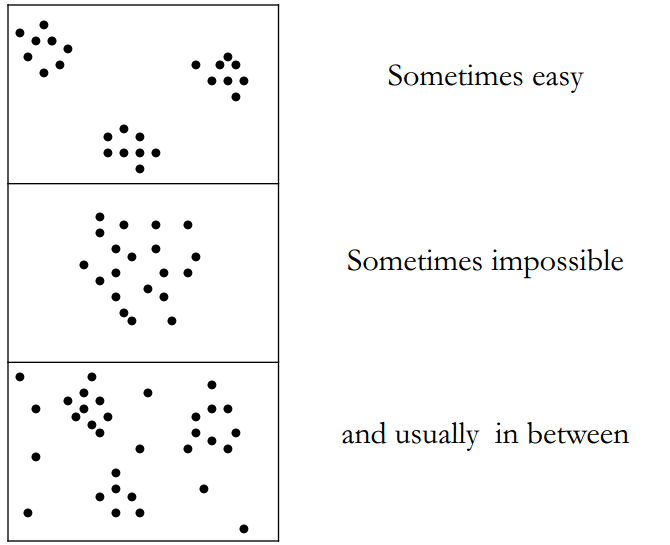
**Clustering and Unsupervised Learning**

[**למידה בלתי מונחית**](http://he.wikipedia.org/w/index.php?title=%D7%9C%D7%9E%D7%99%D7%93%D7%94_%D7%91%D7%9C%D7%AA%D7%99_%D7%9E%D7%95%D7%A0%D7%97%D7%99%D7%AA&action=edit&redlink=1) (unsupervised learning) - מטרת האלגוריתמים היא למצוא ייצוג פשוט וקל להבנה של אוסף הנתונים. שיטות נפוצות מסוג זה הן חלוקה לאשכולות (clustering) והטלה ליריעות ממד נמוך כגון Principle component analysis.

[**למידה מונחית**](http://he.wikipedia.org/wiki/%D7%9C%D7%9E%D7%99%D7%93%D7%94_%D7%9E%D7%95%D7%A0%D7%97%D7%99%D7%AA)**-**  (supervised learning) כל דוגמה מגיעה ביחד עם תווית סיווג. מטרת האלגוריתם היא לחזות את הסיווג של דוגמאות חדשות שאותן לא פגש בתהליך הלמידה. אימון של [רשת עצבית מלאכותית](http://he.wikipedia.org/wiki/%D7%A8%D7%A9%D7%AA_%D7%A2%D7%A6%D7%91%D7%99%D7%AA_%D7%9E%D7%9C%D7%90%D7%9B%D7%95%D7%AA%D7%99%D7%AA) ("רשת נוירונים")מסתמך על אלגוריתמים מסוג זה.

Clustering–אישכול

"זיהוי אוטומטי של מבנה קבוצה בתוך הנתונים."  
בדרך כלל: חלוקת N נקודות נתונים ל –K קבוצות (אשכולות –clusters), כך שהנקודות בכל קבוצה יותר דומות אחת לשנייה מאשר לנקודות בקבוצות אחרות.  
טכניקה תיאורית (לעומת חיזוי)  
זיהוי קבוצות "טבעיות" של אובייקטי נתונים– תיאור באופן איכותי קבוצות של הנתונים.  
עבור ערכי אמת של וקטורים, אשכולות יכולים להיחשב כעננים של נקודות במרחב ה p ממדי.  
נקרא גם unsupervised learning למידה בלתי מונחית

לעיתים קל לחלק לאשכולות

לעיתים בלתי אפשרי

ובדר"כ באמצע

מה הוא ניתוח אשכול?

* תוצאות טובות של ניתוח אשכול הן כאשר: דומים אחד לשני באותו האשכול ושונים מהאובייקטים באשכולות אחרים.
* במילים אחרות: דמיון גבוה בתוך האשכול (שונות נמוכה בתוך האשכול). דמיון נמוך בין האשכולות ( שונות גבוהה בין האשכולות)

מדוע האישכול שימושי?

