Information Theory and Probabilistic Programming

Samuel Cheng

School of ECE
University of Oklahoma

August 20, 2019

About this course

1. Learn some basic information theory (what is it? how is it useful?)
   - Understand basic terminology: what is entropy all about?

About this course

1. Learn some basic information theory (what is it? how is it useful?)
   - Understand basic terminology: what is entropy all about?
2. Statistical inference
   - Bayesian and Monte Carlo techniques

About this course

1. Learn some basic information theory (what is it? how is it useful?)
    - Understand basic terminology: what is entropy all about?
2. Statistical inference
    - Bayesian and Monte Carlo techniques
3. Introduction of probabilistic programming
    - Solve inference problems with programming

What is information theory?

- Study of "information" using probability

What is information theory?

- Study of "information" using probability
- Can be treated as a subfield of applied probability

What is information theory?

- Study of "information" using probability
- Can be treated as a subfield of applied probability
- But it has a huge impact to communications and information science

What is information theory?

- Study of "information" using probability
- Can be treated as a subfield of applied probability
- But it has a huge impact to communications and information science
    - The theoretical basis of the entire telecom industry is built on top of that

What is information theory?

- Study of "information" using probability
- Can be treated as a subfield of applied probability
- But it has a huge impact to communications and information science
  - The theoretical basis of the entire telecom industry is built on top of that
  - Study of extreme cases. What is possible and what is not?

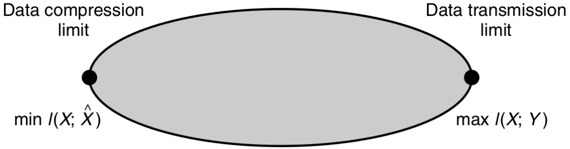

FIGURE 1.2. Information theory as the extreme points of communication theory.

(From Cover and Thomas)

Connection to other fields

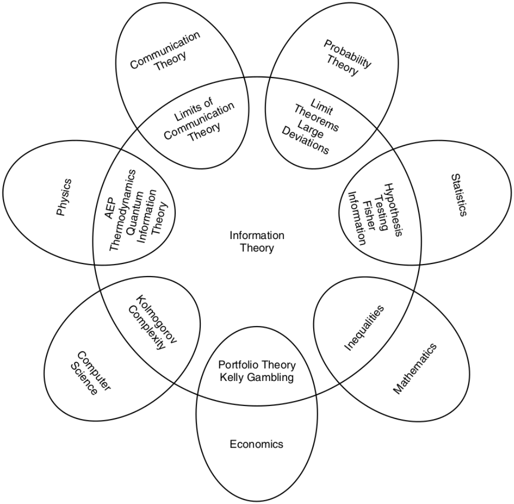

FIGURE 1.1. Relationship of information theory to other fields.

(From Cover and Thomas)

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
    - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly
  - Argue that there is a (capacity) limit of lossless communication under a noisy channel and theoretically we can have lossless communications as long as smaller than the capacity

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
    - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
        - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
        - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly
    - Argue that there is a (capacity) limit of lossless communication under a noisy channel and theoretically we can have lossless communications as long as smaller than the capacity
        - Give the capacity of Gaussian channel as an example

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly
  - Argue that there is a (capacity) limit of lossless communication under a noisy channel and theoretically we can have lossless communications as long as smaller than the capacity
    - Give the capacity of Gaussian channel as an example
- Some similar ideas were explored earlier in Bell Labs by Harry Nyquist and Ralph Hartley. But those results are limited to events with equal probability

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome

What is “information” in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the “knowledge” gained you have knowing that piece of information

What is “information” in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the “knowledge” gained you have knowing that piece of information
  - More information if the outcomes of the event are less predictable

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the "knowledge" gained you have knowing that piece of information
  - More information if the outcomes of the event are less predictable
  - Entropy is a measure of **uncertainty**

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the "knowledge" gained you have knowing that piece of information
  - More information if the outcomes of the event are less predictable
  - Entropy is a measure of **uncertainty**
- A Preview:

$$H(X) = \sum_x p(x) \underbrace{H(X=x)}_{\text{info revealed when } X=x}$$

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the "knowledge" gained you have knowing that piece of information
    - More information if the outcomes of the event are less predictable
    - Entropy is a measure of **uncertainty**
- A Preview:

$$H(X) = \sum_x p(x) \underbrace{H(X=x)}_{\text{info revealed when } X=x}$$

A good guess for $H(X=x)$: $\log \frac{1}{p(x)}$

Computer scientists' treatment

- Kolmogorov complexity (algorithm information theory): quantify a piece of information as the size of smallest program describing it

