**הרצאה 1**

בעיות שפותרים בלמידת מכונה:

**Supervised Learning-**מתאמנים על נתונים שאנו יודעים מה אמור להיות הפלט שלהם

**סיווג**- סיווג נתונים לקבוצות שונות (לזהות כלבים או חתולים בתמונות שונות, האם לתת לאדם משכנתה)

X- הנתונים של שאנו יודעים על הפרט

Y- האופציות שהמערכת יכולה לסווג אליהן (לתת, לתת עם ריבית גבוהה, לא לתת).

לאחר שביצענו את הלמידה נקבל אשר מאפשר לנו לדעת עבור כל פרט לאן לסווג אותו.

את הלמידה מבצעים על נתונים של פרטים שאנו יודעים מראש לאיזה Y הם מסווגים.

נדרש כי Xs שהשתמשנו כדי ללמוד מהם יהיו דומים לXs שאותם נבדוק בעתיד.

הנתונים (הXs) יכולים להיות עם נטייה מסוימת, כיוון שאלו החלטות שביצעו בעבר בני אדם, כמו אפלייה מגדרית או גזענות, לכן צריך לבדוק שהנתונים שאנו משתמשים בהם הם מייצגים את כל טווח הקלטים שאנו רוצים לבדוק.

האלגוריתם שמתפתח מהלמידה הוא קופסא שחורה, אנחנו לא יודעים מה קורה בו, אלא יודעים רק מה הקלט שמכניסים ומה הפלט.

**שערוך**- אותו דבר כמו סיווג רק עם מספרים ממשיים, התוצאה היא ממשית, לא מספר מוגבל של אופציות(משתנה דיסקרטי), כמו הערכה של מחיר של בית מסוים, להעריך מה המשכורת החודשית של בן אדם לפי מה שהוא קונה בסופר (ניתן למכור את המידע לחברות אחרות אשר יכולות לעשות בו שימוש).

**Prediction**- להעריך מה יקרה בעתיד, לדוגמה להעריך אם לקוח עומד לעזוב את החברה, באמצעות בדיקת תיק הלקוח.

**-Affinity grouping** להעריך מנתונים על פרטים מסוימים מה סביר שגם מאפיין אותם. לדוגמה אנשים ששייכים לקבוצות מסוימות בפייסבוק, ככל הנראה יתעניינו בקבוצת אחרות או יהיו כבר שייכים אליהן.

**Unsupervised Learning**-לא יודעים מה אמור להיות הפלט

**Clustering-** לקבץ את הפרטים לקבוצות לפי תכונות(Xs) דומות. מקבלים כפלט את סוגי הקבוצות השונות שיש בנתונים.

**Description and profiling-** גרסה 1: מתאר מה מהפרטים הם במצב נורמלי ומה במצב חריג, גרסה 2: מעבר לסיווג מה נורמלי ומה חריג מסביר גם מה גורם לחריגות. (לדוגמה שמירה על תקינות של מספר שרתים).

לאחר הרצת האלגוריתם ניתן לנסות להבין מה הוביל לתוצאה שקיבלנו, מה הם הגורמים המשמעותיים מהגורמים שאנו נותנים.

למה היום יש שימוש גדול בלמידת מכונה?

היום נשמרים מאגרים מאוד גדולים של מידע בעלות יחסית קטנה, יש עניין גדול בשימור לקוחות, לכן משתמשים במידע שנאסף על הלקוח כדי לשפר את השירות והפרסום שנותנים לו ספציפית אליו (CRM), מידע היום משמש כמוצר שניתן לסחור בו בין חברות, היום כוח מחשוב הוא יותר זמין וזול, קיימות תוכנות זמינות של data mining.

**Data mining-**השם מתחום הסטטיסטיקה לmachine learning. זה אנליזה ועיבוד של מידע בכמויות גדולות, על מנת למצוא דפוסים משמעותיים.

עיבוד של מידע raw על מנת להבין מערכות יחסים עסקיות.

יש כל מיני גישות לdata mining ומספר טכניקות שונות שניתן להשתמש בהם בנפרד או בצורה משולבת.

