr>
</tbody>
</table>

A: No (find the marginal probabilities of p(x) and p(y)).
Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Marginal probability
- Probabilistic Independence
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
Conditional Probability

\[ p(X \mid Y) = \frac{p(X,Y)}{p(Y)} \]
Conditional Probability

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]

\[ p(Y) = \text{marginal probability of } Y \]
Conditional Probability

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]

\[ p(Y) = \int p(X, Y) dX \]
Revisiting Marginal Probability: The Discrete Case

\[ p(y) = \sum_x p(x, y) \]

\[ = \sum_x p(x)p(y \mid x) \]
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Marginal probability

Probabilistic Independence

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Deriving Bayes Rule

Start with conditional

$p(X \mid Y)$
Deriving Bayes Rule

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]
Deriving Bayes Rule

\[ p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \]

\[ p(X, Y) = p(X \mid Y)p(Y) \]

\[ p(X \mid Y) = \frac{p(Y \mid X) \ast p(X)}{p(Y)} \]
Bayes Rule

\[ p(X \mid Y) = \frac{p(Y \mid X) \times p(X)}{p(Y)} \]

- **posterior probability**
- **likelihood**
- **prior probability**
- **marginal likelihood**
- **probability**
Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Marginal probability
- Probabilistic Independence
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
Probability Chain Rule

\[ p(x_1, x_2, \ldots, x_S) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_S \mid x_1, \ldots, x_i) = \prod_{i} p(x_i \mid x_1, \ldots, x_{i-1}) \]

extension of Bayes rule
Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Marginal probability

Probabilistic Independence

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable
Distribution Notation

If $X$ is a R.V. and $G$ is a distribution:

- $X \sim G$ means $X$ is distributed according to ("sampled from") $G$
Distribution Notation

If X is a R.V. and G is a distribution:

• \( X \sim G \) means X is distributed according to ("sampled from") \( G \)
• \( G \) often has parameters \( \rho = (\rho_1, \rho_2, ..., \rho_M) \) that govern its "shape"
• Formally written as \( X \sim G(\rho) \)
Distribution Notation

If X is a R.V. and G is a distribution:

• $X \sim G$ means X is distributed according to ("sampled from") $G$

• $G$ often has parameters $\rho = (\rho_1, \rho_2, \ldots, \rho_M)$ that govern its "shape"

• Formally written as $X \sim G(\rho)$

**i.i.d.** If $X_1, X_2, \ldots, X_N$ are all independently sampled from $G(\rho)$, they are independently and identically distributed
Common Distributions

Bernoulli/Binomial
- Binary R.V.: 0 (failure) or 1 (success)
- $X \sim \text{Bernoulli}(\rho)$
- $p(X = 1) = \rho$, $p(X = 0) = 1 - \rho$
- Generally, $p(X = k) = \rho^k (1 - \rho)^{1-k}$

Categorical/Multinomial

Poisson

Normal

Gamma
Common Distributions

Bernoulli: A single draw
- Binary R.V.: 0 (failure) or 1 (success)
- $X \sim \text{Bernoulli}(\rho)$
- $p(X = 1) = \rho$, $p(X = 0) = 1 - \rho$
- Generally, $p(X = k) = \rho^k (1 - \rho)^{1-k}$

Binomial: Sum of $N$ iid Bernoulli draws
- Values $X$ can take: 0, 1, ..., $N$
- Represents number of successes
- $X \sim \text{Binomial}(N, \rho)$
- $p(X = k) = \binom{N}{k} \rho^k (1 - \rho)^{N-k}$
Common Distributions

Bernoulli/Binomial

- Finite R.V. taking one of K values: 1, 2, ..., K
- $X \sim \text{Cat}(\rho), \rho \in \mathbb{R}^K$
- $p(X = 1) = \rho_1, p(X = 2) = \rho_2, \ldots p(X = K) = \rho_K$
- Generally, $p(X = k) = \prod_j \rho_j^{1[k=j]}$

Categorical/Multinomial

- Vector of size K representing how often value k was drawn
- $X \sim \text{Multinomial}(N, \rho), \rho \in \mathbb{R}^K$
- $1[c] = \begin{cases} 1, & c \text{ is true} \\ 0, & c \text{ is false} \end{cases}$
Common Distributions

Poisson

- Discrete R.V. taking any integer that is $\geq 0$
- $X \sim \text{Poisson}(\lambda), \lambda \in \mathbb{R}$ is the “rate”
- $p(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$
Common Distributions

Normal
- Real R.V. taking any real number
- \( X \sim \text{Normal}(\mu, \sigma) \), \( \mu \) is the mean, \( \sigma \) is the standard deviation

\[
p(X = x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\]
Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

Poisson

Normal

Gamma

Multivariate Normal

- Real vector R.V. $X \in \mathbb{R}^k$
- $X \sim \text{Normal}(\mu, \Sigma)$, $\mu \in \mathbb{R}^K$ is the mean, $\Sigma \in \mathbb{R}^{K \times K}$ is the covariance

$$p(X = x) \propto \exp\left(- (x - \mu)^T \Sigma (x - \mu) \right)$$
Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

Poisson

Normal

Gamma

- Real R.V. taking any positive real number
- $X \sim \text{Gamma}(k, \theta), k > 0$ is the “shape” (how skewed it is), $\theta > 0$ is the “scale” (how spread out the distribution is)

$$p(X = x) = \frac{x^{k-1} \exp(-\frac{k}{\theta})}{\theta^k \Gamma(k)}$$

Probability Prerequisites

- Basic probability axioms and definitions
- Joint probability
- Marginal probability
- Probabilistic Independence
- Definition of conditional probability
- Bayes rule
- Probability chain rule
- Common distributions
- Expected Value (of a function) of a Random Variable
Expected Value of a Random Variable

\[ X \sim p(\cdot) \]
Expected Value of a Random Variable

\[ X \sim p(\cdot) \]

\[ \mathbb{E}[X] = \sum_x x \ p(x) \]
Expected Value: Example

uniform distribution of number of cats I have

\[
\mathbb{E}[X] = \sum_{x} x p(x)
\]

\[
\frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6
\]

= 3.5
Expected Value: Example

uniform distribution of number of cats I have

\[ E[X] = \sum_1^6 x \cdot p(x) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5 \]

Q: What common distribution is this?
Expected Value: Example

uniform distribution of number of cats I have

$E[X] = \sum_{x} x \, p(x)$

$1/6 \times 1 + 1/6 \times 2 + 1/6 \times 3 + 1/6 \times 4 + 1/6 \times 5 + 1/6 \times 6 = 3.5$

