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})

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

Overview of Formulas

Concepts and formulations.

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

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

How do they relate to each other?

- Chain Rule:

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

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

How do they relate to each other?

- Chain Rule:

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

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

- Mutual Information and Entropy

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

How do they relate to each other?

- Chain Rule:

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

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

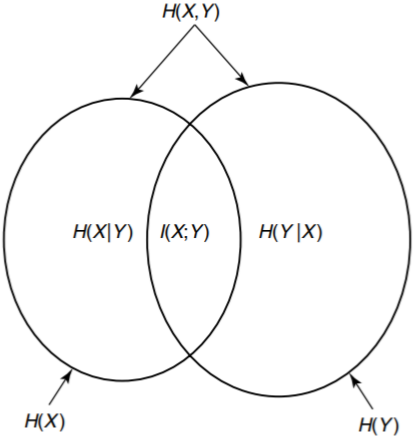
- Mutual Information and Entropy

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

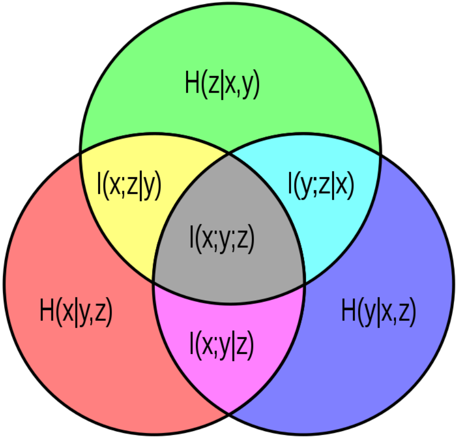
- Apply to 3 variables

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

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 \ 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
- Intuitivelly?

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
- Intuitivelly?
- 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	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

Example - Feature selection

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

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

Example - Feature selection

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

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

Example - Feature selection

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

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

- 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 $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \ldots)$
- Example: $(x_1, y_1) = [(24, 2, \text{excellent}, \ldots), \text{yes}], (x_2, y_2) = [(23, 5, \text{poor}, \ldots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

- 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 $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \ldots)$
- Example: $(x_1, y_1) = [(24, 2, \text{excellent}, \ldots), \text{yes}], (x_2, y_2) = [(23, 5, \text{poor}, \ldots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

- 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, suspectible 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 $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \ldots)$
- Example: $(x_1, y_1) = [(24, 2, \text{excellent}, \ldots), \text{yes}], (x_2, y_2) = [(23, 5, \text{poor}, \ldots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

- 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, suspectible 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 $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \ldots)$
- Example: $(x_1, y_1) = [(24, 2, \text{excellent}, \ldots), \text{yes}], (x_2, y_2) = [(23, 5, \text{poor}, \ldots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

- 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, suspectible 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 $x_i = (\text{age}, \text{semesters at uni}, \text{hw performance}, \ldots)$
- Example: $(x_1, y_1) = [(24, 2, \text{excellent}, \ldots), \text{yes}], (x_2, y_2) = [(23, 5, \text{poor}, \ldots), \text{no}]$

Age \ Exam	Yes	No
22	1	2
23	19	7
24	39	30
25	25	8
...	...	...

HW \ Exam	Yes	No
Poor	1	21
Ok	23	12
Excelent	41	3

Age* \ Exam	Yes	No
22	1	1
23	19	1
24	39	2
25	25	1
...	...	...

HW* \ Exam	Yes	No
Poor	6	5
Ok	23	0
Excelent	41	0
...	...	...

- 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, suspectible 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?

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

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/