סמינר במערכות לומדות וכשלונותיהן
Seminar on Failures Modes in Machine Learning (FAILML)

מספר הקורס: 236839

סמסטר: 2024 אביב

מרצה:	ניר רוזנפלד
שעות הרצאה:	יום שני 13:30-15:30
דרישות קדם:	מבוא למערכות לומדות 236756 (או מקביל) **הקורס מיועד לתלמידי תארים מתקדמים או לתלמידים מתקדמים לתואר ראשון המתעניינים במחקר בתחום.**
אתר הקורס:	יפורסם בהמשך

<u>רישום:</u>

בקשה להרשם לקורס תעשה על ידי מילוי הטופס הבא **<u>בלבד</u>** (ולא דרך מערכת הרישום):
https://forms.gle/Riej7dtnMGcptpDu7
רישום בפועל הוא באישור מרצה בלבד, הודעה על אישור תשלח לקראת תחילת הסמסטר.

<u>תאור הקורס</u>

Supervised machine learning relies on the fundamental assumption that data is sampled iid from the same distribution at train time and at test time. But in virtually any realistic application, this assumption is unlikely to hold. In this seminar we will survey papers that study when, how, and why learning algorithms (such as ERM) can fail when the assumption is violated. We will study various failure modes that stem from different reasons underlying why train and test distribution can differ, including: natural distribution drift, model-induced distribution shift, adversarial manipulation of inputs, and strategic behavior of self-interested users.

<u>Topics</u>: (list of papers will be published towards the beginning of the semester)

- Out-of-distribution (OoD) generalization
- Distribution shift and drift
- Semi-supervised domain adaptation
- Covariate shift and debiasing
- Inverse propensity weighing (IPW)
- Invariant representation and risk minimization (IRM)
- Causality vs. (spurious) correlation
- Decision-dependent distribution shift
- Adversarial learning and robustness
- Strategic classification
- Performative prediction
- Distributional robustness

<u>דרישות הקורס:</u>

- בחירת מאמר להצגה והצגתו בכיתה. ניתן להציג מאמר כיחידים, או שני מאמרים כזוג.
- בחירת מאמר לביקור והצגת הביקורת בכיתה. הצגת ביקורת תעשה באופן יחידני.
- מטלת קריאה שבועית: קריאת המאמרים שיוצגו אל ידי אחרים (מאמר אחד לשבוע; הקריאה תידבק)
- נוכחות בלפחות 11 שיעורים.
- השתתפות פעילה בשיעור (בונוס)

<u>תוצרי למידה:</u>

בסוף הקורס הסטודנטים יוכלו:

- לזהות את האופנים בהן מערכת לומדת נתונה יכולה להכשל
- לשייך את נטייתה להכשל למרכיבים שונים באלגוריתם, בדאטא, או בסביבה
- להציע פתרונות המתאימים לכשלים אלו, ובפרט ביחס ל-distribution shift
- לקרוא באופן ביקורתי ולבקר מאמרים אקדמיים עכשויים בתחום
- לזהות הנחות מרכזיות, הן מפורשות והן סמויות, ואת השלכותיהן
- להציע כיונים מתאימים למחקרי המשחך
- לתכנן ולבצע הצגת תוכן אקדמי בעל פה
- לתת, לקבל, ולקחת חלק בתהליך משוב עמיתים