Q: What common distribution is this?
A: Categorical
Expected Value: Example 2

non-uniform distribution of number of cats a normal cat person has

\[ \mathbb{E}[X] = \sum_{x} x \cdot p(x) \]

\[ \frac{1}{2} \cdot 1 + \frac{1}{10} \cdot 2 + \frac{1}{10} \cdot 3 + \frac{1}{10} \cdot 4 + \frac{1}{10} \cdot 5 + \frac{1}{10} \cdot 6 \]

\[ = 2.5 \]
Expected Value of a Function of a Random Variable

\[ X \sim p(\cdot) \]

\[ \mathbb{E}[X] = \sum_{x} x p(x) \]

\[ \mathbb{E}[f(X)] = ? \]
Expected Value of a Function of a Random Variable

\[ X \sim p(\cdot) \]

\[ \mathbb{E}[X] = \sum_{x} x \cdot p(x) \]

\[ \mathbb{E}[f(X)] = \sum_{x} f(x) \cdot p(x) \]
Expected Value of Function: Example

non-uniform distribution of number of cats I start with

What if each cat magically becomes two?

\[ f(k) = 2^k \]

\[ \mathbb{E}[f(X)] = \sum_x f(x) \ p(x) \]
What if each cat magically becomes two?

\[ f(k) = 2^k \]

\[
\mathbb{E}[f(X)] = \sum_x f(x) p(x) = \sum_x 2^x p(x)
\]

\[
\frac{1}{2} \cdot 2^1 + \frac{1}{10} \cdot 2^2 + \frac{1}{10} \cdot 2^3 + \frac{1}{10} \cdot 2^4 + \frac{1}{10} \cdot 2^5 + \frac{1}{10} \cdot 2^6 = 13.4
\]
# Probability Prerequisites

<table>
<thead>
<tr>
<th>Basic probability axioms and definitions</th>
<th>Bayes rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint probability</td>
<td>Probability chain rule</td>
</tr>
<tr>
<td>Marginal probability</td>
<td>Common distributions</td>
</tr>
<tr>
<td>Probabilistic Independence</td>
<td>Expected Value (of a function) of a Random Variable</td>
</tr>
<tr>
<td>Definition of conditional probability</td>
<td></td>
</tr>
</tbody>
</table>
Example Problem: ITILA Ex. 2.3

➢ Jo has a test for a nasty disease. We denote Jo’s state of health by the variable $a$ ($a=1$: Jo has the disease; $a=0$ o/w) and the test result by $b$.

➢ The result of the test is either ‘positive’ ($b = 1$) or ‘negative’ ($b = 0$).

➢ The test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained.

➢ The final piece of background information is that 1% of people of Jo’s age and background have the disease.

Q: If Jo’s test is positive, what is the probability Jo has the disease?
Example Problem: ITILA Ex. 2.3

Q: If Jo’s test is positive, what is the probability Jo has the disease?

\[ p(a = 1 | b = 1) \]

➢ Jo has a test for a nasty disease. We denote Jo’s state of health by the variable \( a \) (\( a=1 \): Jo has the disease; \( a=0 \) o/w) and the test result by \( b \).

➢ The result of the test is either ‘positive’ (\( b = 1 \)) or ‘negative’ (\( b = 0 \)).

➢ The test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained.

➢ The final piece of background information is that 1% of people of Jo’s age and background have the disease.
Q: If Jo’s test is positive, what is the probability Jo has the disease?

\[
p(a = 1 | b = 1) = \frac{p(b = 1 | a = 1)p(a = 1)}{p(b = 1)}
\]

- Jo has a test for a nasty disease. We denote Jo’s state of health by the variable \(a\) (\(a=1\): Jo has the disease; \(a=0\) o/w) and the test result by \(b\).
- The result of the test is either ‘positive’ (\(b = 1\)) or ‘negative’ (\(b = 0\)).
- The test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained.
- The final piece of background information is that 1% of people of Jo’s age and background have the disease.

\[p(a = 1) = 0.01\]
Example Problem: ITILA Ex. 2.3

Q: If Jo’s test is positive, what is the probability Jo has the disease?

\[
p(a = 1 \mid b = 1) = \frac{p(b = 1 \mid a = 1)p(a = 1)}{p(b = 1)}
\]

- Jo has a test for a nasty disease. We denote Jo’s state of health by the variable \( a \) (a=1: Jo has the disease; a=0 o/w) and the test result by \( b \).
- The result of the test is either ‘positive’ (\( b = 1 \)) or ‘negative’ (\( b = 0 \)).
- The test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained.
- The final piece of background information is that 1% of people of Jo’s age and background have the disease.

\[
\begin{align*}
p(b = 1 \mid a = 1) &= 0.95 \\
p(b = 0 \mid a = 0) &= 0.95 \\
p(a = 1) &= 0.01
\end{align*}
\]
Example Problem: ITILA Ex. 2.3

Q: If Jo’s test is positive, what is the probability Jo has the disease?

- Jo has a test for a nasty disease. We denote Jo’s state of health by the variable \( a \) (\( a = 1 \): Jo has the disease; \( a = 0 \) o/w) and the test result by \( b \).
- The result of the test is either ‘positive’ (\( b = 1 \)) or ‘negative’ (\( b = 0 \)).
- The test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained.
- The final piece of background information is that 1% of people of Jo’s age and background have the disease.

\[
\begin{align*}
p(a = 1 | b = 1) &= \frac{p(b = 1 | a = 1)p(a = 1)}{p(b = 1)} \\
&= \frac{.95 \times .01}{.95 \times .01 + .05 \times .99} \\
&= 0.16
\end{align*}
\]
Probability Topics (High-Level)

Basics of Probability: Prereqs

Philosophy of Probability, and Terminology

Useful Quantities and Inequalities
A Bit of Philosophy and Terminology

What is a probability?

Core terminology
- Support/domain
- Partition function

Some principles
- Generative story
- Forward probability
- Inverse probability
Kinds of Statistics

Descriptive

Confirmatory

Predictive

The average grade on this assignment is 83.
Interpretations of Probability

Past performance
58% of the past 100 flips were heads

Hypothetical performance
If I flipped the coin in many parallel universes...

Subjective strength of belief
Would pay up to 58 cents for chance to win $1

Output of some computable formula?
$p(\text{heads})$ vs $q(\text{heads})$
Camps of Probability

Past performance
58% of the past 100 flips were heads

Hypothetical performance
If I flipped the coin in many parallel universes...

Subjective strength of belief
Would pay up to 58 cents for chance to win $1

Output of some computable formula?
p(heads) vs q(heads)

Frequentists
Bayesians

(my grouping, not too far off though)
Camps of Probability

Past performance
58% of the past 100 flips were heads

Hypothetical performance
If I flipped the coin in many parallel universes...

Subjective strength of belief
Would pay up to 58 cents for chance to win $1

Output of some computable formula?
p(heads) vs q(heads)

Frequentists
Bayesians
ML People

(my grouping, not too far off though)
Past performance
58% of the past 100 flips were heads

Hypothetical performance
If I flipped the coin in many parallel universes...

