

תרגיל 4

להגשה עד 19.6

הערות להגשה:

- יש להשתמש בפייטון ולממש את ה-classifiers מ-scratch. לדוגמא, לא להשתמש בשום ספריה של machine learning. כמובן שאפשר להשתמש ב-numpy ולהתבסס על השקפים בשיעור.
- עבור כל תרגיל יש להגיש את הקוד בפייטון + צילומי מסך של התוצאות. **שימו לב**- תרגילים ללא צילומי מסך לא יבדקו!

Spark + Naïve Bayes:

ראינו בשיעור classifier של Naïve Bayes המבוסס על ספירה של המופעים של מילים ביחס למס' ההודעות הקיים במילון.

- שנו את הקוד וצרו classifier המבוסס על ספירה של מילים ביחס למס' **המילים** הקיימות במילון
- Laplace's smoothing (הוסיפו מנגנון החלקה)

Linear and logistic regression:

- מצ"ב רשימת נתונים על תנאי שכר של מרצים, הכוללים מגדר (1 for female and 0 for male), סוג המרצה (תואר. rank==1 for assistant professor; 2 for associate professor, and 3 for full professor), ותק (מס' שנים) ומשכורת (Highest degree, coded 1 if doctorate, 0 if masters).

קראו את הנתונים ב-python:

- בנו מודל **linear regression** בפייטון אשר חוזה את המשכורת בהתבסס על מגדר, סוג המרצה, תואר וותק. השתמשו ב-70% מהנתונים עבור בניית המודל, ב-30% הנותרים בידקו את המודל שיצרתם.
- דווחו את התוצאות ב-**mean squared-error**. יש להשוות את התוצאות של המודל לנתונים הקיימים.
- הוסיפו עוד features ובידקו האם זה שיפר את התוצאות.
- בנו **logistic regression classifier** בפייטון אשר חוזה את המגדר בהתבסס על סוג המרצה, תואר, ותק ומשכורת. השתמשו ב-70% מהנתונים עבור בניית המודל, ב-30% הנותרים בידקו את המודל שיצרתם.
- accuracy, recall, precision and F-measure**. דווחו את התוצאות במונחים של
- הוסיפו עוד features ובידקו האם זה שיפר את התוצאות.