## How one might use a classification tree to return class probabilities?

בכל עלה בעץ ישנם מופעים לכל תיוגים אפשריים. על מנת להחזיר תיוג משתמשים בפרופורציה שבעלה ומחזירים את התיוג הכי נפוץ. על מנת להחזיר הסתברויות של כל תיוג נחזיר את הפרופורציות עצמן. כלומר, נחזיר   כפי שהוגדרו בספר.

## What are the three measures used for building classification trees? What are their relative merits? What is used for regression trees?

Misclassification error של צומת מסוים שווה לחלק היחסי של התיוגים השגויים באותו צומת ביחס לתיוג הכי נפוץ בצומת. כלומר, מסתכלים על כמות המופעים בצומת אשר התיוג בהם לא זהה לתיוג הכי נפוץ.

Gini index של צומת מסוים מתייחס לחלק היחסי של התיוגים השגויים באותו הצומת אבל לא רק עבור התיוג הכי נפוץ אלא עבור כולם (תוך הסתכלות על הכמות היחסית של התיוגים של אותו התיוג).

Cross entropy מאד דומה ל gini index

היתרונות של השניים האחרונים הם שאפשר לגזור אותם (מקל על שימוש באופטימיזציות) ושהם יותר רגישים מ misclassification error לשינויים בהתפלגות של דוגמאות בצומת (עבור פיצולים שיוצרים misclassification rate זהה, הם יעדיפו את הפיצול שיוצר צמתים יותר "טהורים"-יותר הומוגניים. וזו התנהגות שנעדיף).

על מנת לבנות עצי רגרסיה, נשתמש בקריטריון פיצול שימזער את סכום ההפרשים בין התיוגים האמיתיים של הדוגמאות בצומת לבין הממוצע שלהם (שישמש כערך שנחזה עבור קלט שיגיע לצומת הזו) - RSS:

## How can Gini index be interpreted?

כמה צומת "טהורה". כלומר, כמה ההתפלגות של המופעים בצומת קרובה למצב בו כל המופעים בצומת שייכים לאותו התיוג.

תשובה נוספת (מהספר): אם נגדיר שנקבע את הסיווג של מופע מסוים להיות k בהסתברות אז שגיאה בזמן אימון (training error) של צומת יהיה ה gini index שלה.

## What are the disadvantages of trees?

1. יש להם שונות (variance) גבוהה בגלל המבנה שלהם. שינוי קטן בדאטה יכול לגרור פיצול שונה בשלב מוקדם בבניית העץ, מה שישפיע בהכרח על פיצולים נוספים בהמשך הבנייה.
2. העצים מכלילים בצורה די גסה את התיוג/תוצאת רגרסיה על כל הקלטים שמגיעים לעלה מסוים, מה שיכול לפגוע בתוצאות במקרה של רגרסיה.
3. בצורה הבסיסית, עץ סובל מ overfitting

## What is does pruning aim to solve? How is it done?

הבעיה ש pruning מנסה לפתור היא overfitting. התהליך נעשה בעזרת הורדת עלים שלא תורמים לנו מספיק ביחס לערך מסוים (קבוע או תלוי בערך כמו גודל העץ) מהעץ המקורי. הורדת עלים יוצרת עץ פשוט יותר, מה שמוריד את סיבוכיות (complexity) של המודל.

## What may be the problem with classification decision trees when a categorical feature with many possible values (is used)? Suggest a solution to the problem

הבעיה היא שהחלוקה על המשתנה תניח שיש איזשהו הגיון/סדר באופציות של המשתנה (בדומה למשנה נומרי, שיש לו אינסוף אפשרויות אך קיים ביניהם קשר- כמו גובה או משקל) וזה לא בהכרח המצב (למשל, משתנה המתאר צבע עיניים כאשר 1-חום, 2-כחול ו 3-ירוק. עבור המספור המוצג, ייתכן וצומת תחלק את המידע על פי ערך סף מסוים שלא יהיה תקף עבור סדר שונה, וקביל באותה מידה, של צבעים). פתרון אפשרי הוא ליצור משתנה בינרי לכל אופציה של המשתנה המקורי (במקום צבע עיניים ליצור משתנה עבור כל צבע).