* גילוי של ידע חדש מתוך הנתונים. מנוגדלסיווג מונחה (שבו התוויות ידועות). יכול להיות מאוד שימושי לסכמות של מערכי נתונים גדולים (עבור N גדול ו / או נפח גדול).
* יישומים של אישכול:
* www– אישכול של מסמכים המיוצרים ע"י מנוע חיפוש
* פילוח לקוחות
* ניתוח נתונים מרחביים – אשכולות גאוגרפיים של אירועים: שיעורי סרטן, מכירות וכו'.
* אישכול של גנים עם פרופילי הבעה דומים.

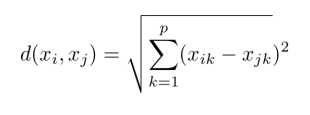
נושאים כלליים באישכול

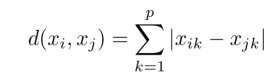
אין אמת שהיא זהב – התשובה לפעמים סובייקטיבית  
ייצוג האשכולות – אילו סוגים או צורות של אשכולות אנו מחפשים? מה מגדיר את האשכול?  
פונקציית המרחק D[x(i),x(j)] היא היבט קריטי של הקיבוץ באשכולות: מרחקשלזוגותבודדיםשלאובייקטים, מרחקשלאובייקטיםבודדיםמאשכולות.  
איך K נבחר?

Distance Measures– מדדי המרחק

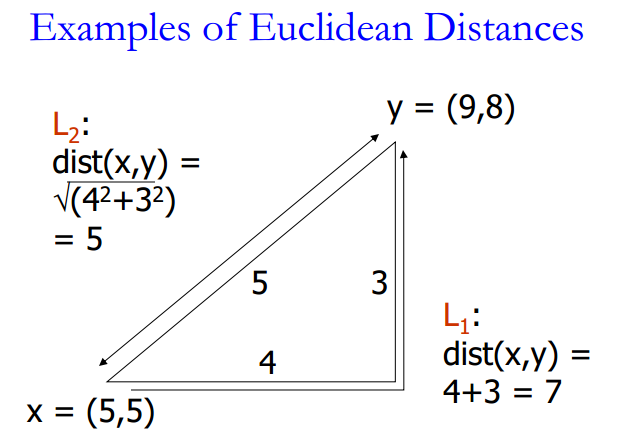
במטרה לקבץ לאשכולות, אנו זקוקים לסוג של "מרחק" בין הנקודות.  
לעיתים המרחקים לא ברורים אבל אנו יכולים ליצור אותם.

Some Distances– מרחקים מסוימים

* Euclidean distance (L2 ): d(x,y) - הרעיון הנפוץ ביותר של המרחק



* Manhattan distance (L1 )– שיטה למרחק רק אם אתה נאלץ לנסוע רק לאורך הקורדינטות.