Subjective strength of belief
Would pay up to 58 cents for chance to win $1

Output of some computable formula?
p(heads) vs q(heads)

“You cannot do inference without making assumptions.”
– ITILA, 2.2, pg 26
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Partition these into two groups...

A  
B  
C  
D

Courtesy Hamed Pirsiavash
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Partition these into two groups

Who selected *red* vs. *blue*?
General ML Consideration: Inductive Bias

What do we know \textit{before} we see the data, and how does that influence our modeling decisions?

Partition these into two groups

Who selected \textit{red vs. blue}?

Who selected \textit{ ○ vs. ▲}?
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Partition these into two groups

Who selected red vs. blue?

Who selected □ vs. △?

Tip: Remember how your own biases/interpretation are influencing your approach

Courtesy Hamed Pirsiavash
Some Terminology

Support

- The valid values a R.V. can take on
- The values over which a pmf/pdf is defined
Some Terminology

Support
- The valid values a R.V. can take on
- The values over which a pmf/pdf is defined

Partition function/normalization function
- The function (or constant) that ensures a pmf/pdf sums to 1
Some Terminology

Support
- The valid values a R.V. can take on
- The values over which a pmf/pdf is defined

Partition function/normalization function
- The function (or constant) that ensures a pmf/pdf sums to 1

Q: What is the support for a Poisson R.V.?
Some Terminology

Support

- The valid values a R.V. can take on
- The values over which a pmf/pdf is defined

Partition function/normalization function

- The function (or constant) that ensures a p{m,d}f sums to 1

Q: What is the support for a Poisson R.V.?

Poisson

- \( X \sim \text{Poisson}(\lambda), \lambda \in \mathbb{R} \) is the “rate”
- \( p(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!} \)

![Graph showing PMF of Poisson distribution for different rates](image)
Some Terminology

Support
- The valid values a R.V. can take on
- The values over which a pmf/pdf is defined

Partition function/normalization function
- The function (or constant) that ensures a pmf/pdf sums to 1

Poisson
- \( X \sim \text{Poisson}(\lambda), \lambda \in \mathbb{R} \) is the “rate”
- \( p(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!} \)

Q: What is the partition function/constant?
Some More Terminology

(Generative) Probabilistic Modeling

Generative Story

Forward probability (ITILA)

Inverse probability (ITILA)
What is (Generative) Probabilistic Modeling?

So far, we’ve (mostly)
   had *labeled* data pairs \((x, y)\), and
   built classifiers \(p(y \mid x)\)
What is (Generative) Probabilistic Modeling?

So far, we’ve (mostly)

had *labeled* data pairs \((x, y)\), and

built classifiers \(p(y \mid x)\)

What if we want to model *both* \(x\) and \(y\) together?

\[ p(x, y) \]
What is (Generative) Probabilistic Modeling?

So far, we’ve (mostly) had *labeled* data pairs \((x, y)\), and built classifiers \(p(y \mid x)\).

What if we want to model *both* \(x\) and \(y\) together?

\[ p(x, y) \]

Q/678 Recap: Where have we used \(p(x, y)\)?
What is (Generative) Probabilistic Modeling?

So far, we’ve (mostly) had *labeled* data pairs \((x, y)\), and built classifiers \(p(y \mid x)\).

What if we want to model *both* \(x\) and \(y\) together?

\[
p(x, y)
\]

**Q/678 Recap: Where have we used \(p(x,y)\)?**

**A: Linear Discriminant Analysis**
What is (Generative) Probabilistic Modeling?

So far, we’ve (mostly) had *labeled* data pairs \((x, y)\), and built classifiers \(p(y \mid x)\).

What if we want to model *both* \(x\) and \(y\) together?

\[ p(x, y) \]

Or what if we only have data but no labels?

\[ p(x) \]

Q: Where have we used \(p(x,y)\)?

A: Linear Discriminant Analysis
Generative Stories

“A useful way to develop probabilistic models is to tell a generative story. This is a *fictional* story that explains how you believe your training data came into existence.” --- CIML Ch 9.5
Generative Stories

“A useful way to develop probabilistic models is to tell a generative story. This is a *fictional* story that explains how you believe your training data came into existence.” --- CIML Ch 9.5

Generative stories are most often used with joint models $p(x, y)$. But despite their name, generative stories are applicable to both generative and conditional models.
p(x, y) vs. p(y | x): Models of our Data

\[ p(x, y) \] is the \textbf{joint} distribution

Two main options for estimating:
1. Directly
2.
p(x, y) vs. p(y | x): Models of our Data

p(x, y) is the **joint** distribution

Two main options for estimating:
1. Directly
2. Using Bayes rule: $p(x, y) = p(x | y)p(y)$

Using Bayes rule *transparently* provides a **generative story** for how our data $x$ and labels $y$ are generated.
p(x, y) vs. p(y | x): Models of our Data

p(x, y) is the **joint** distribution

Two main options for estimating:
1. Directly
2. Using Bayes rule: \( p(x, y) = p(x \mid y)p(y) \)

Using Bayes rule *transparently* provides a **generative story** for how our data x and labels y are generated

p(y | x) is the **conditional** distribution

Two main options for estimating:
1. Directly: used when you *only* care about making the right prediction
   Examples: perceptron, logistic regression, neural networks (we’ve covered)

2.
\( p(x, y) \) vs. \( p(y | x) \): Models of our Data

\( p(x, y) \) is the **joint** distribution

Two main options for estimating:
1. Directly
2. Using Bayes rule: \( p(x, y) = p(x | y)p(y) \)

Using Bayes rule *transparently* provides a **generative story** for how our data \( x \) and labels \( y \) are generated

\( p(y | x) \) is the **conditional** distribution

Two main options for estimating:
1. Directly: used when you *only* care about making the right prediction
   Examples: perceptron, logistic regression, neural networks (we’ve covered)
2. Estimate the joint
Example: Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i) \]
Example: Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i) \]

\( w_1 = 1 \)

\( w_2 = 5 \)

\( w_3 = 4 \)

\( \ldots \)
Generative Story for Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

N different (independent) rolls

\begin{align*}
  w_1 &= 1 \\
  w_2 &= 5 \\
  w_3 &= 4 \\
  \cdots
\end{align*}

Generative Story

for roll \( i = 1 \) to \( N \):
Generative Story for Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i) \]

- \( w_1 = 1 \)
- \( w_2 = 5 \)
- \( w_3 = 4 \)

... 

for roll \( i = 1 \) to \( N \):

\( w_i \sim \text{Cat}(\theta) \)
Generative Story for Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i) \]

N different (independent) rolls

\[ w_1 = 1 \]
\[ w_2 = 5 \]
\[ w_3 = 4 \]
\[ \ldots \]

