עדכונים:

נזכיר -  
הקובץ נתונים שאנחנו בחרנו נלקח מאתר UCI ,שבו מאגרי נתונים ללימוד למידה חישובית . הקובץ מכיל מאגר מידע על אנשים ממדינות שונות. כולל בתוכו 14 מאפיינים ו עמודת תוצאה 32562 שורות. נרצה לבנות מודל שיחזה האם אדם מסוים ירוויח יותר או פחות 50 אלף בשנה.

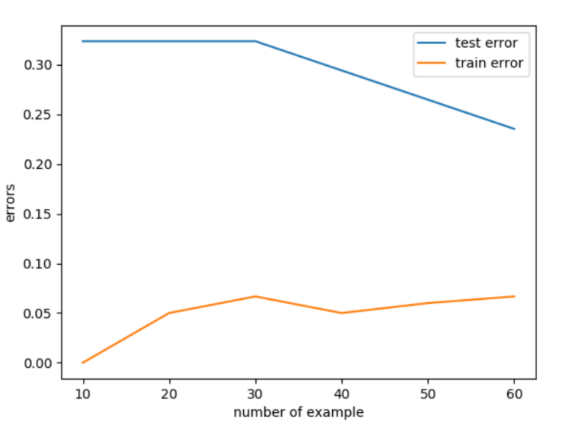
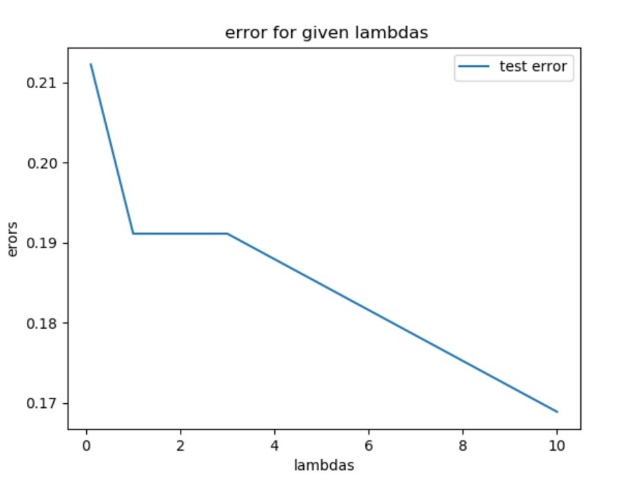
בחרנו בנושא זה כי נושא שמעניין את רובנו, מכיל מידע רב, שברובו נתונים מורכבים המאפשר ביצוע של פילוחים שונים ומגוונים.

בפרוייקט הקודם - סיננו את הדטה כל פעם לפי מדינה אחרת ובדקנו את המודל בנפרד על כל מדינה   
כעת, נציג את המודל שמתקבל על מספר מדינות

Logistic Regression -   
# חיפשנו את ערך ה  האופטימלי, בעזרת k-fold cross validation.  
 והצגנו בגרף את ערך השגיאה לכל .  
# עבור ה  הכי טוב ציירו את הגרף של ציר ה-x מספר הדוגמאות וציר ה-y   
 השגיאה על קבוצת הלמידה והשגיאה על קבוצת test קבועה

בפרוייקט הקודם,  
הגרפים שהצגנו בדוח היו על דטה - שאחרי סינון לפי מדינה אחת - הכילה 95 דוגמאות.  
הגרף השמאלי מראה את השגיאות עבור כל למדה, הלמדה שהשגיאה בה הכי מינימלית היא 10=, אבל זה לא אומר שהיא האופטימלית ביותר,  
יכול להיות שעבור למדה גדולה יותר הגרף ימשיך לרדת.

הגרף מימין מראה שהשגיאות של המודל עבור הנתונים שהוא מכיר נמוכות בהרבה מהשגיאות על הנתונים שהוא לא מכיר, כלומר overfitting

****

הצגת השיפורים של עדכון הפרוייקט בעמוד הבא

כעת נריץ על שלוש מדינות (מספר הדוגמאות 322)  
בגרף השמאלי – הלמדה האופטימלית היא 3  
ואפשר לדעת בוודאות שזו הלמדה האופטימלית, מכיוון שעבור למדה נמוכה יותר או גבוהה יותר – השגיאות גבוהות יותר.

בגרף הימני – רואים שהפער בין השגיאות עבור הנתונים שהוא מכיר לבין השגיאות של הנתונים שהוא לא מכיר – הולך ומצטמצם ככל שמספר הדוגמאות עולה