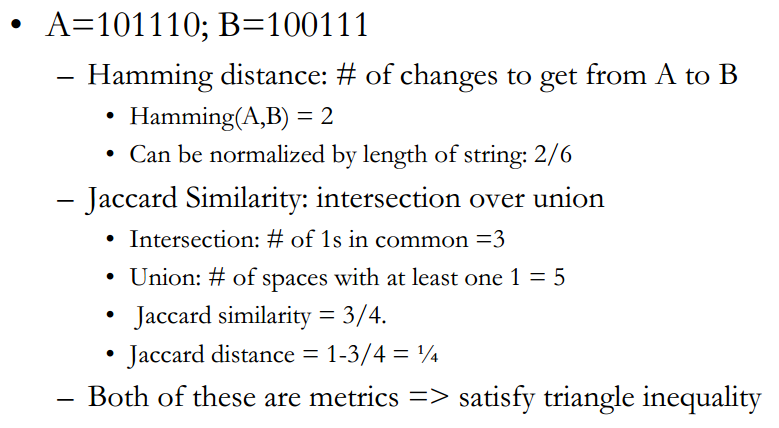


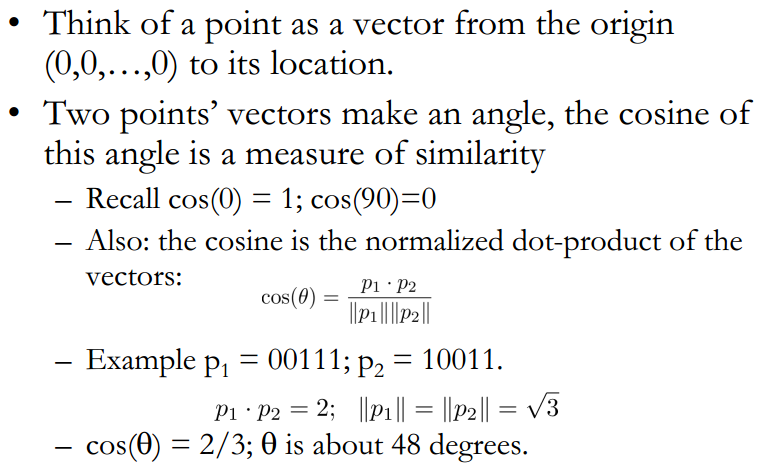
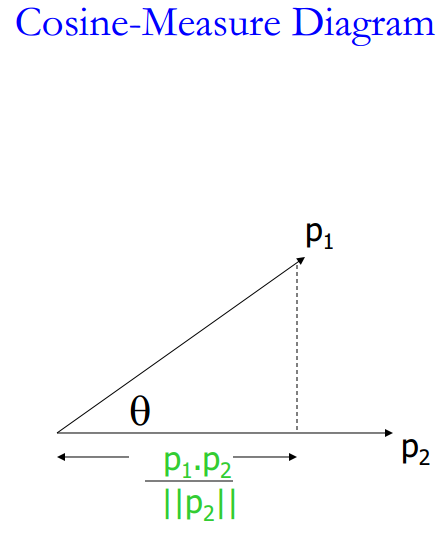
Non-Euclidean Distances– מרחקים לא אוקלידים

תצפיות מסוימות אינן מתאימות למרחק אוקלידי:  
- וקטורים בינאריים: 10011 לעומת 11000  
- strings: sadistics לעומת Statistics  
- Ordinal variables: M.S לעומת B.A  
- קטגוריים: blue לעומת green

איך נחשב מרחק עבור משתנים כאלו?