“for each” loop becomes a product

Generative Story

for roll \( i = 1 \) to \( N \):
\[ w_i \sim \text{Cat}(\theta) \]

Calculate \( p(w_i) \) according to provided distribution
Generative Story for Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

N different (independent) rolls

\( w_1 = 1 \)

\( w_2 = 5 \)

\( w_3 = 4 \)

\( \ldots \)

"for each" loop becomes a product

Generative Story

for roll \( i = 1 \) to \( N \):

\( w_i \sim \text{Cat}(\theta) \)

Calculate \( p(w_i) \) according to provided distribution

\[ \sum_{k=1}^{6} \theta_k = 1 \]

\[ 0 \leq \theta_k \leq 1, \forall k \]
Some More Terminology

(Generative) Probabilistic Modeling

Generative Story

Forward probability (ITILA)

Inverse probability (ITILA)
Forward & Inverse Probabilities

Forward Probability

- Given some data that is “generated” according to some generative model, compute a data-dependent distribution or other quantity
- Involves probabilistic computation for things produced by the story
- Example (ITILA Ex 2.4): Urn problem
  - Urn with B black and W white balls. For N draws with replacement, find distribution over \( n_b \) (the number of times a black ball is drawn)
Forward & Inverse Probabilities

Forward Probability
- Given some data that is “generated” according to some generative model, compute a data-dependent distribution or other quantity
- Involves probabilistic computation for things produced by the story
- Example (ITILA Ex 2.4): Urn problem
  - Urn with B black and W white balls. For N draws with replacement, find distribution over $n_b$ (the number of times a black ball is drawn)

Inverse Probability
- Given some data that is “generated” according to some generative model, compute the conditional (posterior) probability of an unobserved variable in the model
- The typical ML learning/inference problem
Forward & Inverse Probabilities

Forward Probability
- Given some data that is “generated” according to some generative model, compute a data-dependent distribution or other quantity
- Involves probabilistic computation for things produced by the story
- Example (ITILA Ex 2.4): Urn problem
  - Urn with B black and W white balls. For N draws with replacement, find distribution over \( n_b \) (the number of times a black ball is drawn)

Inverse Probability
- Given some data that is “generated” according to some generative model, compute the conditional (posterior) probability of an unobserved variable in the model
- The typical ML learning/inference problem
- Rely of Bayes rule
  - \( p(\text{latent}|\text{obs}) \propto p(\text{obs}|\text{latent})p(\text{latent}) \)
- Example (ITILA Ex 2.6)
  - Multiple urns, each with their own number of black and white balls
  - N balls are drawn, but the selected urn is unobserved/not given
Probability Topics (High-Level)

Basics of Probability: Prereqs

Philosophy of Probability, and Terminology

Useful Quantities and Inequalities
Probabilistic Quantities

• Many quantities involve *expectations*
• Difficulty level varies:
  – Sometimes, they’re easy to compute
  – Sometimes, they look hard to compute but are easy
  – Sometimes, they’re hard to compute
Probabilistic Quantities

• Many quantities involve *expectations*

• Difficulty level varies:
  – Sometimes, they’re easy to compute
  – Sometimes, they look hard to compute but are easy
  – Sometimes, they’re hard to compute

Exponential family formalism helps here (we’ll come back to this later)
Entropy

\[ H(X) = \mathbb{E}_p[-\log p(X)] \]
Entropy

\[ H(X) = \mathbb{E}_p[-\log p(X)] = \sum_x p(x) \log p(x) \]

Marginalize over the support of \( p \)
Entropy

\[ H(X) = \mathbb{E}_p[- \log p(X)] \]

Discrete RV

\[ = \sum_x p(x) \log p(x) \]

Contin. RV

\[ = \int_x dp(x) \log p(x) \]
Entropy

\[ H(X) = \mathbb{E}_p[-\log p(X)] \]

• \( H(X) \geq 0 \)
• By convention, For any \( x \) s.t. \( p(x) = 0 \), \( p(x) \log p(x) = 0 \)
• Sometimes written as \( H(p) \)
• Low entropy \( \rightarrow \) “peaky” distribution
• High entropy \( \rightarrow \) more uniform distribution
Entropy

\[ H(X) = \mathbb{E}_p[-\log p(X)] \]

- \( H(X) \geq 0 \)
- By convention, For any \( x \) s.t. \( p(x) = 0 \), \( p(x) \log p(x) = 0 \)
- Sometimes written as \( H(p) \)
- Low entropy \( \rightarrow \) “peaky” distribution
- High entropy \( \rightarrow \) more uniform distribution

Ex: If \( p \) is a Bernoulli distribution, what is \( H(p) \)?
Joint Entropy

For $X, Y \sim p$:

$$H(X, Y) = \mathbb{E}_p[-\log p(X, Y)]$$

Q: If $X$ & $Y$ are independent, what is $H(X,Y)$?
Joint Entropy

Q: If \( X \) & \( Y \) are independent, what is \( H(X,Y) \)?

A: \( H(X) + H(Y) \)

For \( X, Y \sim p \):

\[
H(X, Y) = \mathbb{E}_p[- \log p(X, Y)]
\]
**Kullback-Leibler (KL) Divergence**

- Measures how “dissimilar” two distributions are

- \( D_{KL}(p||q) \geq 0 \)
  - \( D_{KL} = 0 \) iff \( p = q \)
  - Higher \( D_{KL} \) \( \rightarrow \) more dissimilar

- KL is **not** symmetric
  - \( D_{KL}(p||q) \neq D_{KL}(q||p) \)

\[
D_{KL}(p||q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right]
\]
Kullback-Leibler (KL) Divergence

• Measures how “dissimilar” two distributions are

\[ D_{KL}(p||q) \geq 0 \]
  – \( D_{KL} = 0 \) iff \( p = q \)
  – Higher \( D_{KL} \rightarrow \) more dissimilar

• KL is not symmetric
  – \( D_{KL}(p||q) \neq D_{KL}(q||p) \)

\[ D_{KL}(p||q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right] \]

\[ \text{Discrete RV} = \sum_x p(x) \log \frac{p(x)}{q(x)} \]

\[ \text{Contin. RV} = \int_x d\mu(x) \log \frac{p(x)}{q(x)} \]
Kullback-Leibler (KL) Divergence

• Measures how “dissimilar” two distributions are
  \[ D_{KL}(p||q) \geq 0 \]
  – \( D_{KL} = 0 \) iff \( p = q \)
  – Higher \( D_{KL} \) → more dissimilar
• KL is not symmetric
  \[ D_{KL}(p||q) \neq D_{KL}(q||p) \]

\[
D_{KL}(p||q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right]
\]

