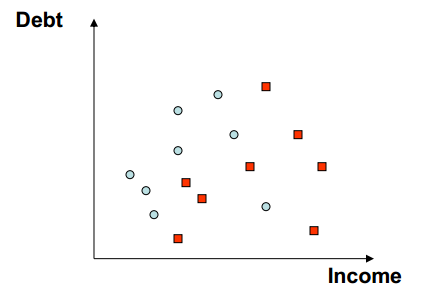
**Classification (Decision) Trees– עצי החלטה**

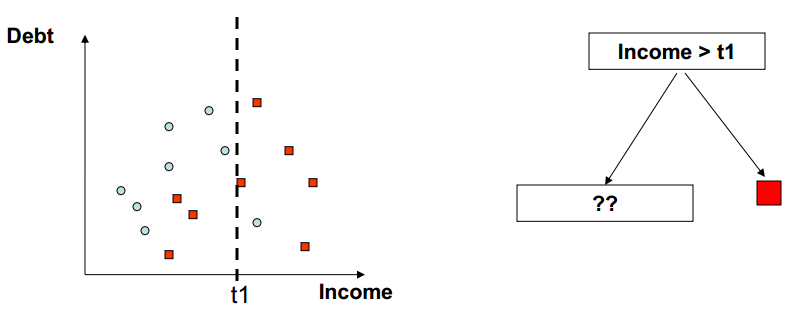
* עצים הם בין הפופולריים והשימושיים במודלים של כריית נתונים
* גרסה אלגוריתמית של סיווג (classification)
* יתרונות:
* אין הנחות על התפלגות כלשהי
* יכול להתמודד עם קלטים נומינליים ואמיתיים
* מהיר ובעל הדרגה
* חסין עבור ערכים חריגים וחסרים
* פשוטים להבנה ופירוש
* קומפקטיות של כללי הסיווג
* חסרונות:
* בעל פירוש?
* מספר פרמטרים מכוונים כדי להגדיר עם הדרכה קטנה
* גבול ההחלטה הוא לא רציף



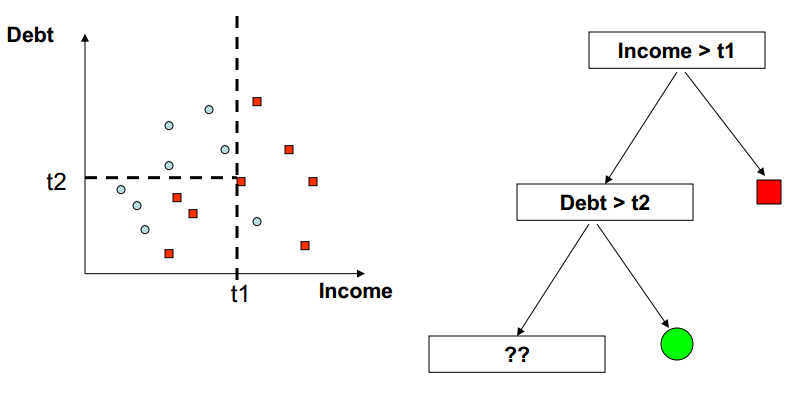
לדוגמא: האם האנשים משלמים חשבונות?

העיגולים הכחולים – הכנסה

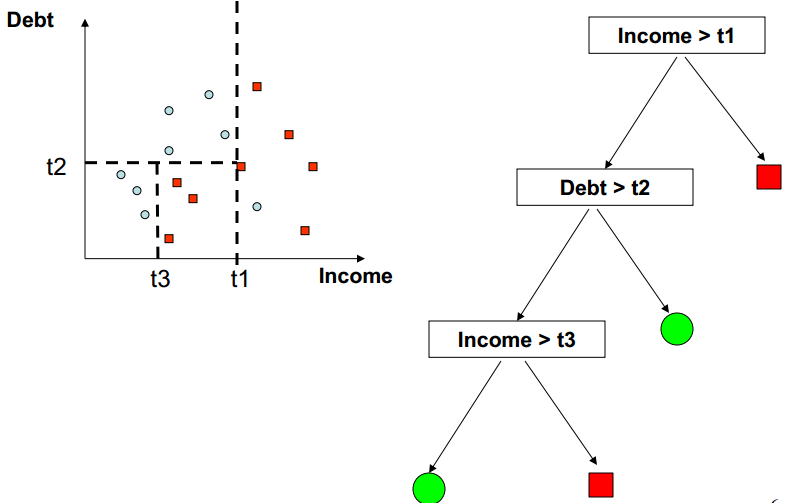
הריבועים האדומים – חוב



לוקחים קבוע כלשהו (t1), אם המשכורת קטנה מ t1 אז לא משלם חשבונות, ואם המשכורת גדולה מ t1 ממשיכים לשאלה הבאה.