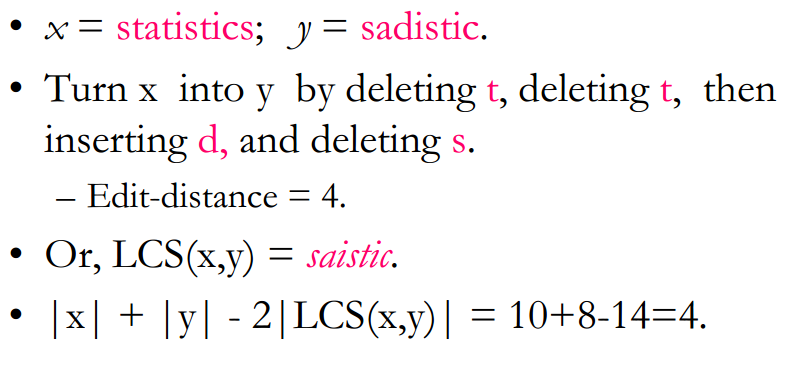
מרחקים עבור וקטורים בינאריים

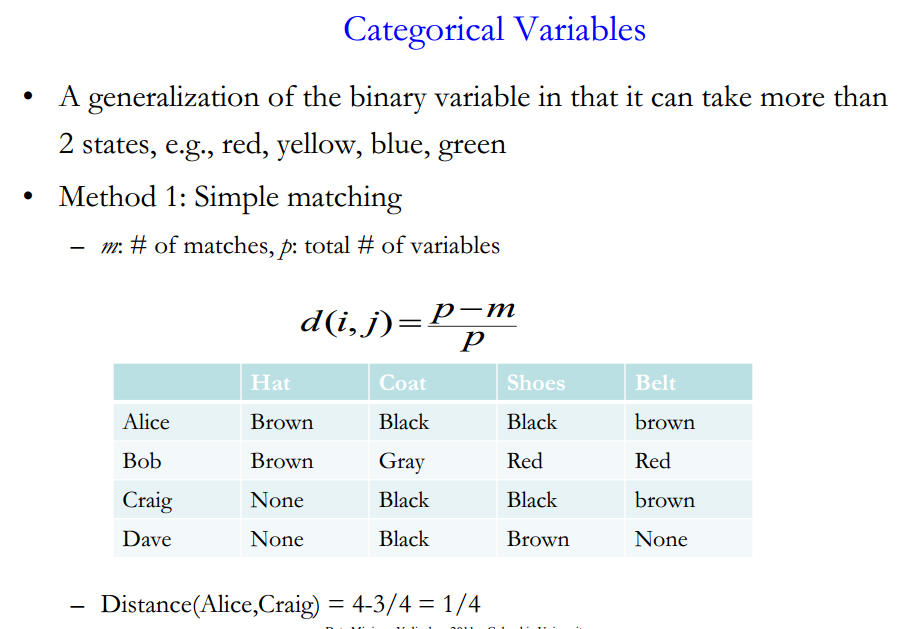


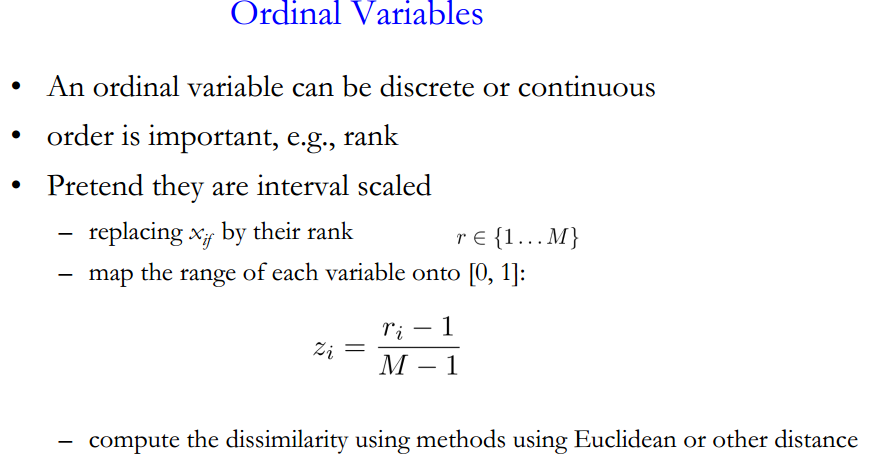
מרחק הקוסינוס

עריכת המרחק עבור מחרוזות

מרחק המינג (Hamming) למחרוזות: מספר התווים המוכנסים והמחוקים צריכים להסתובב אחד לתוך השני.  
באופן שקול: d(x,y) = |x| + |y| - 2|LCS(x,y)|  
LCS - longest common subsequence– המחרוזת הארוכה ביותר שהתקבלה ע"י מחיקה מ-x ומ-y.

דוגמא:



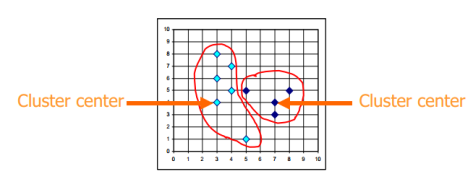


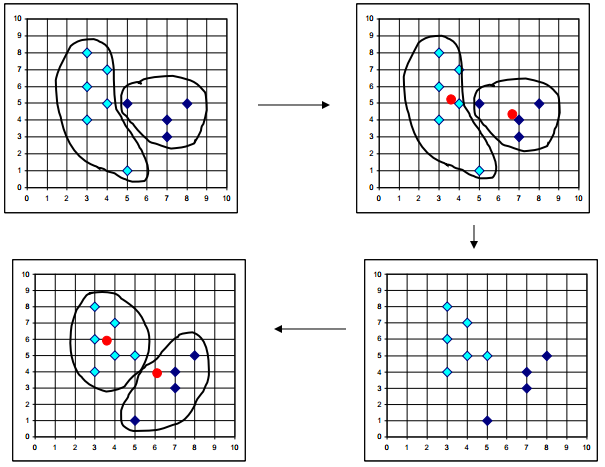
שיטות האישכול - Clustering Methods

יש לנו עכשיו מטריצה (n X n) של מרחקים.