**Discrete RV**
\[
= \sum_x p(x) \log \frac{p(x)}{q(x)}
\]

**Contin. RV**
\[
= \int dp(x) \log \frac{p(x)}{q(x)}
\]

Ex 1: \( D_{KL}(p||q) \) if \( p \) & \( q \) are both distributions for rolling a die; one is uniform, one has low entropy
Kullback-Leibler (KL) Divergence

• Measures how “dissimilar” two distributions are

\[ D_{KL}(p||q) \geq 0 \]

– \( D_{KL} = 0 \) iff \( p = q \)
– Higher \( D_{KL} \) \( \rightarrow \) more dissimilar

• KL is **not** symmetric

– \( D_{KL}(p||q) \neq D_{KL}(q||p) \)

\[ D_{KL}(p||q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right] \]

**Discrete RV**
\[ = \sum_x p(x) \log \frac{p(x)}{q(x)} \]

**Contin. RV**
\[ = \int_x d p(x) \log \frac{p(x)}{q(x)} \]

Ex 1: \( D_{KL}(p||q) \) if \( p \ & q \) are both distributions for rolling a die; one is uniform, one has low entropy

Ex 2: \( D_{KL}(p||q) \) if \( p \ & q \) are both Gamma distributions
Outline

Basics of Learning

Probability

Maximum Likelihood Estimation
Learning:
Maximum Likelihood Estimation (MLE)

Core concept in intro statistics:
• Observe some data $\mathcal{X}$
• Compute some distribution $g(\mathcal{X})$ to \{predict, explain, generate\} $\mathcal{X}$
• Assume $g$ is controlled by parameters $\phi$, i.e., $g_{\phi}(\mathcal{X})$
  – Sometimes written $g(\mathcal{X}; \phi)$
• Learning appropriate value(s) of $\phi$ allows you to \textsc{Generalize} about $\mathcal{X}$
Learning: Maximum Likelihood Estimation (MLE)

Core concept in intro statistics:
• Observe some data $\mathcal{X}$
• Compute some distribution $g(\mathcal{X})$ to \{predict, explain, generate\} $\mathcal{X}$
• Assume $g$ is controlled by parameters $\phi$, i.e., $g_\phi(\mathcal{X})$
  – Sometimes written $g(\mathcal{X}; \phi)$
• Learning appropriate value(s) of $\phi$ allows you to \textbf{GENERALIZE} about $\mathcal{X}$

\textit{How do we “learn appropriate value(s) of $\phi$?”}

Many different options: a common one is \textbf{maximum likelihood estimation (MLE)}

• Find values $\phi$ s.t. $g_\phi(\mathcal{X} = \{x_1, ..., x_N\})$ is maximized
• Independence assumptions are very useful here!
• Logarithms are also useful!
Learning:
Maximum Likelihood Estimation (MLE)

Core concept in intro statistics:
• Observe some data \( \mathcal{X} \)
• Compute some distribution \( g(\mathcal{X}) \) to \{predict, explain, generate\} \( \mathcal{X} \)
• Assume \( g \) is controlled by parameters \( \phi \), i.e., \( g_\phi(\mathcal{X}) \)
  - Sometimes written \( g(\mathcal{X}; \phi) \)
• MLE: Find values \( \phi \) s.t.
  \( g_\phi(\mathcal{X} = \{x_1, ..., x_N\}) \) is maximized

Example: How much does it snow?
• \( \mathcal{X} = \{x_1, x_2, ..., x_N\} \) are snowfall values from the previous N storms
• Goal: learn \( \phi \) such that \( g \) correctly models, as accurately as possible, the amount of snow likely
Learning:
Maximum Likelihood Estimation (MLE)

Core concept in intro statistics:
• Observe some data $\mathcal{X}$
• Compute some distribution $g(\mathcal{X})$ to \{predict, explain, generate\} $\mathcal{X}$
• Assume $g$ is controlled by parameters $\phi$, i.e., $g_\phi(\mathcal{X})$
  – Sometimes written $g(\mathcal{X}; \phi)$
• MLE: Find values $\phi$ s.t.
  $g_\phi(\mathcal{X} = \{x_1, \ldots, x_N\})$ is maximized

Example: How much does it snow?
• $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous N storms
• Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely
• Assumption: each $x_i$ is independent from all others

$$\max_{\phi} \sum_{i=1}^{N} \log g_\phi(x_i)$$
MLE Snowfall Example

Example: How much does it snow?

- $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous $N$ storms
- Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely
- Assumption: each $x_i$ is independent from all others

$$\max_{\phi} \sum_{i=1}^{N} \log g_{\phi}(x_i)$$

Q: Why is taking logarithms okay?

Q: What other assumptions, or decisions, do we need to make?
MLE Snowfall Example

Example: How much does it snow?

- $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous $N$ storms
- Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely
- Assumption: each $x_i$ is independent from all others, but all from $g$

$$\max_{\phi} \sum_{i=1}^{N} \log g_{\phi}(x_i)$$

Q: Why is taking logarithms okay?

Q: What other assumptions, or decisions, do we need to make?

$x_i$ is positive, real-valued. What’s a faithful probability distribution for $x_i$?

- Normal? $\times$
- Gamma? $\checkmark$
- Exponential? $\checkmark$
- Bernoulli? $\times$
- Poisson? $\times$
MLE Snowfall Example

Example: How much does it snow?

• $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ are snowfall values from the previous $N$ storms

• Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely

• Assumption: each $x_i$ is independent from all others, but all from $g$

$$\max_{\phi} \sum_{i=1}^N \log g_{\phi}(x_i)$$

Q: Why is taking logarithms okay?

Q: What other assumptions, or decisions, do we need to make?

$x_i$ is positive, real-valued. What’s a faithful probability distribution for $x_i$?

• Normal? $\times$
• Gamma? $\checkmark$ $p(X = x) = \frac{x^{k-1}\exp(-\frac{k}{\theta})}{\theta^k \Gamma(k)}$
• Exponential? $\checkmark$
• Bernoulli? $\times$
• Poisson? $\times$
Example: How much does it snow?

• $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous N storms

• Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely

• Assumption: each $x_i$ is independent from all others, but all from $g$

$$\max_{\phi} \sum_{i=1}^{N} \log g_{\phi}(x_i)$$

Q: Why is taking logarithms okay?

Q: What other assumptions, or decisions, do we need to make?

$x_i$ is positive, real-valued. What’s a faithful/nice-to-compute-and-good-enough probability distribution for $x_i$?

• Normal? $\times$ $\check{\square}$
• Gamma? $\check{\square}$ ?
• Exponential? $\check{\square}$ ?
• Bernoulli? $\times$ $\times$
• Poisson? $\times$ $\times$
MLE Snowfall Example

Example: How much does it snow?

