# מבוא לבינה מלאכותית: מבחן בית - הפורמט הקצר

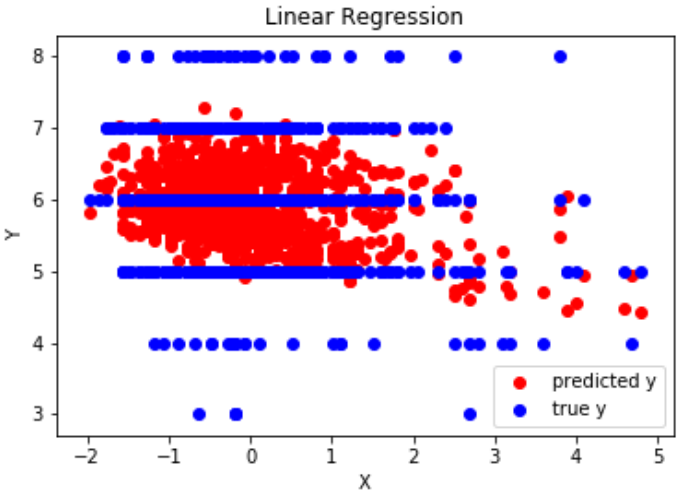
יוסי לוי ת.ז. 037597028

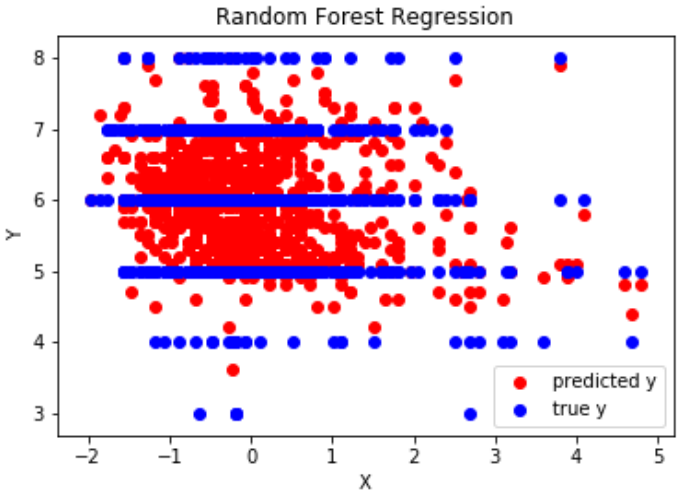
13.08.2020

## 1

כל ערכי הפיצ'רים הם ערכים מנורמלים. בכל חלקי השאלה

1. בחלק זה נתבקשנו להריץ את הנתונים על שני מודלים ולחזות את איכות היין. המודל הראשון שנבחר הוא רגרסיה ליניארית והשני הוא Random Forest.

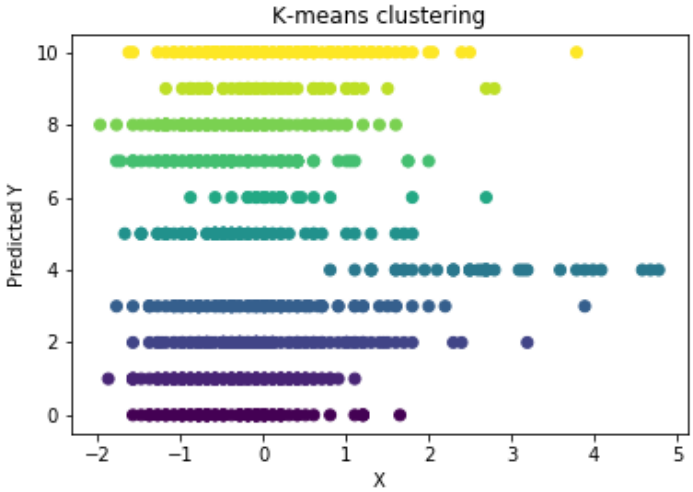



הגרפים מציגים את היכולת לחזות את איכות היין (PREDICTED Y) אל מול הערך האמיתי של איכות היין (TRUE Y).
בוצע מבחן TSET-T לבלתי תלויים על סמך ריבוע השגיאות של ערך החיזוי של כל דגימה אל מול הערך האמיתי של איכות היין (PREDICTED Y - TRUE Y). ערך ה-P VALUE

של מבחן זה יצא 1.3067221214172147e-05 כשממוצע ריבועי השגיאות של הרגרסיה הליניארית יצא 0.5690247717229263 ושל המודל random forest יצא 0.382377551020408 ולכן מבין שתי השיטות הללו מודל הרגרסיה הטוב יותר הוא מודל ה-random forest בו סך ממוצעי ריבועי השגיאות קטן באופן מובהק מזה של מודל הרגרסיה הליניארית.