דוגמה בCRM:

באמזון ממליצים לאדם ספרים לפי התחום ספרים שהוא קנה בעבר (classification), או בודקים איזה אנשים קנו גם את הספרים שהוא קנה וממליצים לו את הספרים האלו (Affinity groups).

על מנת ליצור מערכת יחסים של למידה של חברה עם הלקוחות שלה נדרש:

**לשים לב** מה הלקוח עושה

**לשמור** מה הלקוח עשה בעבר

**ללמוד** ממה שהלקוח ביצע בעבר

**לפעול** על פי מה שנלמד

לדוגמה בגוגל בודקים מה חיפשת בעבר, אילו אופציות חיפוש בחרת, כמה זמן עבר מהבחירה שלך לבחירה הבאה, לפי זה ניתן להמליץ על מילות השלמה לחיפוש ולדרג תוצאות חיפוש לפי יעילות.

נתונים על פי טרנזקציות- איסוף נתונים על פי הקניות שבוצעו, השיחות שבוצעו, ההעברות שבוצעו.

זיהוי ושמירה של יחסים בין נתונים הוא המפתח לאלגוריתם data mining יעיל.

היסטוריית data mining

התחיל לפני יותר מ40 שנים, בעיקר בתחום הסטטיסטיקה עם אלגוריתם k-means. עם הזמן התקדם לתחומים שונים. בשנות ה80 פותחו אלגוריתמים של רשתות נוירונים.

סוגי data mining

**Directed(supervised)-** ניסיונות לקטלג או להסביר נתונים למטרה מסוימת מוגדרת מראש, עם קבוצת נתונים מסווגים מראש.

**Undirected(unsupervised)-** ניסיונות למצוא דפוסים ודמיון בין קבוצות של נתונים בלי שימוש במטרה מסוימת או קבוצה של נתונים שכבר מסווגים.

האתגרים שעומדים בפני data mining

* התמודדות עם כמויות אדירות של מידע וקושי לסכם אותם למידע רלוונטי
* לבחור באיזה נתונים להשתמש
* לקבל תוצאה סופית שיש לה שימוש מסוים

**שימוש בשאלונים על מנת לאסוף מידע על לקוחות:**

שואלים שאלות רלוונטיות על מנת לקבל מידע על הלקוחות

הבעיות:  
-מדגם לא תמיד מייצג טוב את כל הלקוחות

-לקוחות עלולים להיות לא כנים ולתת מידע לא נכון

- פעולה מסוימת יכול לקרות ממספר סיבות שונות

מעגל הפעולות של Data Mining

1. **זיהוי ההזדמנות העסקית**- תהליכים עסקיים שונים הם אופציות טובות אפשריות, לבדוק מה התבצע מאמצים קודמים בData mining
2. **כריית מידע כדי להפוך אותו למידע שניתן לבצע איתו פעולות**- הצלחה היא להפוך את המידע לשימוש להיגיון עסקי. בעיות שיכולות להיות: פורמטים לא טובים לנתונים כמו ערכים חסרים, שדות נתונים מבלבלים כמו 2 שדות שונים שהם שמות נרדפים לאותו הדבר, חוסר בפונקציונליות עם המידע, הגבלות משפטיות (פרטיות), גורמים בארגון, כמו חוסר רצון של הארגון להשתנות, בעיות עמידה בלוחות זמנים- לדוגמה לקוח קיבל בטעות תוכנית יקרה יותר ממה שהוא אמור לשלם בצורה אוטומטית, לכן נדרש ליצור איתו קשר מספיק בזמן כדי להודיע על הטעות ולשנות כדי שהלקוח לא יעזוב.
3. **לבצע פעולות באמצעות המידע**-לשנות את האינטראקציה עם הלקוחות \ ספקים, לשנות את רמות המלאי של מוצרים שונים לפי הביקוש וההיצע בתקופות שונות, שינוי תהליכי שירות, התרחבות של העסק
4. **למדוד את התוצאות**-למדוד מה השתנה בעקבות הפעולות שהתבצעו, האם השינוי הביא לתוצאה הרצויה