

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
- Mutual Information (D_{KL})

Overview of Formulas

Concepts and formulations.

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

- $I(x) = -\log p(x)$

Overview of Formulas

Concepts and formulations.

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formula Relations

- $H(X, Y) - H(Y)$
- $H(X) - H(X|Y)$
- $H(Y) - H(Y|X)$
- $H(p, q) - H(p)$

Overview of Formula Relations

- $H(X, Y) - H(Y)$
- $H(X) - H(X|Y)$
- $H(Y) - H(Y|X)$
- $H(p, q) - H(p)$
- Conditional entropy $H(X|Y)$

Overview of Formula Relations

- $H(X, Y) - H(Y)$
- $H(X) - H(X|Y)$
- $H(Y) - H(Y|X)$
- $H(p, q) - H(p)$
- Conditional entropy $H(X|Y)$
- Mutual information $I(X, Y)$

Overview of Formula Relations

- $H(X, Y) - H(Y)$
- $H(X) - H(X|Y)$
- $H(Y) - H(Y|X)$
- $H(p, q) - H(p)$
- Conditional entropy $H(X|Y)$
- Mutual information $I(X, Y)$
- Mutual information $I(X, Y)$

Overview of Formula Relations

- $H(X, Y) - H(Y)$
- $H(X) - H(X|Y)$
- $H(Y) - H(Y|X)$
- $H(p, q) - H(p)$
- Conditional entropy $H(X|Y)$
- Mutual information $I(X, Y)$
- Mutual information $I(X, Y)$
- KL divergence $D(p||q)$

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:

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

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

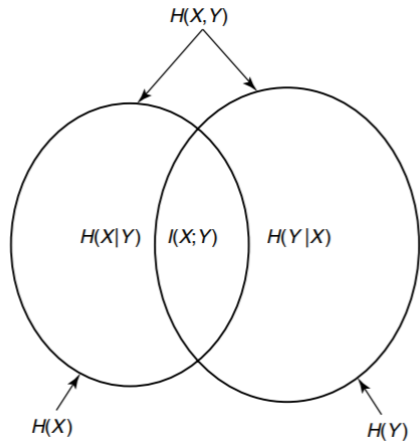
- Mutual Information and Entropy

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

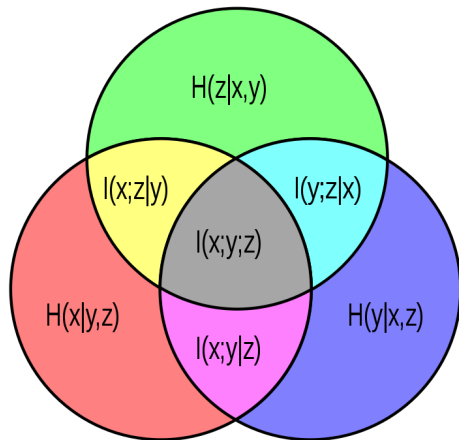
- Apply to 3 variables

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

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

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

- Which one is true? (1) $H(Y|X) \leq H(Y)$, (2) $H(Y|X) \geq H(Y)$ or (3) No systematic bound

Example - Conditional vs. basic

- Which one is true? (1) $H(Y|X) \leq H(Y)$, (2) $H(Y|X) \geq H(Y)$ or (3) No systematic bound
- Intuitively?

Example - Conditional vs. basic

- Which one is true? (1) $H(Y|X) \leq H(Y)$, (2) $H(Y|X) \geq H(Y)$ or (3) No systematic bound
- Intuitively?
- Formally?

Example - Feature selection

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

Age \ Exam			HW \ Exam			Age* \ Exam			HW* \ Exam		
	Yes	No		Yes	No		Yes	No		Yes	No
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

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			HW \ Exam			Age* \ Exam			HW* \ Exam		
	Yes	No		Yes	No		Yes	No		Yes	No
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

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			HW \ Exam			Age* \ Exam			HW* \ Exam		
Yes	No		Yes	No		Yes	No		Yes	No	
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

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			HW \ Exam			Age* \ Exam			HW* \ Exam		
Yes	No		Yes	No		Yes	No		Yes	No	
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

- 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			HW \ Exam			Age* \ Exam			HW* \ Exam		
Yes	No		Yes	No		Yes	No		Yes	No	
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

- 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	HW \ Exam	Yes	No	Age* \ Exam	Yes	No	HW* \ Exam	Yes	No
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

- 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 between equally bad (or good) features, susceptible to imbalance.

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			HW \ Exam			Age* \ Exam			HW* \ Exam		
Yes	No		Yes	No		Yes	No		Yes	No	
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

- 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 between equally bad (or good) features, susceptible to imbalance.
- 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	HW \ Exam	Yes	No	Age* \ Exam	Yes	No	HW* \ Exam	Yes	No
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

- 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 between equally bad (or good) features, susceptible to imbalance.
- 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	HW \ Exam	Yes	No	Age* \ Exam	Yes	No	HW* \ Exam	Yes	No
22	1	2	Poor	1	21	22	2	1	Poor	6	5
23	19	7	Ok	23	12	23	19	1	Ok	23	0
24	39	30	Excelent	41	3	24	39	2	Excelent	41	0
25	25	8				25	25	1
...			

- 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 between equally bad (or good) features, susceptible to imbalance.
- 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/>