

# Assignment 2 + Information Theory

## (SNLP Tutorial 3)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

11th, 12th May 2021

# Assignment 2

- Exercise 1: Perplexity Calculation
- Exercise 2: Formulating n-gram models
- Exercise 3: Perplexity Calculation for n-grams
- Bonus: Alternative metric to perplexity

# Overview of Formulas

Concepts and formulations.

- Information Content
- Entropy
- Joint entropy
- Conditional entropy
- Mutual Information (IG)
- Cross-entropy
- KL-Divergence

# Overview of Formulas

Concepts and formulations.

- Information Content
  - Entropy
  - Joint entropy
  - Conditional entropy
  - Mutual Information (IG)
  - Cross-entropy
  - KL-Divergence
- $I(x) = -\log p(x)$

# Overview of Formulas

Concepts and formulations.

- Information Content
  - Entropy
  - Joint entropy
  - Conditional entropy
  - Mutual Information (IG)
  - Cross-entropy
  - KL-Divergence
- $I(x) = -\log p(x)$
  - $H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)$

# Overview of Formulas

Concepts and formulations.

- Information Content
  - Entropy
  - Joint entropy
  - Conditional entropy
  - Mutual Information (IG)
  - Cross-entropy
  - KL-Divergence
- $I(x) = -\log p(x)$
  - $H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)$
  - $H(X, Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(x, y)$

# Overview of Formulas

Concepts and formulations.

- Information Content
  - Entropy
  - Joint entropy
  - Conditional entropy
  - Mutual Information (IG)
  - Cross-entropy
  - KL-Divergence
- $I(x) = -\log p(x)$
  - $H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)$
  - $H(X, Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(x, y)$
  - $H(X | Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(y | x)$

# Overview of Formulas

Concepts and formulations.

- Information Content
- Entropy
- Joint entropy
- Conditional entropy
- Mutual Information (IG)
- Cross-entropy
- KL-Divergence

- $I(x) = -\log p(x)$
- $H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)$
- $H(X, Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(x, y)$
- $H(X | Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(y | x)$
- $I(X; Y) = \sum_{x, y} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)}$



# Overview of Formulas

Concepts and formulations.

- Information Content
- Entropy
- Joint entropy
- Conditional entropy
- Mutual Information (IG)
- Cross-entropy
- KL-Divergence

- $I(x) = -\log p(x)$
- $H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)$
- $H(X, Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(x, y)$
- $H(X | Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(y | x)$
- $I(X; Y) = \sum_{x, y} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)}$
- $H(p, q) = -\sum_x p(x) \cdot \log q(x)$

# Overview of Formulas

Concepts and formulations.

- Information Content
- Entropy
- Joint entropy
- Conditional entropy
- Mutual Information (IG)
- Cross-entropy
- KL-Divergence

- $I(x) = -\log p(x)$
- $H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)$
- $H(X, Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(x, y)$
- $H(X | Y) = -\sum_{x \in X, y \in Y} p(x, y) \cdot \log p(y | x)$
- $I(X; Y) = \sum_{x, y} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)}$
- $H(p, q) = -\sum_x p(x) \cdot \log q(x)$
- $D(p \| q) = -\sum_x p(x) \cdot \log \frac{p(x)}{q(x)}$

## How do they relate to each other?

- Chain Rule:

$$H(X, Y) = H(X) + H(Y|X)$$

$$H(X_1 \dots X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1})$$

## How do they relate to each other?

- Chain Rule:

$$H(X, Y) = H(X) + H(Y|X)$$

$$H(X_1 \dots X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1})$$

- Mutual Information and Entropy

$$I(X; Y) = H(X) - H(X | Y) = H(X) + H(Y) - H(X, Y)$$

## How do they relate to each other?

- Chain Rule:

$$H(X, Y) = H(X) + H(Y|X)$$

$$H(X_1 \dots X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1})$$

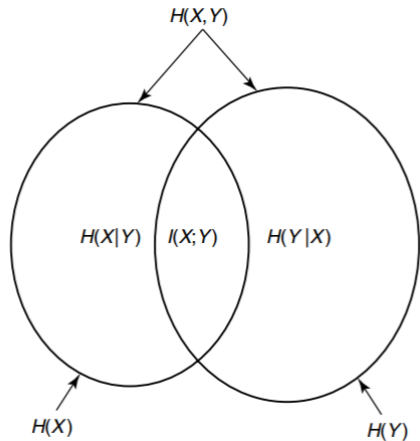
- Mutual Information and Entropy

$$I(X; Y) = H(X) - H(X | Y) = H(X) + H(Y) - H(X, Y)$$

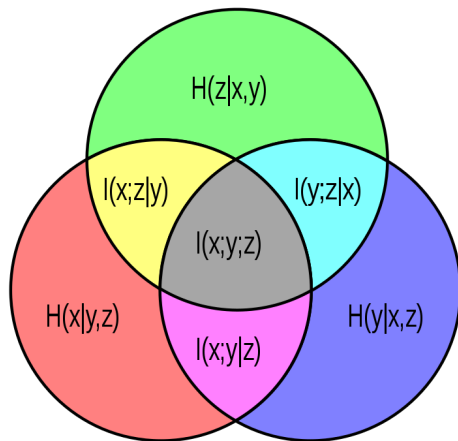
- Apply to 3 variables

$$I(X; Y | Z) = I((X; Y)|Z) = H(X | Z) - H(X | Y, Z)$$

# How do they relate to each other?



Source: <https://syncedreview.com/2020/11/30/synced-tradition-and-machine-learning-series-part-1-entropy/>



Source: [https://en.wikipedia.org/wiki/Information\\_diagram](https://en.wikipedia.org/wiki/Information_diagram)

## Example - Entropy calculation

$X \setminus Y$	0	1
0	$1/2$	$1/6$
1	$1/3$	0

Find

- $H(X), H(Y)$
- $H(X, Y)$
- $H(X|Y), H(Y|X)$
- $I(X; Y)$
- $I(X; Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

## Example - Entropy of functions

What is the (in)equality relationship between  $H(X)$  and  $H(Y)$  when

- $y = f(x)$  (general case)
- $y = 2^x$
- $y = \sin(x)$



## Example - Conditional vs. basic

- Is this true?  $H(Y|X) \leq H(Y)$

## Example - Conditional vs. basic

- Is this true?  $H(Y|X) \leq H(Y)$
- Intuitively?

## Example - Conditional vs. basic

- Is this true?  $H(Y|X) \leq H(Y)$
- Intuitively?
- Formally?

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

- Q: Is age a better predictor for  $y$  than hw performance? How do we measure this?

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

- Q: Is age a better predictor for  $y$  than hw performance? How do we measure this?
- Idea: decide majority class, compute accuracy



## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

- Q: Is age a better predictor for  $y$  than hw performance? How do we measure this?
- Idea: decide majority class, compute accuracy
- Issue: no consideration for split weight (39,10) vs (26,23)

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

- Q: Is age a better predictor for  $y$  than hw performance? How do we measure this?
- Idea: decide majority class, compute accuracy
- Issue: no consideration for split weight (39,10) vs (26,23)
- A:  $I(\text{exam}; \text{hw performance})$

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

- Q: Is age a better predictor for  $y$  than hw performance? How do we measure this?
- Idea: decide majority class, compute accuracy
- Issue: no consideration for split weight (39,10) vs (26,23)
- A:  $I(\text{exam}; \text{hw performance})$
- Q: Can we use conditional entropy instead?

## Example - Feature selection

- Task: Predict if a student  $i$  will pass the exam ( $y_i \in \{\text{no}, \text{yes}\}$ ).
- Input: Massive feature vector  $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \dots)$
- Example:  $(x_1, y_1) = [(24, 2, \text{excellent}, \dots), \text{yes}]$ ,  $(x_2, y_2) = [(23, 5, \text{poor}, \dots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	10
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	22
Excelent	41	3

- Q: Is age a better predictor for  $y$  than hw performance? How do we measure this?
- Idea: decide majority class, compute accuracy
- Issue: no consideration for split weight (39,10) vs (26,23)
- A:  $I(\text{exam}; \text{hw performance})$
- Q: Can we use conditional entropy instead?
- A: Yes, but!

# KL-divergence

Question: Can we use the chain rule on KL-Divergence?

# KL-divergence

Question: Can we use the chain rule on KL-Divergence?

$$D(p(x, y) \parallel q(x, y)) = D(p(x) \parallel q(x)) + D(p(y \mid x) \parallel q(y \mid x))$$

Applications of KL Divergence:

- Bayesian inference
- Compression techniques
- Variational autoencoders

# Assignment 3

- Exercise 1: Understanding entropy in languages
- Exercise 2: Entropy as a measure of uncertainty
- Exercise 3: KL Divergence properties
- Bonus: Computation of KL Divergence

# Resources

- 1 <http://csustan.csustan.edu/~tom/sfi-csss/info-theory/info-lec.pdf>
- 2 <https://www.cs.cmu.edu/~odonnell/toolkit13/lecture20.pdf>
- 3 <https://syncedreview.com/2020/11/30/synced-tradition-and-machine-learning-series-part-1-entropy/>