- מצגת 1.1 -

**תוכנות לכריית נתונים**

* תוכנה

ניתן להשתמש בכל תוכנה שאתה אוהב– חייב לדעת איך להכניס קלט, לתפעל, להכין גרף, ולנתח נתונים.

התוכנה המועדפת: R

תוכנות נוספות: sas, weka, spss, syntat וכו'

תוכנה לא מומלצת: excel, c

* מה זה R?

תוכנה סטטיסטית בעלת קוד פתוח שגדלה מ S/SPlus

מופץ הרבה ע"י המשתמשים ב-CRAN

קהילת משתמשים פעילה, מועילה

סביבת העבודה של R זמינה ברשת

גרפים מעולים

**מה זה כריית נתונים?**

לא מוגדר היטב..

אף אחד לא יכול להסכים על מה זה כריית נתונים! למעשה, למומחים יש תיאורים שונים מאוד:

"מציאת מבנה מעניין במאגרי מידע" - Fayyad, Chaduriand

"התהליך הלא טריוויאלי של זיהוי של זיהוי מבנים במידע חוקי, חדש, שימושי" –Fayyad

"תהליך של גילוי ידע של חילוץ מידע שלא היה ידוע, מעשי ממאגרים גדולים של מידע" –Zorne

"תהליך שעושה שימוש בכלים שונים של ניתוח נתונים כדי לגלות דפוסים ויחסים בנתונים שניתן להשתמש בהם כדי להפוך את התחזיות לתוקף" –Edelstein

מ –Zaiane:

כריית נתונים, הידוע גם בכינויו גילוי ידע במאגרי מידע

תהליך גילוי ידע במסדי נתונים מורכב מכמה צעדים מובילים מאוספי נתונים גולמיים לצורה כלשהי של ידע חדש.התהליך החוזר והנשנה מורכב מהשלבים הבאים:

- ניקוי נתונים

- שילוב נתונים

- בחירת נתונים

- שינוי נתונים

- כריית נתונים: זהו הצעד המכריע בטכניקות המתוחכמות המיושמות כדי לחלץ דפוסים שעשויים להיות שימושיים.

- הערכת דגם

- ייצוג הידע

מה הסמכות אומרת?

כריית נתונים היא התהליך של מציאת דפוסים סמויים מן הנתונים.

כריית נתונים היא התהליך של גילוי דפוסים חדשים ממערכות נתונים גדולות הכוללות שיטות של סטטיסטיקה ובינה מלאכותית, אבל גם ניהול מסדי נתונים.

Hand, Mannila, Smyth:

"כריית נתונים הוא הניתוח של נתונים תצפיתיים כדי למצוא מערכות יחסים בלתי צפויות ולסכם את הנתונים בדרכים חדשות, כי הן מובנות ושימושיות לבעלי המידע"

זה לא כמו סטטיסטיקה?

כריית נתונים לעומת סטטיסטיקה

Snark:סטטיסטיקה + שיווק = כריית נתונים

סטטיסטיקה ידועה ב:

השערות מוגדרות היטב שמשתמשים בהם כדי ללמוד על אוכלוסייה ספציפית שנבחרה, לומדים באמצעות נתונים שנאספו בקפידה ומתן מסקנות.

כריית נתונים לא זהירה כזו. היא:

גילוי נתונים מונעים ממודלים ודפוסים ממאגרי מידע מסיביים ותצפיתיים.

הסטטיסטיקה המסורתית:

קודם השערות, אח"כ לאסוף מידע, ולאחר מכן ניתוח לעיתים מונחה מודל. מתמקד בהבנה.

כריית נתונים (גם מכונת למידה):

מעט השערות מראש אם יש

המידע בדר"כ כבר נאסף מראש

הניתוח הוא בדר"כ תלוי נתונים ולא תלוי השערות

לעיתים קרובות algorithm-oriented יותר מ- model-oriented

מתמקד בניבוי

אבל –

רעיונות סטטיסטיים מאוד שימושיים בכריית נתונים. למשל, באימותאםהידעשהתגלהשימושי.

הגברתהחפיפהבגבולשלסטטיסטיקהוDM.