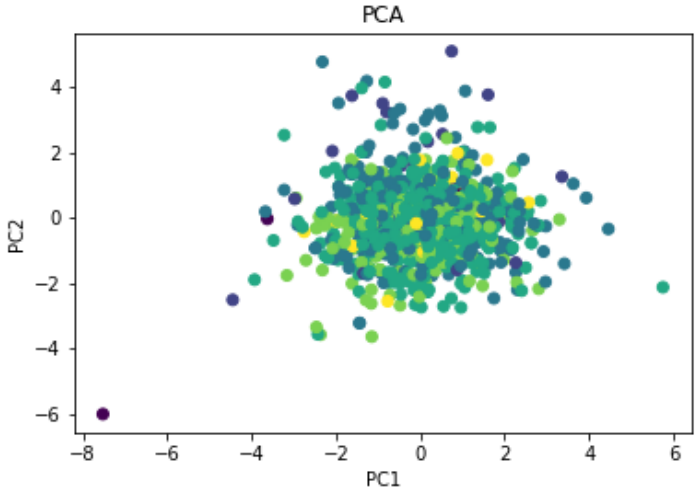
2. בחלק זה הדרישה היתה לבצע אשכול של הנתונים על סמך שני מודלים שונים. המודל הראשון שנבחר הוא k-means והמודל השני הוא GMM(gaussian mixture model).בחלק זה טיב האשכול הוערך על ידי silhouette score.ההרצה של כל אחת משיטות ה-clustering הנ"ל בוצעה 30 פעמים ולכל אחת חולץ מדד זה. הממוצע k-mean היה 0.11178440152783391 וממוצע GMM היה 0.0016525120223483622. כבר מהשוואה זו ניתן לראות כי ששיטת k-mean עדיפה לחלוקה לאשכולות מאחר והיא בעלת הערך הגבוה יותר והקרובה ל-1 (הערך הגבוה ביותר למדד). מעבר לכך בוצע מבחן t-test גם כאן וערך ה-p-value יצא 1.260634256993649e-56, גם כאן ההבדלים יצאו מובהקים.

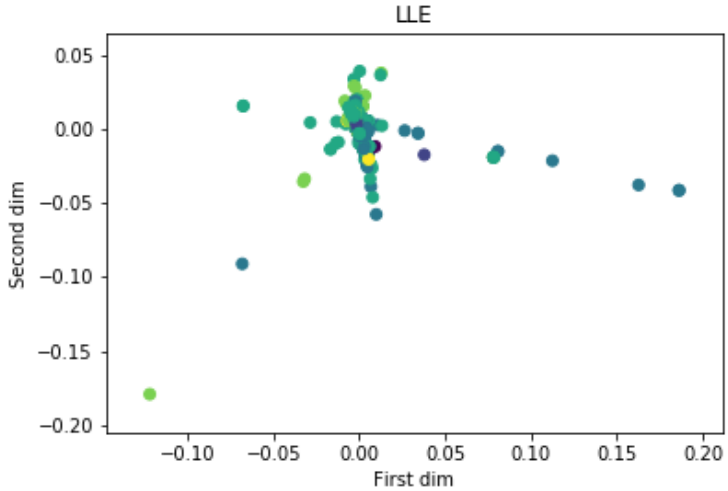




הגרף העליון מראה את שיטת החלוקה לאשכולות על סמך GMM והגרף התחתון לפי שיטת k-means. הגרפים מציגים את החלוקה לאשכולות על סמך הפיצ'ר הראשון כאשר ציר ה-y מסמן את האשכול של כל דגימה ונע בטווח 0-10.

3. בחלק זה הדרישה היתה ליישם שתי שיטות להורדת מימדים. השיטה הראשונה שנבחרה היא שיטת LLE(locally linear embedding), והשיטה השניה היא PCA (principal component analysis). המדד להשוואה בין השיטות הוא reconstruction minimum error. מדד זה ב-LLE הוא 9.209806608190625e-19 וב-PCA הוא 0.46563534972005655 ולכן מבין שניהם ההטלה של LLE היא קורלטיבית יותר לאיכות היין.המשמעות היא שהורדת המימדים שמרה על עיקר המידע ונוכל להסתמך על

המודל בכדי לחזות בהצלחה גבוהה יותר על סמך LLE.




שני הגרפים הנ"ל מראים את האשכולות לפי שיטת PCA ושיטת LLE.