Computer scientists' treatment

- Kolmogorov complexity (algorithm information theory): quantify a piece of information as the size of smallest program describing it
- Nice philosophically but doesn't go much anywhere

Computer scientists' treatment

- Kolmogorov complexity (algorithm information theory): quantify a piece of information as the size of smallest program describing it
- Nice philosophically but doesn't go much anywhere
- We will take the probabilistic view (electrical/communication engineers treatment here) to quantify information theory who usually study with Bayesian models

Neumann-Shannon Anecdote

When Shannon discovered this function he was faced with the need to name it, for it occurred quite often in the theory of communication he was developing. He considered naming it “information” but felt that this word had unfortunate popular interpretations that would interfere with his intended uses of it in the new theory. He was inclined towards naming it “uncertainty” and discussed the matter with the late John Von Neumann. Von Neumann suggested that the function ought to be called “entropy” since it was already in use in some treatises on statistical thermodynamics (e.g. ref. 12). Von Neumann, Shannon reports, suggested that there were two good reasons for calling the function “entropy”. ”It is already in use under that name,” he is reported to have said, ”and besides, it will give you a great edge in debates because nobody really knows what entropy is anyway.” Shannon called the function “entropy” and used it as a measure of “uncertainty,” interchanging the two words in his writings without discrimination.

–From wikipedia

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x, y) = p(y)$

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x, y) = p(y)$
- Conditional probability: $p(x|y) = \frac{p(x,y)}{p(y)}$
  - N.B. $\sum_x p(x|y) = 1$ but $\sum_y p(x|y) \neq 1$

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x, y) = p(y)$
- Conditional probability: $p(x|y) = \frac{p(x,y)}{p(y)}$
  - N.B. $\sum_x p(x|y) = 1$ but $\sum_y p(x|y) \neq 1$
- Chain rule: $p(x, y, z) = p(x)p(y|x)p(z|x, y)$

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x,y) = p(y)$
- Conditional probability: $p(x|y) = \frac{p(x,y)}{p(y)}$
  - N.B. $\sum_x p(x|y) = 1$ but $\sum_y p(x|y) \neq 1$
- Chain rule: $p(x,y,z) = p(x)p(y|x)p(z|x,y)$
- Independence: $p(x,y) = p(x)p(y)$, $X \perp\!\!\!\perp Y$

Some probability basic

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x, y) = p(y)$
- Conditional probability: $p(x|y) = \frac{p(x,y)}{p(y)}$
  - N.B. $\sum_x p(x|y) = 1$ but $\sum_y p(x|y) \neq 1$
- Chain rule: $p(x, y, z) = p(x)p(y|x)p(z|x, y)$
- Independence: $p(x, y) = p(x)p(y)$, $X \perp\!\!\!\perp Y$
- Markov property and conditional independence: $p(x, y|z) = p(x|z)p(y|z)$, $X \perp\!\!\!\perp Y|Z, X \leftrightarrow Z \leftrightarrow Y$

Independence but not conditional independence

Consider flipping two coins with outcomes store as X and Y, say 1 represents a head and 0 represents a tail

- In general the two outcomes should be independent (maybe unless if you are some professional/magical gambler), so we have $X \perp\!\!\!\perp Y$
- Now, let $Z = X \oplus Y$, where $\oplus$ is the exclusive or operation $(1 \oplus 0 = 0 \oplus 1 = 1$ and $1 \oplus 1 = 0 \oplus 0 = 0)$
  - Even though $X \perp\!\!\!\perp Y$, $X \not\perp\!\!\!\perp Y|Z$
  - Actually given Z, X “depends” very much on Y since from $X = Y \oplus Z$, we can find out X precisely given Y
  - We can also check the condition $X \perp\!\!\!\perp Y|Z$ by comparing the probability $p(x|z,y)$ with $p(x|z)$
    - For example, $p_{X|Z}(0|0) = 0.5 \neq 1 = p_{X|Z,Y}(0|0,0)$. Thus $X \perp\!\!\!\perp Y|Z$ cannot be true

