כריית נתונים - מעבדה 2 - תורת האינפורמציה של שאנון ואנטרופיה

בשנת 1948 פרסם קלוד שאנון (אחד מהוגי המחשב הספרתי ואבי תורת המידע) את מאמרו "תאוריה מתמטית של התקשורת"
שאנון הגדיר את מושג האנטרופיה עבור מציאת כמות ביטים ממוצעות לקידוד סימן בהודעה העוברת בתקשורת. מנקודת מבטו של מדען הנתונים זהו מדד לחוסר הוודאות באינפורמציה - איך מכמתים (נותנים ערך מספרי) חוסר וודאות בסדרת נתונים.

$$H = -\sum_i p_i \log_2(p_i)$$

$$p_i = \frac{\text{number of times ith character appears}}{\text{length of message}} = \frac{\text{number of times ith value appears}}{\text{num of all values}}$$

מעבדה זו יש לממש ולהגיש בפייתון על גבי ג'ופיטר.
בחלק של המימוש העצמי אין להשתמש בפונקציות או ספריות מוכנות - יש לממש בעזרת פייתון והספריות הסטנדרטיות בלבד.

קיראו וסכמו על המושגים הבאים במילים שלכם

1. אנטרופיה

$$H = -\sum_i p_i \log_2(p_i)$$

2. אנטרופיה מותנית

$$H(Y/X) = -\sum p(x,y) \cdot \log(p(y/x))$$

3. אנטרופיה הדדית (Mutual Information)

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

4. הגבר מידע (Information Gain)

Information gain = entropy (parent) − [weightes average] ★ entropy (children)

(מתוך https://miro.medium.com/max/1400/1*bVGWGFTTprZhSphr7sXFVw.png)

1. כיתבו פונקציה בפייתון המקבלת סדרת ערכים ומחזירה את ערך האנטרופיה שלה

2. כיתבו פונק' בפייתון המקבלות שתי סדרות מקבילות של ערכים
- הסדרה הראשונה תייצג מאפיין
- השניה תייצג עמודת סיווג
    1. ממשו חישוב לאנטרופיה מותנית
    2. ממשו חישוב לאנטרופיה הדדית
    3. ממשו חישוב להגבר מידע

3. מיצאו פונקציות ספריה מוכנות המבצעות את ארבע הפונקציות שהתבקשתם לממש הדגימו והשוו את פעולתן (ייתכן ואין את כל הפונקציות מן המוכן ובמקרה כזה יש לציין היכן חיפשתם)

חומרים

הסבר על הפונקציה המוכנה מתוך scipy
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html

פונקציות ספריה מוכנות
https://pypi.org/project/pyitlib/

הסבר על פונקציות המודדות חוסר וודאות (impurity בהסבר)
https://www.bogotobogo.com/python/scikit-learn/scikt_machine_learning_Decision_Tree_Learning_Informatioin_Gain_IG_Impurity_Entropy_Gini_Classification_Error.php

הסבר על אנטרופיה
https://victorzhou.com/blog/information-gain/

הסבר על אנטרופיה מותנית, הדדית והגבר מידע
https://machinelearningmastery.com/information-gain-and-mutual-information/

הסבר נוסף
https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01

שימוש באנטרופיה למחקר במיפוי
https://www.mdpi.com/1099-4300/15/4/1464