# Probability for Computer Science

Spring 2021

Lecture 23

Prof. Claire Monteleoni

# Today

Introduction to Information Theory

- Huffman Coding

Statistical Inference

- Hypothesis Testing
  - Maximum Likelihood
  - Maximum A Posteriori (MAP)
- Parameter Estimation
  - Maximum Likelihood
  - Maximum A Posteriori (MAP)

# Optimal codes

Fixed-length codes are optimal when all symbols occur with equal probability.

When the symbols have different probabilities, the optimal code will be a variable-length code.

- but not *any* variable length code; some achieve worse information rates than others.

# Huffman Code: Information Rate bound

The information rate of the Huffman code is upper bounded as follows:

$$R(A_1, \ldots, A_n) \leq H(A_1, \ldots, A_n) + 1$$

This is optimal for prefix codes.

And remember, for any code which uniquely encodes each symbol,

$$H(A_1, \ldots, A_n) \leq R(A_1, \ldots, A_n)$$

# Huffman algorithm

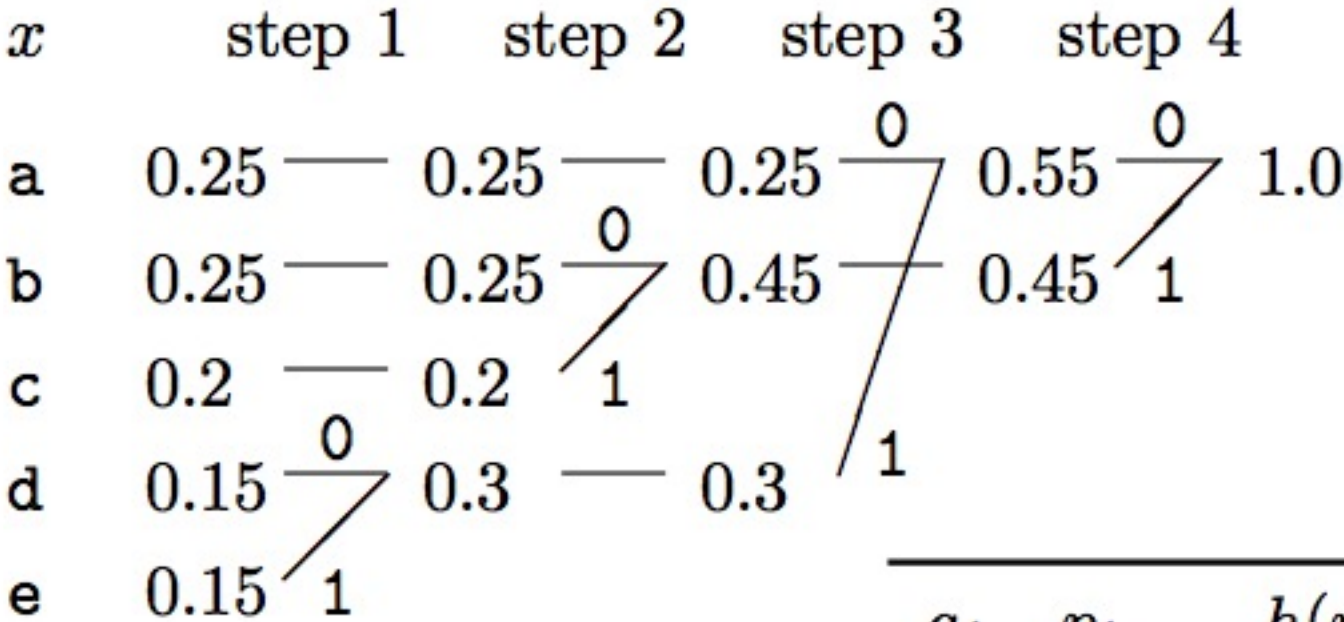
1. Take the two least probable symbols in the alphabet. These two symbols will be given the longest codewords, which will have equal length, and differ only in the last digit.

2. Combine these two symbols into a single symbol, and repeat.

At the end, codewords are read from root to leaf.