תרבויותיכולותללמודאחדמהשני

חזקמאודבעתשימושיחד

**Data Mining Enablers**

* פיצוץ של נתונים
* חישוב ואכסון מהיר וזול: - העיבוד מכפיל את עצמו כל שנתיים – האחסון מכפיל את עצמו כל 9 חודשים – טכנולוגיית מסד נתונים
* לחץ תחרותי בעסקים – למידע יש ערך, ההצלחות מקבלות פרסום נרחב.
* מוצרים תחרותיים - SAS, SPSS, Google Analytics, IBM, Oracle

מוצרי קוד פתוחים –R, WEKA

**גילוי מונע נתונים - Data-Driven Discovery**

נתונים תצפיתיים:

זול יחסית לנתונים ניסיוניים. דוגמאות: חנויות קמעונאיות, חברות תעופה וכו' אמזון, גוגל וכו'.

הגיוני למנף נתוניםזמינים ותצפיתיים.

מה הן הסכנות של נתוני תצפיות?

קל לעשות ניסויים מדומים.

נתונים תצפיתיים גם יכולים לעזור בגיבוש השערה.

**כריית נתונים: מפגש של דיסציפלינות רבות**

DataMining

Database

Technology

Statistics

Other

Disciplines

Information

Science

Machine

Learning

Visualization

לתחומים שונים יש תצוגות שונות של מה היא כריית נתונים.

**סוגים של נתונים:**

**קובץ שטוח או נתוני וקטור**

****

n

p

שורות = אובייקטים, תצפיות

עמודות = מדידות על אובייקטים

מציג כל שורה כ- pוקטורממדי, איפה ש- p היא ממדית, למעשה, הטמעת האובייקטים שלנו ב- p, מרחב וקטורים ממדי.

גם n וגם p מאוד גדולות ב DM.

מטריצה יכולה להיות דלילה.

**נתוני טקסט –** יכול להיות מיוצג כמטריצה דלילה

**טרנזקציות נתונים –**אירועי תאריך מוטבע (בלוגים, שיחות טלפון), יכול להיות מיוצג כסדר תזמן.

**נתונים יחסיים –**רוב קבוצות נתונים גדולות מאוחסנות במערכות נתונים יחסיים,שפת שאילתה נתונים מיוחדת: SQL.

**נתוני סדרות זמן –**סדרה לעתים קרובות הרבהזמן, סדרת זמן רב, או סדרה רב-משתנית זמן.

**נתוני E-bay**

**נתוני תמונה**

**דוגמאותלהצלחותשל כרייתנתונים**

סלשוק (Walmart)

מערכות המלצה (Amazon.com)

איתור הונאה בתקשורת (AT & T)

יעדשיווק / CRM

שווקיםפיננסיים

DNA Microarrayניתוח (אושזה?)

תעבורתאינטרנט / ניתוחבלוג

גוגל- חברה שנבנתה על כריית נתונים, Page Rankממוקש באינטרנט כדיל בנות חיפוש טוב יותר,בודק איות, זיהוי פנים, צובר חדשות, מקום למודעות.

**תהליך כריית נתונים**

נקרא לעתים קרובות KDD(Knowledge Discovery in Databases) – גילוי ידעב מאגרי מידע.

הניתוח הוא רק חלק אחד מהתהליך.

איסוף ואחסון נתונים

ניקוי הנתונים

דגימת הנתונים

ניתוח

קבלתהחלטות

**משימות שונות לכריית נתונים**

חקר ניתוח הנתונים

מודלים תיאוריים

מודלים חיזויים

גילוי דפוסים וחוקים ועוד..

**חקר ניתוח הנתונים - Exploratory Data Analysis**

לפני המודל צריך לעשות-

חייב לבדוק את הנתונים – חישוב סטטיסטיקות סיכום: טווח, מקסימום, דקות, ממוצע, חציון, שונות, צידוד. ערכים חסרים, תצפיות חריגות. לבדוק אילו סוגי משתנים יש לך.

ההדמיה נעשית בשימוש נרחב: היסטוגרמות 1D, מגרשי פיזור 2D. שיטות ממדיות גבוהות.

חקר ניתוח פשוט יכול להיות יקר מאוד. תמיד "תסתכל" על הנתונים שלך לפני יישום אלגוריתם כלשהו של כריית נתונים.

**מודלים תיאוריים - Descriptive Modeling**

המטרה היא לבנות מודל תיאורים, לדוגמא: מודל שיכול לדמות את הנתונים במידת הצורך.

