## What is the idea of bagging?

יוצרים מספר דאטהסטים נפרדים בעזרת bootstrapping ומאמנים מודל על כל אחד מהם. כעת, נשתמש בתוצאות של כל מודל על מנת לספק תוצאה סופית (במקרה של רגרסיה נוכל לעשות ממוצע של התוצאות ובמקרה של סיווג נוכל לבצע הצבעה). הרעיון מאחורי bagging הוא שאם נשלב כמה מודלים בלתי תלויים, נקבל מודל שהשונות שלו תהיה נמוכה מזו של מודל יחיד.

## What is the trade off in choosing the number m of features to sample in every node?

ככל שנאפשר בחינה של מספר פיצ'רים גדול יותר, כך המודל יוכל להשתמש ביותר מידע וכנראה יניב תוצאות יותר טובות כמודל בודד. מצד שני, מודל שיורכב ממספר מודלים כאלה כנראה יסבול ממודלים תלויים אחד בשני מכיוון שתהליך הבנייה של כל אחד מהם יהיה דומה. התוצר של מצב כזה הוא שהשונות לא תצטמצם.

ככל שנאפשר לו פחות פיצ'רים, כך המודל יהיה יותר גנרי ולכן המודל אשר יורכב ממספר רב של מודלים מצומצמים כאלה (בהנחה שלא נותנים להם את אותם הפיצ'רים) יהיה בעל שונות נמוכה יותר (כי המודלים יהיו פחות תלויים זה בזה).

## What’s Out Of Bag Sampling? What is the size of the Out Of Bag set when the dataset is large?

OOB samples אלה הרשומות בדאטה המקורי אשר לא נכללו בדאטה שממנו נוצר הדאטה החדש (בעזרת bootstrapping). הכמות של OOB samples עבור מאגר נתונים גדול מספיק היא בערך 0.37 מהדאטה:

ההסתברות לא לבחור דוגמה מסויימת מהמאגר עבור יצירת דאטה בגודל N כאשר גם גודל הדאטה המקורי הוא N היא: והגבול של נוסחה זו הוא . כלומר, עבור N מספיק גדול, ההסתברות של דוגמה לא להיבחר היא 0.368 ולכן כמות דוגמאות ה OOB היא .

## How can one measure feature importance in random forest?

בכל מודל נשמור את "השיפור" שכל פיצ'ר תרם במהלך הבנייה (למשל RSS ברגרסיה ו entropy בסיווג). הפיצ'רים המשמעותיים ביותר יהיו אלה שתרמו הכי הרבה בממוצע על כל המודלים שמרכיבים את היער.

## Solve exercise 15.1. What is its implication on random forest?

Need to prove that the variance of the average of B identically distributed random variables, each with variance of and a correlation of equals to.

כלומר

מתקיים גם

נסמן משתנה ב במקום ב

מתקיים גם

בהנחה שעובדים עם קורלציית pearson, מתקיים

כלומר

ולכן

ובסה"כ כנדרש.

בשלב בו חישבנו את מתקיים כי ולכן .

## Demonstrate in code that when the number of variables is large, but the fraction of relevant variables is small, random forests are likely to perform poorly with small m.

קוד במחברת בשם random\_forest\_theoretical\_6