Credit: D. MacKay 2003

# Example 1




| $a_i$ | $p_i$ | $h(p_i)$ | $l_i$ | $c(a_i)$ |
|---|---|---|---|---|
| a | 0.25 | 2.0 | 2 | 00 |
| b | 0.25 | 2.0 | 2 | 10 |
| c | 0.2 | 2.3 | 2 | 11 |
| d | 0.15 | 2.7 | 3 | 010 |
| e | 0.15 | 2.7 | 3 | 011 |

Now: Exercise 4

Credit: D. MacKay 2003

# Post-class

## Post-class exercises:

- Propose a symbol distribution and then design a Huffman code for it. Compute its information rate.
- Optional: Compare with info. rate of code constructed top-down (Starting from all events, recurse, making equiprobable splits).

# Statistical Inference

In Probability Theory, a probabilistic model is defined for an experiment, and the relevant probabilities are fully specified.

- Any question about the outcome of the experiment has a unique right answer.
  - E.g. what is the probability of 2 heads in 2 fair coin flips?

In Statistical Inference, we are only given observations.

- There may not be a single “right” answer.
  - E.g. given access to a collection of old emails, how likely is it that some new email is spam?

# Types of Inference

- Hypothesis Testing: Given some data, determine which out of a set of hypotheses is more likely to be true.
  - Given the words/characters in an email, determine whether it is more likely to be spam or not.
    - E.g. H1 = “This message is spam”, H2 = “This message is not spam”
  - Given a student’s test score, determine whether s/he studied or not.
    - E.g. H1 = “This student studied for the exam”, H2 = “This student did not study for the exam”

- Parameter Estimation: Have a model that is fully specified except some unknown parameters we need to estimate.
  - Estimate the heads probability of a coin, from repeated flips.
  - Estimate the fraction of the population who prefers candidate A to candidate B, from polling data.

# Hypothesis Testing

- Let D be the event that we observe some particular data.
  - E.g., let D be the event that an email contains the strings `‘ca$h’` and `‘!!!!!!!!!!’`
- Let $H_1$, …, $H_n$ be a set of events that partition the sample space. We call these hypotheses.
  - E.g., H1 = “This message is spam”

    H2 = “This message is not spam”
- How do we use D to decide which hypothesis, $H_i$, is most likely? This problem is called hypothesis testing.

# Maximum Likelihood

Suppose we know (or can compute) the probability $P(D \mid H_i)$ of observing data, D, given each hypothesis $H_i$.

The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely.

$$\boxed{H^{\mathrm{ML}} = \arg\max_i P(D|H_i)}$$

# Example 1

There are 2 boxes of cookies:

- Box 1 contains half chocolate chip cookies and half oatmeal raisin cookies.
- Box 2 contains 1/3 chocolate chip cookies and 2/3 oatmeal raisin cookies

- I select a box and choose a random cookie from the box.
- If you only observe the cookie I chose (not my box selection), and it is chocolate chip, which box is it most likely that I chose from?

P(chocolate chip | Box 1) = ½

P(chocolate chip | Box 2) = 1/3

So Box 1 is the Maximum Likelihood Hypothesis.

# Example 2

- If I drive to work, there's a 60% chance that I'll be late.
- If I bike to work, there's a 20% chance that I'll be late.

If I tell you that I was late to work, then what is the maximum likelihood hypothesis for how I got to work?

$P(\text{Late} \mid \text{Drove}) = 0.6$

$P(\text{Late} \mid \text{Biked}) = 0.2$

So the maximum likelihood hypothesis is that I drove to work.

# The problem with Maximum Likelihood

In Example 2, suppose I also tell you that I only drive to work 5% of the time.

- Does it still seem likely that I drove to work today?
- No!

→ We need to be able to incorporate this kind of information into our reasoning.

# Bayesian Reasoning

If we know $P(H_i)$ and $P(D \mid H_i)$ for each $H_i$, then we can use Bayes Rule to compute $P(H_i \mid D)$ for each hypothesis .

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)} = \frac{P(D|H_i)P(H_i)}{\sum_j P(D|H_j)P(H_j)}$$

- $P(H_i)$ is called the prior probability of $H_i$.
- $P(H_i \mid D)$ is called the posterior probability of $H_i$.

The posterior probability is a refinement of our prior belief about $H_i$, in light of the observed data, D.

# Maximum A Posteriori (MAP)

Suppose we know (or can compute) the probability $P(D \mid H_i)$ of observing data, D, given each hypothesis $H_i$, as well as the prior probability of each hypothesis $P(H_i)$.