More formal treatment: probability space

- More rigorously, a probability model is defined by the **probability space** composed of the triple $(\Omega, \mathcal{F}, p)$
  - Ω is the **sample space** containing all possible outcomes
  - $\mathcal{F}$ is a "σ-field", which is a collection of subsets (events) of Ω
  - p is the (non-negative) **probability measure** on elements of $\mathcal{F}$
- E.g., probability model of unbiased dice
  - $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1, 2\}, \{1, 3\}, \cdots, \{1, 2, 3, 4, 5, 6\}\}$
  - $p(S)$ is the probability of an event
    - $p(\{1\}) = p(\{2\}) = p(\{3\}) = p(\{4\}) = p(\{5\}) = p(\{6\}) = 1/6$
    - $p(\{1, 2\}) = p(\{1, 3\}) = \cdots = p(\{5, 6\}) = 2/6$
    - $\cdots$
    - $p(\{1, 2, 3, 4, 5, 6\}) = 1$
- N.B. It could be confusing at first. Be careful that events $\neq$ outcomes. **An event is actually a set of outcomes**

σ-algebra

- The purpose of σ-field (aka σ-algebra) is to impose restriction on what we can and cannot query regarding probability
- Namely, we can only measure the probability of something inside the σ-field $\mathcal{F}$ (i.e., an event)
- Formal definition of σ-field:
    - **σ-field has to satisfied the following: 1) containing empty set $\varnothing$, 2) closed under complement, countable union, and countable intersection of its element**
- E.g., let $\Omega = \{1, 2, 3, 4\}$
    1. $\{\varnothing, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ is a valid σ-field
    2. $\{\varnothing, \{1\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ is NOT a valid σ-field
- N.B., A complement, countable union, or countable intersection of Ω is call a **Borel set**
    - $\varnothing, \{1\}, \{1, 2\}$ are example of Borel sets (an event is a Borel set)
    - Collection of all Borel sets forms a σ-algebra (aka Borel (σ-)algebra)

Probability measure

- Probability measure p is a **measure**. Along with $\mathcal{F}$, the tuple $(\mathcal{F}, p)$ forms a **measure space**. For $\mathbb{P}$ to be a valid probability measure, it has to satisfy the following
  - Requirements to be a measure (in the context of measure theory):
    1. $p(\varnothing) = 0$
    2. Countably additive: $p(\cup_{i\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} p(A_i)$
  - And since p is a probability measure, it also has to satisfy $p(\Omega) = 1$
- The above constraints are sometimes known as the axioms of probability theory

Some properties of probability measure

From the axioms described in the last slides, one can show that probability measure has to satisfies the following:

1. $p(A^c) = 1 - p(A)$
2. $p(A) \leq p(B)$ if $A \subset B$
3. Union bound: $p(\cup_i A_i) \leq \sum_i p(A_i)$
    - Proof hint: use 2) and induction
4. Inclusion-exclusion formula: $p(\cup_{i=1}^n A_i) = \sum_{i=1}^n p(A_i) - \sum_{i<j} p(A_i \cap A_j) + \sum_{i<j<k} p(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} p(\cap_{i=1}^n A_i)$
    - Proof hint: show $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ and then use induction. ($p(A \cup B) = p(A) + p(B \setminus A)$ and $p(B) = p(A \cap B) + p(B \setminus A)$).

Why so complex?

- Consider X a uniform random variable defined between $[0, 1]$
- Define $Y = \begin{cases} 1 & \text{if } X \text{ is rational} \\ 0 & \text{otherwise} \end{cases}$
- Y is a random variable since X is random. It is reasonable to ask what is the probability that $Y = 1$. From undergrad probability class,

$$Pr(Y = 1) = \int_{\{x|x\in[0,1]\cap\mathbb{Q}\}} dx = ?$$

  - The integral above is actually undefined according to undergrad calculus, where the integral is known as a Riemann integral
- Instead, we have to incorporate the idea of "measure" (Lesbeque integral)

$$Pr(Y = 1) = \int_{\{x|x\in[0,1]\cap\mathbb{Q}\}} dp(x) = 0$$

  - The Lesbeque integral above is 0 since the measure of $\{x|x \in [0, 1] \cap \mathbb{Q}\} = 0$

Some remarks on notation

- In general, we can write

$$p(\Omega') = \int_{\Omega'} dp(\omega)$$

and

$$E[f(X)] = \int_{\Omega} f(X(\omega))dp(\omega)$$

- E.g.,

$$E[X] = \int_{\Omega} X(\omega)dp(\omega) = \int_{\Omega} X(\omega)\ dp = \int_{\Omega} Xdp$$