- \( X = \{x_1, x_2, ..., x_N\} \) are snowfall values from the previous \( N \) storms
- Goal: learn \( \phi \) such that \( g \) correctly models, as accurately as possible, the amount of snow likely
- Assumption: each \( x_i \) is independent from all others, but all from \( g \)

\[
\max_{\phi} \sum_{i=1}^{N} \log g_{\phi}(x_i)
\]

\[
x_i \sim \text{Normal}(\mu, \sigma^2)
\]

\[
\max_{(\mu, \sigma^2)} \sum_{i=1}^{N} \log \text{Normal}_{\mu, \sigma^2}(x_i) =
\]
MLE Snowfall Example

Example: How much does it snow?

• \( \mathcal{X} = \{x_1, x_2, \ldots, x_N\} \) are snowfall values from the previous N storms

• Goal: learn \( \phi \) such that \( g \) correctly models, as accurately as possible, the amount of snow likely

• Assumption: each \( x_i \) is independent from all others, but all from \( g \)

\[
\max_{\phi} \sum_{i=1}^{N} \log g_\phi(x_i)
\]

\[
x_i \sim \text{Normal}(\mu, \sigma^2)
\]

\[
\max_{\mu, \sigma^2} \sum_{i=1}^{N} \log \text{Normal}_{\mu, \sigma^2}(x_i) = \max_{\mu, \sigma^2} \left[ \sum_{i=1}^{N} \left[ -\frac{(x_i - \mu)^2}{\sigma^2} \right] \right] - N \log \sigma = F
\]
Example: How much does it snow?

- $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ are snowfall values from the previous N storms
- Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely
- Assumption: each $x_i$ is independent from all others, but all from $g$

$$\max_{\phi} \sum_{i=1}^{N} \log g_{\phi}(x_i)$$

$$x_i \sim \text{Normal}(\mu, \sigma^2)$$

$$\max_{(\mu, \sigma^2)} \sum_{i=1}^{N} \log \text{Normal}_{\mu, \sigma^2}(x_i) = \max_{(\mu, \sigma^2)} \sum_{i=1}^{N} \left[ \frac{-(x_i - \mu)^2}{\sigma^2} \right] - N \log \sigma = F$$

Q: How do we find $\mu, \sigma^2$?
MLE Snowfall Example

Example: How much does it snow?

- $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous $N$ storms
- Goal: learn $\phi$ such that $g$ correctly models, as accurately as possible, the amount of snow likely
- Assumption: each $x_i$ is independent from all others, but all from $g$

$$
\max_{\phi} \sum_{i=1}^{N} \log g_{\phi}(x_i)
$$

$$
\max_{(\mu, \sigma^2)} \sum_{i=1}^{N} \log \text{Normal}_{\mu, \sigma^2}(x_i) =
$$

$$
\max_{(\mu, \sigma^2)} \sum_{i=1}^{N} \left[ -\frac{(x_i - \mu)^2}{\sigma^2} \right] - N \log \sigma = F
$$

Q: How do we find $\mu, \sigma^2$?

A: Differentiate and find that

$$
\hat{\mu} = \frac{\sum_i x_i}{N}
$$

$$
\sigma^2 = \frac{\sum_i (x_i - \hat{\mu})^2}{N}
$$
Learning: Maximum Likelihood Estimation (MLE)

Central to machine learning:

- Observe some data \((X, Y)\)
- Compute some function \(f(X)\) to \{predict, explain, generate\} \(Y\)
- Assume \(f\) is controlled by parameters \(\theta\), i.e., \(f_\theta(X)\)
  - Sometimes written \(f(X; \theta)\)
Learning:
Maximum Likelihood Estimation (MLE)

Central to machine learning:
• Observe some data \((X, Y)\)
• Compute some function \(f(X)\) to \{predict, explain, generate\} \(Y\)
• Assume \(f\) is controlled by parameters \(\theta\), i.e., \(f_\theta(X)\)
  – Sometimes written \(f(X; \theta)\)
• Parameters are learned to minimize error (loss) \(\ell\)

\[
\min_{\theta} \ell(Y^*, f_\theta(X))
\]
Learning: Maximum Likelihood Estimation (MLE)

Central to machine learning:
• Observe some data \((X, Y)\)
• Compute some function \(f(X)\) to \{predict, explain, generate\} \(Y\)
• Assume \(f\) is controlled by parameters \(\theta\), i.e., \(f_\theta(X)\)
  – Sometimes written \(f(X; \theta)\)
• Parameters are learned to minimize error (loss) \(\ell\)

Seen in CMSC 678: linear regression, Naïve Bayes, logistic regression, neural networks, SVMs, PCA, k-means, ...
Learning: Maximum Likelihood Estimation (MLE)

Central to machine learning:

• Observe some data \((X, Y)\)
• Compute some function \(f(X)\) to \{predict, explain, generate\} \(Y\)
• Assume \(f\) is controlled by parameters \(\theta\), i.e., \(f_\theta(X)\)
  – Sometimes written \(f(X; \theta)\)
• Parameters are learned to \textbf{minimize error (loss)} \(\ell\)

Seen in CMSC 678: linear regression, Naïve Bayes, logistic regression, neural networks, SVMs, PCA, k-means, ...

We’ll get back to this in more depth on Wednesday
Learning:
Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

• $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous N storms
• $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$ are closure results from the previous N storms
• Goal: learn $\theta$ such that $f$ correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - $y_{n+1}^*$ from $x_{n+1}$

• If we assume the output of $f$ is a probability distribution on $\mathbf{Y}|\mathbf{X}$...
  ➢ $f(\mathbf{X}) \rightarrow \{p(\text{yes}|\mathbf{X}), p(\text{no}|\mathbf{X})\}$

• Then re: $\theta$, {predicting, explaining, generating} $\mathbf{Y}$ means... *what*?
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

- $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ are snowfall values from the previous N storms
- $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ are closure results from the previous N storms
- Goal: learn $\theta$ such that $f$ correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - $y^*_{n+1}$ from $x_{n+1}$

- If we assume the output of $f$ is a probability distribution on $\mathcal{Y} | \mathcal{X}$...

- Then re: $\theta$, \{predicting, explaining, generating\} $\mathcal{Y}$ means... \textit{what}?
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

- $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous N storms
- $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ are closure results from the previous N storms
- Goal: learn $\theta$ such that $f$ correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - $y_{n+1}^*$ from $x_{n+1}$

- If we assume the output of $f$ is a probability distribution on $\mathcal{Y}|\mathcal{X}$...