בניית מודל לתהליך הבסיסי.

דוגמאות: הערכת צפיפות, ניתוח אשכולות, מודלים תלותיים בין משתני p.

**מודלים חיזויים - Predictive Modeling**

חיזוי שמשתנה אחד y נותן קבוצה של משתנים אחרים x. כאן x יכול להיות p וקטור ממדי.

סיווג (Classification):Yהואקטגורי. רגרסיה: (Regression): Yהוא ערך אמיתי.

למעשה זהו קירוב פונקציה, ללמוד את מערכת היחסים בין Y ל- X.

בכריית נתונים, הדגש הוא על דיוק ניבוי, לא בהבנת המודל

איתור הונאה-

זיהוי הונאת תקשורת – הונאות עולות לחברות כמיליון דולר לשנה, מעט עסקאות הן הונאה אבל הן יקרות. הגישה: לכל עסקה מאמידים הונאה, בהתבסס על הונאה ידועה או התנהגות משתמש ידועה.

דוגמאות למודלים: פרופיל שימוש בכרטיסאשראי, שינוי בהתנהגות, אשמה על ידי עמותה.

**גילוי דפוס - Pattern Discovery**

מטרה היא לגלות דפוסים "מקומיים" מעניינים בנתונים ולא כדי לאפיין את הנתונים בעולם.

לפי נתונים של השוק אנו יכולים לגלות: אם לקוחות קונים יין ולחם ולאחר מכן הם קונים גבינה בהסתברות 0.9, זה ידוע כ"חוקי עמותה". זו הדרך שבה כריית נתונים נולדה. אסטרונומיה, כלכלה.

**חסרונות בכריית נתונים**

האם כריית נתונים תמיד הכרחית – רק בגלל שיש לך טרה-בייט לא אומר שאתה צריך להשתמש בו.

ענייני פרטיות – שונה על ידי מדינה, תעשייה, יישום, דור.

משמעות של דפוסים ברורים – פרדוקס הריין, טרור, ל - DMיש הרבה מה ללמוד מסטטיסטיקה!

- מצגת 1.2 –

**כריית נתונים**

ניתוח הנתונים מונעים לחשוף דפוסים אחרים לא ידועים אך שימושיים במערכי נתונים גדולים, לגלות ידע חדש ולפתח מודלים חיזוי, הפיכת נתונים ומידע לידע וחוכמה (יום אחד אולי), באופן ובמועד.

זהו יישום של:

למידת מכונות (Machine Learning)

סטטיסטיקה

הנדסת תוכנה ותכנות עם הנתונים

תקשורת ואינטואיציה יעילה

למערכי נתונים המשתנים על ידי נפח, מהירות, גיוון, ערך, מהימנות

לגלות ידע חדש

לשפר תוצאות עסקיות

לספק שירותים טובים יותר מותאמים

אלגוריתמים בסיסיים של כריית נתונים

ניתוח אשכולות - Cluster Analysis

אגודת ניתוח - Association Analysis

Linear Discriminant Analysis

Logistic Regression

Decision Trees

Random Forests

Boosted Stumps

Neural Networks

Support Vector Machines

מיומנויות R מושכות משכורות טובות

בחירה ב –R למדעי הנתונים

Rהוא זמין באופן חופשי לכל לניתוח נתונים

ל –Rיש את החבילה המקיפה ביותר של פונקציונליות זמינה

שום דבר אחר הוא יותר אפילו קרוב

למה צריך ממש משתמש למדע הנתונים ב- R?

סטטיסטיקה יכולה להיות מורכבת וממולכדת

ב-R כלים רבים כדי לספק תובנות

ניתוחים יעילים צריכים להיות מבוססי script

Scriptingנדרש גם להדירות

R היא שפה כדי לתכנת עם הנתונים

משתמשים של Ratlle:

היום, Ratlle נמצא בשימוש בעולם הרחב בתעשיות רבות: ניתוחי בריאות, פילוח לקוחות ושיווק, איתור הונאות, ממשלה.

הוא משמש ב: אוניברסיטאות ללימוד כריית נתונים, עבודות מחקר לניתוחים בסיסיים, יועצים ואנליסטיים.

זהוי ישאר זמין באופן חופשי.