הפקולטה למדעי הנתונים וההחלטות
MAKING SENSE OF IT

הטכניון מכון טכנולוגי לישראל

מעבדה בניתוח והצגת נתונים (094295) אביב תשפ"ג 2023

תרגיל בית 3

Data Analysis and Prediction on Academic Citation Networks

תאריך הגשה: 04/07/2023

מבוא

בתרגיל זה, תתנסו בשימוש באלגוריתמי למידת מכונה וכלים סטטיסטים עבור ניתוח רשתות.

משימתכם העיקרית תורכב מניתוח, חקירה וסיווג של סט נתונים המתאר רשת ציטוטים אקדמאיים.

הנתונים

סט הנתונים הוא גרף מכוון, המייצג רשת ציטוטים אקדמאיים.

הגרף מכיל כ- 100,000 קודקודים כאשר כל קודקוד בגרף מייצג מאמר וכל קשת מכוונת בין קודקוד א' לקודקוד ב' מייצגת שקודקוד א' ציטט את קודקוד ב'.

כל מאמר מיוצג על ידי וקטור פיצ'רים שנוצר ממיצוע של כל ייצוגי המילים (נוצרו על ידי מודל skip-gram) באבסטרקט והכותרת שלו.

כמו כן, סט הנתונים מכיל עבור כל מאמר את שנת הפרסום שלו ומספר המייצג את הקטגוריה אליו הוא שייך.

את סט הנתונים ניתן להשיג באופן אוטומטי דרך יצירת אובייקט מהמחלקה HW3Dataset (בdataset.py המצורף באתר הקורס) שיורשת מהמחלקה Dataset של הספרייה PyTorch Geometric (בה מומלץ להשתמש בכל משימות התרגיל).

שימו לב - אין לשנות את מבנה התיקיות של קבצי הנתונים .

סט הנתונים החבוי, אשר יוזכר בהמשך, יושג באותו האופן וישתמש באותה היררכיית תיקיות.

אובייקט הData המוכל בתוך אובייקט הDataset של סט הנתונים הגלוי יכיל את הנתונים הבאים:

```
Data(x= [100000, 128], edge_index= [2, 444288], y= [100000, 1], node_year= [100000, 1], train_mask= [80000], val_mask= [20000])
```

x: Feature vectors for each node, shape [100000, 128].

edge_index: Graph edge connectivity, shape [2, 444288]

y: Node-level target/label values, shape [100000, 1]

node_year: Year associated with each node, shape [100000, 1]

train_mask: Indices for training nodes, shape [80000]

val_mask: Indices for for validation nodes, shape [20000]

שימו לב – train_mask וval_mask ישמשו אתכם להפרדה בין סט האימון והולידציה.

משימה

בשנים האחרונות קצב פרסומי המאמרים עולה ויש צורך בכלים אוטומטיים שממיינים ומנגישים את המאמרים לחוקרים.

כחלק ממטרה זו, משימתכם בתרגיל זה היא לחזות את קטגוריות המאמרים.

דרישות ומדידת הביצועים

המדד המרכזי למדידת ביצועי המודל יהיה **Accuracy** ודרישת המינימום על סט הולידציה היא **0.5**.

מה מגישים

1. דו"ח בן 5-7 עמודים (אפשר לצרף אפנדיקס). מבנה הדו"ח גמיש אך עליו להיות מסודר ולהכיל הסברים איכותיים לפעולות אשר ביצעתם לאורך התרגיל.
2. קישור ל GitHub Repository שלכם המכיל את כל קבצי הקוד שכתבתם ואת predict.py.

סקריפט predict.py

על הסקריפט להשתמש בפקודה הבאה ליצירת סט הנתונים:

```
dataset = HW3Dataset(root='data/hw3/')
```

שימו לב כי אותו הפקודה תשמש גם ליצירת אובייקט הDataset לסט הנתונים החבוי.