ישנם 2 סוגים עיקריים של אלגוריתם האישכול:  
1. חלוקה (partitioning) – חלוקת הנתונים לאשכולות עם גבולות מוגדרים. להציב כל נקודה באשכול הקרוב אליה.  
2. היררכיות (hierarchical) – כל נקודה באשכול משלה, שילוב איטרטיבי. מחלוקת: כל הנתונים באשכול אחד, איטרטיבי לניתוח.

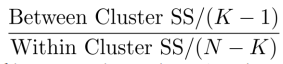
k –Means Algorithm(s)

* הנח שהמרחב הוא אוקלידי
* התחל ע"י בחירת k, מספר האשכולות (קבע מס'k מרכזים באקראי / לא באקראי)
* אתחל את האשכולות ע"י בחירת נקודה אחת לכל אשכול (בדרךכלל ישkנקודותאקראיות)
* באופן שרירותי תבחר k אובייקטים מהנתונים להיות מרכז של כל אשכול.
* לכל אחד מהאובייקטים שנותרו: הקצה כל אובייקט לאשכול שהמרכז שלה הכי קרוב אליה.
* חזור על שלושת הצעדים הבאים עד שהאשכולות יתקבצו(ללא שינוי באשכולות):

1. חישוב המרכז החדש של האשכולות הנוכחיים
2. להקצות כל אובייקט לאשכול שהמרכז שלה הכי קרוב אליה
3. חזור לשלב 1 או הפסק אם המרכז לא משתנה

דוגמא:

להשיג k נכון

- קשה, לפעמים נעשה באופן סובייקטיבי  
- נסה k שונים, הסתכל על השינוי בממוצע המרחק של נקודת המרכז ככל ש k גדל.  
- חפש אחר איזון בין השונות שבתוך האשכול לבין השונות שבין האשכולות.  
Calinski Index–

- הממוצע יורד במהירות עד ש k נכון, לכן תשנה במעט את ה k.

תגובות על שיטת k-means

חוזקות:  
- יחסית יעיל: קל ליישום , לעיתים מגיע עם תוצאות טובות אם לא הטובות ביותר  
- אינטואיטיבי

חולשות:  
- צריך לציין k, מספר האשכולות מראש  
- לא יכול להתמודד עם בנתונים רועשים או בחריגים  
- אינו מתאים לגלות אשכולות עם צורות שאין קמורות  
- די רגיש לנקודות המאותחלות בהתחלה. עשה זאת מספר פעמים וראה כיצד התוצאות משתנות.

וריאציות על k-means

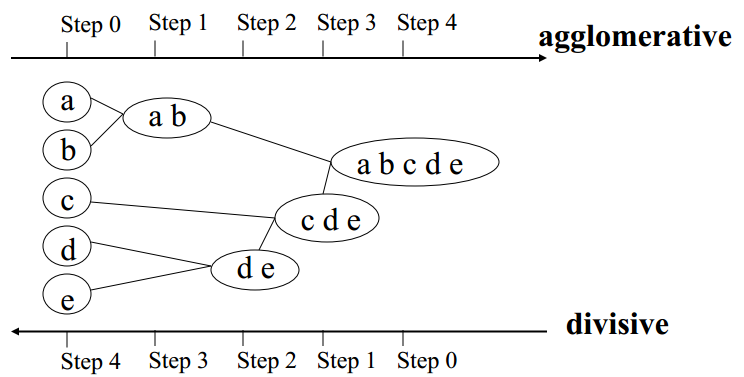
עשה זאת טוב יותר ע"י שימוש ב k-modes או k-mediods.  
k-mediods: mediodsזה האובייקט הממוקם במיקום המרכזי ביותר באשכול

הפוך את האתחול לטוב יותר:  
- קח מדגם קטן אקראי וקבץ ממנו אשכול כדי למצוא נקודת התחלה  
- תבחר k נקודת על ברשת  
- עשה מספר ריצות עם נקודות התחלה שונות

Hierarchical clustering

אינו דורש את מספר האשכולות k כקלט.  
שתי קצוות: כל המידע באשכול אחד. כל נקודת נתונים באשכול משלו.