  - Note that p is the probability measure (often people use upper case P instead)
  - People often omit ω as above when context is clear

Bayes' rule (with model type)

- $p(\theta, o) = p(o)p(\theta|o) = p(\theta)p(o|\theta)$

Bayes' rule (with model type)

- $p(\theta, o) = p(o)p(\theta|o) = p(\theta)p(o|\theta)$
- Let's add model type M,
  $p(\theta, o|M) = p(o|M)p(\theta|o, M) = p(\theta|M)p(o|\theta, M)$

Bayes' rule (with model type)

- $p(\theta, o) = p(o)p(\theta|o) = p(\theta)p(o|\theta)$
- Let's add model type M,
  $p(\theta, o|M) = p(o|M)p(\theta|o, M) = p(\theta|M)p(o|\theta, M)$

$$\underbrace{p(\theta|o, M)}_{posterior} = \frac{\overbrace{p(\theta|M)}^{prior}\overbrace{p(o|\theta, M)}^{likelihood}}{\underbrace{p(o|M)}_{model\ \ evidence}}$$

- M: model type
- θ: model parameter
- o: observation

Inference

o: Observed variable, θ: Parameter, x: Latent variable

Maximum Likelihood (ML)

$\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(o|\theta)$

Maximum A Posteriori (MAP)

$\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(\theta|o)$

Bayesian

$$\hat{x} = \sum_x x \underbrace{\sum_\theta p(x|\theta)p(\theta|o)}_{p(x|o)}$$

where $p(\theta|o) = \frac{p(o|\theta)p(\theta)}{p(o)} \propto p(o|\theta)\underbrace{p(\theta)}_{prior}$

Coin Flip

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Prior: Probability of a hypothesis before we make any observations

(Slide credit: University of Washington CSE473)

Coin Flip

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Uniform Prior: All hypothesis are equally likely before we make any observations

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} \qquad P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.6$

Posterior: Probability of a hypothesis given data

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

(Slide credit: University of Washington CSE473)

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin:

C_2 

Best estimate for P(H)

$P(H|C_2) = 0.5$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

(Slide credit: University of Washington CSE473)

Your Estimate?

Maximum Likelihood Estimate: The best hypothesis that fits observed data assuming uniform prior

Most likely coin:

C_2 

Best estimate for P(H)

$P(H|C_2) = 0.5$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

(Slide credit: University of Washington CSE473)

Using Prior Knowledge

- Should we always use Uniform Prior?
- Background knowledge:
  - Heads => you go first in Abalone against TA
  - TAs are nice people
  - => TA is more likely to use a coin biased in your favor

C_1

C_2

C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Using Prior Knowledge

We can encode it in the prior:

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

C_1

C_2

C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H)$ = ? $P(C_2|H)$ = ? $P(C_3|H)$ = ?

$$P(C_1|H) = \alpha P(H|C_1)P(C_1)$$

C_1

C_2

C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.006$ $P(C_2|H) = 0.165$ $P(C_3|H) = 0.829$

ML posterior after Exp 1:

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.600$

C_1	C_2	C_3
		
$P(H|C_1) = 0.1$	$P(H|C_2) = 0.5$	$P(H|C_3) = 0.9$
$P(C_1) = 0.05$	$P(C_2) = 0.25$	$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 0.05$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 0.25$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 0.05$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 0.25$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin:

C_3 

Best estimate for P(H)

$P(H|C_3) = 0.9$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Your Estimate?

Maximum A Posteriori (MAP) Estimate: The best hypothesis that fits observed data assuming a non-uniform prior

Most likely coin:

C_3 

Best estimate for P(H)

$P(H|C_3) = 0.9$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Did We Do The Right Thing?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Did We Do The Right Thing?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_2 and C_3 are almost equally likely

C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

A Better Estimate

Recall: $P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$ = 0.680

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error,
given data and (generally) assuming a non-uniform prior

$$P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i) = 0.680$$

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Comparison

ML
- Easy to compute

Comparison

ML	• Easy to compute
MAP	• Still relatively easy to compute • Incorporate prior information

Comparison

ML	• Easy to compute
MAP	• Still relatively easy to compute • Incorporate prior information
Bayesian	• Minimizes expected error $\Rightarrow$ especially shines when little data available • Potentially much harder to compute