לאחר מכן, על הסקריפט לטעון מודל מאומן, לסווג כל קודקוד לאחת מ40 הקטגוריות (מיוצגות על ידי מספרים שלמים בין 0ל39) ולשמור קובץ prediction.csv אשר יכיל את תוצאות החיזוי.

קובץ ה prediction.csv יכיל שתי עמודות – idx וprediction כך שהidx הוא האינדקס המתאים לקודקוד בdataset[0].x והקטגוריה מיוצגת על ידי מספר שלם (int).

idx	prediction
0	27
1	0
2	39

שימו לב – predict.py ישמש להסקה בלבד ללא אימון וללא גישה לתיוגי האמת.

חלק תחרותי

בנוסף לעמידה בדרישת המינימום על ביצועי המודל על סט הולידציה (אשר מוגדר דרך val_mask) המודל שלכם יבחן גם על סט נתונים חבוי שיורכב מאותם הרכיבים כמו הסט הגלוי (dataset[0]) יכיל את אותם רכיבים אבל **ללא val_mask** ו**train_mask**).

ההצלחה של המודל על סט זה תימדד באופן יחסי לתוצאות שאר משתתפי הקורס.

מבנה הציון

- דו"ח מסכם (רמת פירוט, רגורסיות, מתודולוגיה נכונה) - 80%
- חלק תחרותי - 20%

הערות לתרגיל

- מקורות מומלצים:
  - Colab Notebooks and Video Tutorials — pytorch_geometric documentation (pytorch-geometric.readthedocs.io)
  - CS224W: Machine Learning with Graphs | 2021 | Lecture 6.1 - Introduction to Graph Neural Networks - YouTube
  - A Gentle Introduction to Graph Neural Networks (distill.pub)
- אתם יכולים להשתמש בכל שיטת קלסיפיקציה שעולה על רוחכם (במגבלת משאבי השרת).
- על הקוד להיות קריא ולהכיל הערות בעת הצורך.
- **אין להתקין ספריות נוספות מלבד אלה המוכלות בסביבה (environment.yml) שקיבלתם.**
- **אין לשנות את המחלקה HW3Dataset .**
- **על האימון להיות בפרדיגמת Transductive Learning – מותר לפעפע מידע מכל הקודקודים, כולל מהולידציה, אבל אסור להשתמש בתיוגים של הולידציה בזמן תהליך הלמידה (כמובן שבהסקה אסור לעשות שימוש בתיוגים בכללי).**
- אין לעלות את הדאטא repository.
- אין לאמן מודלים בסקריפט **predict.py**. תוכלו לשמור מודל מאומן מראש ב repository שלכם ולקרוא אותו מתוך הסקריפט.
- **אסור להעשיר את הסט הנתונים ע"י מקורות מידע חיצוניים**. עליכם לספק את כל קוד האימון למודל שלכם ב repository. במידה ונחשוד שהמודל שאימנתם קיבל דאטה ממקורות חיצוניים, נאמן אותו בשנית ונבדוק האם יש הבדל מובהק בביצועים. הגשה שלא תעמוד בהנחיה זו (אי צירוף קוד האימון ו/או שימוש במקור חיצוני) **תיפסל.**
- יש להגיש דוח מוקלד.
- יש לתת קרדיט לכל מקור חיצוני שנעשה בו שימוש.
- הגשת התרגיל תתבצע באתר הקורס על ידי אחד השותפים בלבד.
- איחור בהגשה ייקנס ב10% לכל יום איחור לא מוצדק. על איחורים מוצדקים (כדוגמת שירות מילואים), יש לדווח לפני מועד הגשת העבודה. ניתן יהיה להגיש עבודות באיחור אשר לא יעלה על שבוע.
- הקוד אמור להיות מסוגל לרוץ ב-Azure. יש לעקוב אחרי הנחיות מדריך ה DevOps להקמת סביבה.
- הקוד אמור לרוץ ובזמן סביר.

בהצלחה!