Assignment 3 + Compression

(SNLP Tutorial 4)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

18th, 20th May 2021

Assignment 3

- Exercise 1: Entropy Intuition
- Exercise 2: Uncertainty of events
- Exercise 3: KL Divergence
- Bonus: KL Divergence calculation

Compression

- Prefix Codes: No whole code word is a prefix of any other code word
- Uniquely decodable codes: Each word maps to one and only one code word

Compression

- Prefix Codes: No whole code word is a prefix of any other code word
- Uniquely decodable codes: Each word maps to one and only one code word

Prefix codes are a subset of uniquely decodable codes!

Optimal length of code words

$$l_i = -\log_D p(w_i)$$

Kraft's Inequality

$$\sum_{i=1}^{m} D^{-l_i} \leq 1$$

What does the sum < 1 imply?
What does the sum $= 1$ imply?
What does the sum > 1 imply?
What does this tell us about uniquely decodable and prefix codes?

Exercise: Test Kraft's Inequality on Morse Code

(Hint: What is the encoding alphabet?)

Encoding

Task

Create encoding (binary) for the following recipe:

```
apple apple banana cherries apple dark_chocolate eggplant banana cherries banana ...
```

Encoding

Task

Create encoding (binary) for the following recipe:

`apple apple banana cherries apple dark_chocolate eggplant banana cherries banana ...`

Fixed-width encoding

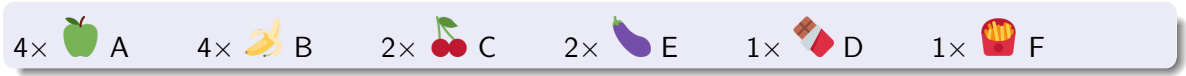
 000  001  010  011  100  101

Length $= 14 \times 3 = 42$

Issues?

- Encoding for  and  ?
- What do 110 and 111 mean?

Encoding - Huffman



Huffman Bonus

- When will the Huffman tree be balanced?
- How do we store the tree? Does the efficiency of this matter?
- Are there undefined sequences of bits when using Huffman encoding?
- Does the result of Huffman encoding depend on the text ordering?
  E.g. 🍏🍌🍌🍫 vs. 🍌🍫🍏🍌
- Can there be two equally good Huffman encodings?

Long Range Dependencies

- Correlation
- Conditional entropy

Resources

1. Twitter emojis
2. https://www.ics.uci.edu/~dan/pubs/DC-Sec1.html
3. https://en.wikipedia.org/wiki/Shannon%27s_source_coding_theorem