The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability.

$$\boxed{H^{\mathrm{MAP}} = \arg\max_i P(D|H_i)P(H_i)}$$

# MAP vs. MLE

- If you the priors, $P(H_i)$, then MAP will always give you a better estimate.
- Use MLE if you don't know the priors $P(H_i)$
  - [And of course when the question requests it]

- When do they give the same answer?
  - When all hypotheses are equally likely. I.e., if there are k hypotheses, s.t. $P(H_i) = 1/k$ for each, then:

$$H^{MAP} = \arg\max_i P(D|H_i)P(H_i) = \arg\max_i \frac{1}{k}P(D|H_i) = \arg\max_i P(D|H_i) = H^{ML}$$

# Example 2, revisited

- If I drive to work, there's a 60% chance that I'll be late.
- If I bike to work, there's a 20% chance that I'll be late.
- I drive 5% of the days (and bike the rest of the days)

If I tell you that I was late to work, then what is the MAP hypothesis for how I got to work?

P(Late | Drove)P(Drove)= (0.6)(0.05)=0.03

P(Late | Biked)P(Biked) = (0.2)(0.95)=0.19

So the MAP hypothesis is that I biked to work.

# Example 1, revisited

There are 2 boxes of cookies:

- Box 1 contains half chocolate chip cookies and half oatmeal raisin cookies.
- Box 2 contains 1/3 chocolate chip cookies and 2/3 oatmeal raisin cookies
- You know that I pick box 2 with probability 90%

- I select a box and choose a random cookie from the box.
- If you only observe the cookie I chose (not my box selection), and it is chocolate chip, which box is it most likely that I chose from?

A: Choosing the random cookie means choosing a cookie uniformly at random from the box.

$P(\text{chocolate chip} \mid \text{Box 1})\, P(\text{Box 1}) = (\tfrac{1}{2})(1/10) = 1/20$

$P(\text{chocolate chip} \mid \text{Box 2})\, P(\text{Box 2}) = (1/3)(9/10) = 9/30 = 3/10$

So Box 2 is the MAP hypothesis.

# Parameter Estimation

We have previously introduced Maximum Likelihood and MAP for Hypothesis Testing

Now we'll study Parameter Estimation and look at Maximum Likelihood and MAP techniques for that problem.

# Parameter Estimation

- Suppose we would like to estimate the unknown bias, p, of a coin, based on observations of the outcomes of n independent tosses of the coin
- Or suppose we want to estimate the approval rating of the President, by randomly polling n people and asking if they approve or disapprove.
- We can define analogs of both Maximum Likelihood and MAP for this problem of parameter estimation

# Parameter Estimation

Consider n observations, $X_i$, of outcomes of a random variable that is parameterized by some unknown θ.

- Suppose the observations are: $X_1=k_1$, $X_2=k_2$, …, $X_n=k_n$.

The maximum likelihood (ML) estimate is the parameter value that makes the data most likely.

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n \boxed{; \theta)}$$

"Parameterized by θ"

θ: the parameters of the probabilistic model. We don't assume θ to be random.

# Parameter Estimation

The maximum likelihood (ML) estimate is the parameter value that makes the data most likely.

$$\hat{\theta} = \arg\max_{\theta} \boxed{P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)}$$

If the $X_i$ are independent observations, then:

$$\hat{\theta} = \arg\max_{\theta} \boxed{\prod_{i=1}^{n} P(X_i = k_i; \theta)}$$

"likelihood"

# Parameter Estimation

If the $X_i$ are independent observations, then:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

$$= \arg\max_{\theta} \log \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

$$= \arg\max_{\theta} \boxed{\sum_{i=1}^{n} \log P(X_i = k_i; \theta)}$$

"log-likelihood"

# Maximum Likelihood is Consistent

Consistency: if θ is the true value of the parameter, and $\theta_n$ is the maximum likelihood estimate after n observations, then for any ε > 0,

$$\lim_{n\to\infty} P(|\theta_n - \theta| \geq \epsilon) = 0$$

In other words: as the number of observations grows large, the maximum likelihood estimate gets closer and closer to the true parameter value, as desired.

# Exercise

What is the Maximum Likelihood Estimate (MLE) of the unknown bias, *p*, of a coin, based on observations of the outcomes of *n* independent tosses of the coin, $X_1, \ldots, X_n$?

# MAP Parameter Estimation

What if we do want to assume θ is a random variable(s)?

If there is a prior probability distribution over values of the parameter θ, one can derive a maximum a posteriori (MAP) estimate for the parameter.

Useful in cases where n is small.