- Then re: $\theta$, {predicting, explaining, generating} $\mathcal{Y}$ means finding a value for $\theta$ that maximizes the probability of $\mathcal{Y}$ given $\mathcal{X}$, according to $f$
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

- \( X = \{x_1, x_2, ..., x_N\} \) are snowfall values from the previous N storms
- \( Y = \{y_1, y_2, ..., y_N\} \) are closure results from the previous N storms
- Goal: learn \( \theta \) such that \( f \) correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - \( y^*_{n+1} \) from \( x_{n+1} \)

- If we assume the output of \( f \) is a probability distribution on \( Y | X \)...
- Then re: \( \theta \), \{predicting, explaining, generating\} \( Y \) means finding a value for \( \theta \) that maximizes the probability of \( Y \) given \( X \), according to \( f \):
  \[
  \max_{\theta} f_\theta(x) \rightarrow \max_{\theta} p(Y|X)
  \]
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

• \( \mathcal{X} = \{x_1, x_2, \ldots, x_N\} \) are snowfall values from the previous N storms
• \( \mathcal{Y} = \{y_1, y_2, \ldots, y_N\} \) are closure results from the previous N storms
• Goal: learn \( \theta \) such that \( f \) correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  – \( y^*_{n+1} \) from \( x_{n+1} \)

• If we assume the output of \( f \) is a probability distribution on \( \mathcal{Y}|\mathcal{X} \)...
• Then re: \( \theta \), \{predicting, explaining, generating\} \( \mathcal{Y} \) means finding a value for \( \theta \) that maximizes the probability of \( \mathcal{Y} \) given \( \mathcal{X} \), according to \( f \)
  \[ \max_{\theta} f_{\theta}(x) \rightarrow \max_{\theta} p(Y|X) \]

We’ll get back to this in more depth in next few days
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

- $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous N storms
- $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$ are closure results from the previous N storms
- Goal: learn $\theta$ such that $f$ correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - $y_{n+1}^*$ from $x_{n+1}$

The 678 approach focused most on $\mathbf{Y}$

What if we also care about $\mathbf{X}$?
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

- $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ are snowfall values from the previous N storms
- $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$ are closure results from the previous N storms
- Goal: learn $\theta$ such that $f$ correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - $y^*_{n+1}$ from $x_{n+1}$

- Assume $f$ is a probability distribution on $\mathcal{Y} | \mathcal{X}$
- [Change] Assume there is $g$, a probability distribution on $\mathcal{X}$
Learning: Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

- \( \mathcal{X} = \{x_1, x_2, ..., x_N\} \) are snowfall values from the previous N storms
- \( \mathcal{Y} = \{y_1, y_2, ..., y_N\} \) are closure results from the previous N storms
- Goal: learn \( \theta \) such that \( f \) correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - \( y_{n+1}^* \) from \( x_{n+1} \)

- Assume \( f \) is a *probability distribution* on \( \mathcal{Y} | \mathcal{X} \)
- Assume there is \( g \), a *probability distribution* on \( \mathcal{X} \)
- We also need to learn the distribution \( g \)
Learning:
Maximum Likelihood Estimation (MLE)

Example: Can I sleep in the next time it snows/is school canceled?

• $X = \{x_1, x_2, \ldots, x_N\}$ are snowfall values from the previous N storms
• $Y = \{y_1, y_2, \ldots, y_N\}$ are closure results from the previous N storms
• Goal: learn $\theta$ such that $f$ correctly predicts, as accurately as possible, if UMBC will close in the next storm:
  - $y_{n+1}^*$ from $x_{n+1}$

• Assume $f$ is a probability distribution on $Y | X$
• Assume there is $g$, a probability distribution on $X$
• We also need to learn the distribution $g$

Core design problem: how does $f$ use $g$?

This is task-dependent!
Outline

Basics of Learning

Probability

Maximum Likelihood Estimation
Extended examples of MLE
Learning Parameters for the Die Model

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

maximize (log-) likelihood to learn the probability parameters

Q: Why is maximizing log-likelihood a reasonable thing to do?
Learning Parameters for the Die Model

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

maximize (log-) likelihood to learn the probability parameters

Q: Why is maximizing log-likelihood a reasonable thing to do?
A: Develop a good model for what we observe
Learning Parameters for the Die Model: Maximum Likelihood (Intuition)

$$p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i)$$

maximize (log-) likelihood to learn the probability parameters

If you observe these 9 rolls...

...what are “reasonable” estimates for $p(w)$?

$$p(1) = ? \quad p(2) = ?$$

$$p(3) = ? \quad p(4) = ?$$

$$p(5) = ? \quad p(6) = ?$$
Learning Parameters for the Die Model: Maximum Likelihood (Intuition)

If you observe these 9 rolls...

...what are “reasonable” estimates for $p(w)$?

$p(1) = 2/9 \quad p(2) = 1/9$  
$p(3) = 1/9 \quad p(4) = 3/9$  
$p(5) = 1/9 \quad p(6) = 1/9$

$p(w_1, w_2, ..., w_N) = p(w_1) p(w_2) \cdots p(w_N) = \prod_{i} p(w_i)$

maximize (log-) likelihood to learn the probability parameters
Learning Parameters for the Die Model: Maximum Likelihood (Math)

\[ p(w_1, w_2, \ldots, w_N) = \prod_{i} p(w_i) \]

N different (independent) rolls

\[ \begin{align*}
  w_1 &= 1 \\
  w_2 &= 5 \\
  w_3 &= 4 \\
  \vdots &
\end{align*} \]

Generative Story

for roll \( i = 1 \) to \( N \):

\[ w_i \sim \text{Cat}(\theta) \]

Maximize Log-likelihood

\[ \mathcal{L}(\theta) = \sum_{i} \log p_{\theta}(w_i) \]

\[ = \sum_{i} \log \theta_{w_i} \]
Learning Parameters for the Die Model: Maximum Likelihood (Math)

\[
p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i)
\]

Generative Story

for roll \( i = 1 \) to \( N \):

\( w_i \sim \text{Cat}(\theta) \)

Maximize Log-likelihood

\[ \mathcal{L}(\theta) = \sum_i \log \theta_{w_i} \]

Q: What's an easy way to maximize this, as written \textit{exactly} (even without calculus)?
Learning Parameters for the Die Model: Maximum Likelihood (Math)

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Generative Story
for roll \( i = 1 \) to \( N \):
\( w_i \sim \text{Cat}(\theta) \)

Maximize Log-likelihood
\[ \mathcal{L}(\theta) = \sum_i \log \theta_{w_i} \]

Q: What’s an easy way to maximize this, as written exactly (even without calculus)?