סוגים של אישכול היררכי:  
- **Agglomerative** (bottom up) **clustering**– בונה את ה dendrogram מהרמה התחתונה וממזג את הזוג הדומה ביותר (או הקרוב ביותר) של האשכולות. עוצר את התהליך כאשר כל נקודות המידע מוזגו לאשכול יחיד.  
- **Divisive** (top down) **clustering**– מתחיל עם כל נקודות הנתונים באשכול אחד,השורש. מפצלאתהשורשלתוךמערכתשלאשכולותילדים. כלילד אשכולמחולקבאופןרקורסיבינוסף. עוצר כאשר נותרו אשכולות בודדים של נקודות נתונים אישיות. כלומר, כל אשכול עם נקודה אחת בודדת.



* ייצוג: עץ של אשכולות מקוננים.
* אלגוריתם גרידי: מצא שתי נקודות דומות ביותר, חבר אותם, חזור על השלבים שוב.
* יכול גם לרוץ אחורה: divisive
* הדמיה אפקטיבית ע"י dendrogram
* האלגוריתם דורש את מדד המרחק עבור מרחקים בין אשכולות או בין נקודה ואשכול.

מרחקים בין אשכולות

קישור יחיד: המרחק הקטן ביותר בין הנקודות. השכן הכי קרוב. יכול להיות רגיש לחריגים.

קישור מלא: המרחק הגדול ביותר בין הנקודות. אוכף את הקומפקטיות.

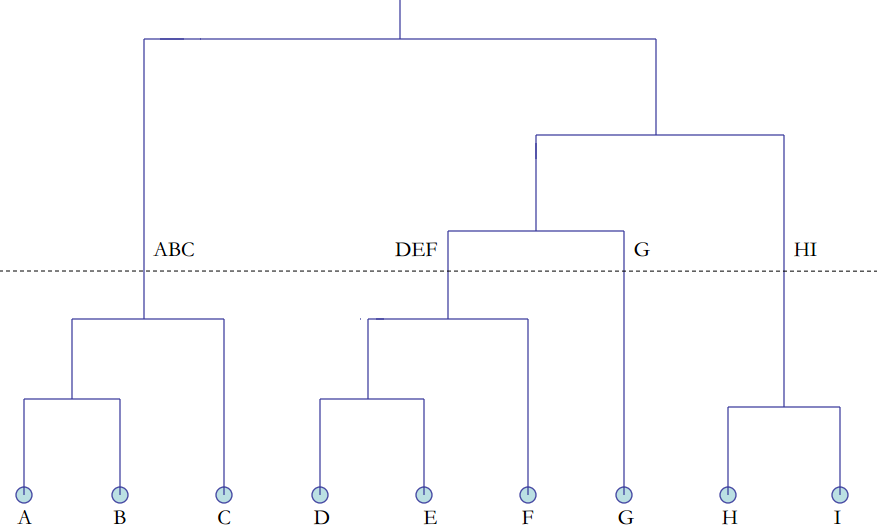
קישור ממוצע: ממוצע – מקבל את התנהגות הממוצע.נקודת המרכז חזקה יותר

המדד של וורד: מיזוג אשכולות שמקטינים את העלייה במרחקים בתוך האשכול:

D(i,j) = ( (SS(Ci+j) - SS(Cj ) - SS(Ci ) )

Dendrograms

ע"י חיתוך Dendrograms ברמה הרצויה ואז כל רכיב מחובר יוצר אשכול.



יתרונות:

- לאצריךלצייןkמראש  
- ייצוגחזותישלאשכולשונים  
- מאפייניםמdendogram

חסרונות:

- אפשרויות הצמדה שונות מקבל תוצאות מאוד שונות.