****

במידה והמשכורת הייתה גדולה יותר מ t1, לוקחים קבוע נוסף t2 ושואלים האם החוב גדול מ t2, אם לא אז כן משלם חשבונות. אם החוב כן גדול מ t2 אז ממשיכים לשאלה הבאה.

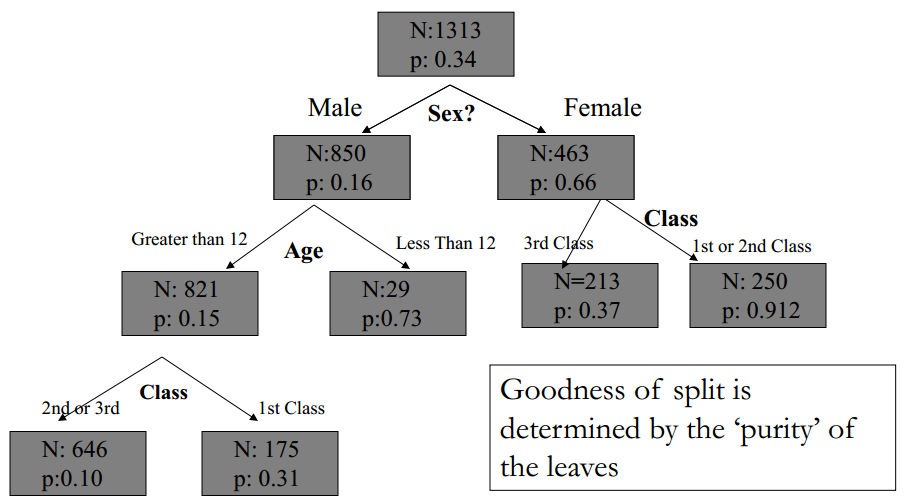
****

במידה והחוב היה גדול מ t2, לוקחים קבוע נוסף t3 ושואלים האם ההכנסה גדולה מ t3, אם ההכנסה לא גדולה מ t3 אז לא משלם חשבונות ואם ההכנסה גדולה מ t3 אז כן משלם חשבונות.

דוגמא: נתוני הטיטאניק

- 1313 נוסעים  
- 34% שרדו  
- האם זה היה מדגם אקראי?  
- או האם ההישרדות הייתה תלויים בתכונות של הפרט? מין, גיל, מחלקה

בפיצול הראשון תחליט מהו המשתנה הטוב ביותר כדי ליצור הפרדה בין השורדים לבין הלא שורדים.  
בדוגמא ניתן לראות שהפיצול הראשון נעשה בין גברים לנשים.



האינדוקציה של עצי ההחלטה

* אלגוריתם בסיסי (אלגוריתם חמדני - בחירת האפשרות הטובה ביותר הנראית לעין בשלב הנוכחי, מבלי לקחת בחשבון את ההשפעה של צעד זה על המשך הפתרון)

- העץ נבנה באופן רקורסיבי מלמעלה למטה.  
- בהתחלה, כל דוגמאות האימון נמצאות בשורש

-הדוגמאות מחולקות למחיצות באופן רקורסיבי כדי למקסם את הטוהר.  
- הערכת כל הפיצולים האפשריים. החלטת הפיצול הטובה ביותר, כלומר הפיצול עם רווח המידע הגבוה ביותר, נבחר כדי לחלק את הנתונים לשתי תת קבוצות (הפרד ומשול) והשיטה נקראת באופן רקורסיבי.