A: Just keep increasing \( \theta_k \) (we know \( \theta \) must be a distribution, but it’s not specified)
Learning Parameters for the Die Model: Maximum Likelihood (Math)

N different (independent) rolls

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Maximize Log-likelihood (with distribution constraints)

\[ \mathcal{L}(\theta) = \sum_i \log \theta_{w_i} \text{ s.t. } \sum_{k=1}^{6} \theta_k = 1 \]

solve using Lagrange multipliers

(we can include the inequality constraints \(0 \leq \theta_k\), but it complicates the problem and, right now, is not needed)
Learning Parameters for the Die Model: Maximum Likelihood (Math)

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Maximize Log-likelihood (with distribution constraints)

\[ F(\theta) = \sum_i \log \theta_{w_i} - \lambda \left( \sum_{k=1}^{6} \theta_k - 1 \right) \]

\[ \frac{\partial F(\theta)}{\partial \theta_k} = \sum_{i:w_i=k} \frac{1}{\theta_{w_i}} - \lambda \quad \frac{\partial F(\theta)}{\partial \lambda} = - \sum_{k=1}^{6} \theta_k + 1 \]

(we can include the inequality constraints \(0 \leq \theta_k\), but it complicates the problem and, right now, is not needed)
Learning Parameters for the Die Model: Maximum Likelihood (Math)

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Maximize Log-likelihood (with distribution constraints)

\[ \mathcal{F}(\theta) = \sum_i \log \theta_{w_i} - \lambda \left( \sum_{k=1}^{6} \theta_k - 1 \right) \]

\[ \theta_k = \frac{\sum_{i:w_i=k} 1}{\lambda} \quad \text{optimal } \lambda \text{ when } \sum_{k=1}^{6} \theta_k = 1 \]
Learning Parameters for the Die Model: Maximum Likelihood (Math)

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

Maximize Log-likelihood (with distribution constraints)

\[ F(\theta) = \sum_i \log \theta_{w_i} - \lambda \left( \sum_{k=1}^{6} \theta_k - 1 \right) \]

\[ \theta_k = \frac{\sum_{i:w_i=k} 1}{\sum_k \sum_{i:w_i=k} 1} = \frac{N_k}{N} \]

Optimal \( \lambda \) when \( \sum_{k=1}^{6} \theta_k = 1 \)

(N different (independent) rolls)

(we can include the inequality constraints \( 0 \leq \theta_k \), but it complicates the problem and, right now, is not needed)
Example: Conditionally Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N) \]

\[ = \prod_i p(w_i|z_i)p(z_i) \]

*add complexity to better explain what we see*
Example: Conditionally Rolling a Die

\[ p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N) \]

\[ = \prod_i p(w_i|z_i) p(z_i) \]

First flip a coin...

\[ z_1 = T \]

\[ z_2 = H \]

...
Example: Conditionally Rolling a Die

\[ p(w_1, w_2, \ldots, w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_i p(w_i) \]

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N) \]

\[ = \prod_i p(w_i|z_i) p(z_i) \]

First flip a coin...

...then roll a different die depending on the coin flip

\[ z_1 = T \quad w_1 = 1 \]

\[ z_2 = H \quad w_2 = 5 \]

...
Learning in Conditional Die Roll Model: Maximize (Log-)Likelihood

\[
p(w_1, w_2, ..., w_N) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{i} p(w_i)
\]

Add complexity to better explain what we see.

\[
p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = p(z_1)p(w_1|z_1) \cdots p(z_N)p(w_N|z_N)
\]
\[
= \prod_{i} p(w_i|z_i) p(z_i)
\]

If you observe the \(z_i\) values, this is easy!
Learning in Conditional Die Roll Model: Maximize (Log-)Likelihood

\[ p(z_1, w_1, z_2, w_2, ..., z_N, w_N) = \prod_i p(w_i | z_i) p(z_i) \]

If you observe the \( z_i \) values, this is easy!

First: Write the Generative Story

\( \lambda = \) distribution over coin (z)

\( \gamma^{(H)} = \) distribution for die when coin comes up heads

\( \gamma^{(T)} = \) distribution for die when coin comes up tails

for item \( i = 1 \) to \( N \):

\[ z_i \sim \text{Bernoulli}(\lambda) \]

\[ w_i \sim \text{Cat}(\gamma^{(z_i)}) \]
Learning in Conditional Die Roll Model: Maximize (Log-)Likelihood

\[
p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = \prod_{i} p(w_i|z_i) p(z_i)
\]

If you observe the \( z_i \) values, this is easy!

First: Write the Generative Story

\begin{align*}
\lambda &= \text{distribution over coin (z)} \\
\gamma^{(H)} &= \text{distribution for H die} \\
\gamma^{(T)} &= \text{distribution for T die} \\
\text{for item } i = 1 \text{ to } N: \\
z_i &\sim \text{Bernoulli}(\lambda) \\
w_i &\sim \text{Cat}(\gamma^{(z_i)})
\end{align*}

Second: Generative Story \rightarrow Objective

\[
\mathcal{F}(\theta) = \sum_{i=1}^{n} (\log \lambda_{z_i} + \log \gamma_{w_i}^{(z_i)})
\]

Lagrange multiplier constraints
Learning in Conditional Die Roll Model: Maximize (Log-)Likelihood

\[
p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = \prod_{i} p(w_i|z_i) p(z_i)
\]

If you observe the \(Z_i\) values, this is easy!

First: Write the Generative Story

\[\lambda = \text{distribution over coin (z)}\]
\[\gamma^{(H)} = \text{distribution for H die}\]
\[\gamma^{(T)} = \text{distribution for T die}\]

for item \(i = 1\) to \(N\):

\[z_i \sim \text{Bernoulli}(\lambda)\]
\[w_i \sim \text{Cat}(\gamma^{(z_i)})\]

Second: Generative Story \(\rightarrow\) Objective

\[\mathcal{F}(\theta) = \sum_{i}^{n} (\log \lambda_{z_i} + \log \gamma^{(z_i)}_{w_i})\]
\[-\eta \left( \sum_{k=1}^{2} \lambda_k - 1 \right) - \sum_{k=1}^{2} \delta_k \left( \sum_{j=1}^{6} \gamma^{(k)}_j - 1 \right)\]
Learning in Conditional Die Roll Model: Maximize (Log-)Likelihood

\[ p(z_1, w_1, z_2, w_2, \ldots, z_N, w_N) = \prod_i p(w_i | z_i) p(z_i) \]

If you observe the \( z_i \) values, this is easy!

But if you don’t observe the \( z_i \) values, this is not easy!

First: Write the Generative Story

\[ \lambda = \text{distribution over coin (z)} \]
\[ \gamma^{(H)} = \text{distribution for H die} \]
\[ \gamma^{(T)} = \text{distribution for T die} \]
for item \( i = 1 \) to \( N \):
\[ z_i \sim \text{Bernoulli}(\lambda) \]
\[ w_i \sim \text{Cat}(\gamma^{(z_i)}) \]

Second: Generative Story → Objective

\[ \mathcal{F}(\theta) = \sum_i^n (\log \lambda_{z_i} + \log \gamma^{(z_i)}_{w_i}) \]
\[ -\eta \left( \sum_{k=1}^2 \lambda_k - 1 \right) - \sum_{k=1}^2 \delta_k \left( \sum_{j=1}^6 \gamma^{(k)}_j - 1 \right) \]