* תנאים להפסקת החלוקה:  
  - כל הדגימות שייכות לאותה מחלקה  
  - צומת העלה קטן יותר מהסף שנקבע  
  - איזון בין מורכבות להכללה

הרחבות של עצי הסיווג

* ניתן להשתמש בפיצולים שאינם בינאריים: multi way, נוטה להגדיל את המורכבות באופן משמעותי אך לא לשפר את הביצועים, פיצולים בינאריים ניתנים לפירוש אפילו לא ע"י מומחים, קל לחישוב ולדמיון.
* יכול גם לקחת בחשבון פיצולים של קומבינציה לינארית: יכול לשפר ביצועים חזויים, קשה לייעול.
* פונקציית הפסד: טעויות מסוימות יכולות להיות יקרות יותר מאחרות, ניתן לשילוב בחישוב ג'יני.
* עציםישניםרגיליםבדרךכללעובדיםדיטוב.
* העצים יכולים גם להיות בשימוש עבור רגרסיה כאשר התגובה היא בעלת ערך אמיתי: מועיל עם חיזויים קטגוריים, ניתן להשתמש בשונות כמדד טהור, חיזוי העלים הוא ממוצע הערך במקום הערכת הסתברות המחלקה.

למה עצים נמצאים בשימוש מורחב?

* העצים יכולים להתמודד עם נתונים ממדיים גבוהים (בונה מודל תוך שימוש בממד 1 בזמן)
* יכול להתמודד עם כל סוג של משתנים המוכנסים כקלט (חיזוי קטגורי)
* קבוע לשינויים מונוטוניים של קלטי המשתנים (שימוש ב x, 10x + 2, log(x), 2^x לא ישנו את העץ)
* העצים הם ניתנים לפירוש במידה מסוימת.
* אלגוריתם העץ קל לקידוד ולבדיקה

מגבלות של העץ

* קושי בדגימת מבנה לינארי
* חוסר בחלקלקות
* שונות גבוהה :

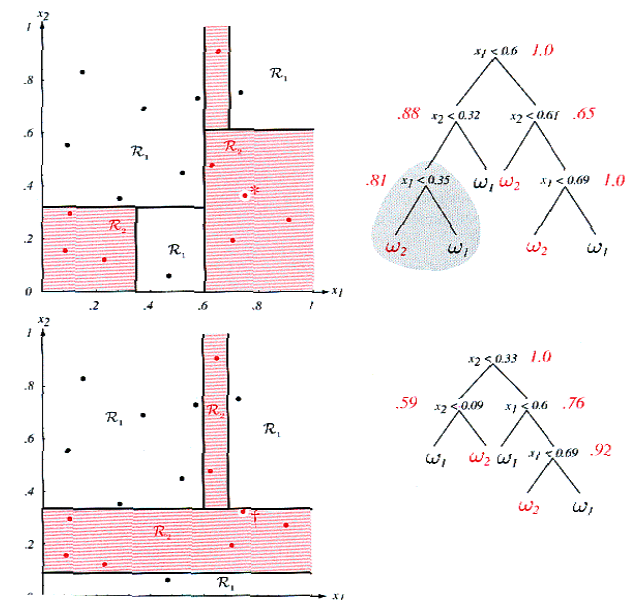
- עציםיכוליםלהיות "לאיציבים" כפונקציהשלהמדגם.

- גורם לשתי בעיות: 1. שונות גבוהה תורמת לשגיאות ניבוי. 2. שונות גבוהה מפחיתה את ההבנה של הפירוש.

* העצים הם מועמדים טובים למודלים משולבים.

עצי החלטה הם לא יציבים

העברה קצת של דוגמא אחת עלולה להוביל לעצים שונים לגמרי ומחיצת המרחב.

חוסר היציבות מול ההפרעות הקטנות של הנתונים

יערות אקראיים - Random Forests

צורה נוספת של עצים-  
העצים רגישים לפיצול הראשוני, מה שעלול להוביל את העץ לכיוונים בלתי הולמים / לא מתאימים. דרך אחת לראות את זה היא: התאם עץ למדגם אקראי או לנתונים של המדגם.

פתרון-  
- יערות אקראיים: שילוב של עצי החלטה  
- כל עץ בנוי על תת קבוצה אקראית של הנתונים מסט האימון.  
- בכל נקודת פיצול רק תת קבוצה אקראית של ניבויים נבחרים.  
- החיזוי הוא פשוט וברוב קולות של העצים

ליערות יש את היתרון של העצים, עם יותר חוסן וכלל החלטה חלק יותר.