*מבוא למדעי הנתונים – Introduction to Data Science:*

*סיכום כפיר גולדפרב*

**שיעור ראשון -** *מבוא לסטטיסטיקה והסתברות:*

בחבילת קלפים יש 52 קלפים רגילים ו-2 ג'וקרים (54 סה"כ), מה ההסתברות שאחרי ערבוב שרירותי של כל הקלפים יהיו 2 ג'וקרים בראש החבילה?

* בהנחה שהם דומים (לא ניתן להבדיל בניהם) .
* בהנחה שהם קלפים שונים: .

אם אטיל קוביה איסוף פעמים הסיכוי שייצא מה שרציתי חחלך ויגדל לפי הנוסחה: .

כאשר הוא מאורע ו- הוא הסתברות כלומר פירושו הוא מה ההסתברות שמאורע יקרה?

**אקסיומות ההסתברות:**

1. , אנו תתופסים כי הסתברות של מאורע הוא מספר ממשי בין 0 ל-1.
2. , אנו מצפים שלפחות אחד מהאירועים הבסיסיים שבמרחב המדם יתקיים תמיד – בכל ניסוי שנערוך, – מרחב המדגם, קבוצת כל התוצאות האפשרויות בניסוי, למשל בהטלת מטבע יש שני אפשרויות עץ / פלי.
3. , הסתברות שמאורע לא יתקיים שווה ל-1 פחות המשלים – ההסתברות שכן יקרה.
4. , כאשר המאורעות זרים מתקיים .

**דוגמה:**

בקוביה , - הסיכוי לקבל 2 או 3, - הסיכוי לקבל מספר זוגי.

*כלומר , , הסיכוי לקבל מספר שיצא בשני הקבוצות .*

1. *חיסור / הפרש מאוררעות: המאורע פחות המאורע – הוא המאורע שכולל את כל התוצאות האפשרויות של מאורע ולא של מאורע , סימון: .*

***דוגמה:***

*נתון: .*

***שאלה:***

*בתהליך ייצור של מפעל קיימים 2 סוגים של פגמים: ו- , הסיכוי ש- יתרחש בתהליך הייצור הוא 0.1, הסיכוי ש- יתרחש בתהליך הייצור הוא 0.2, הסיכוי שגם וגם יתרחשו הוא 0.05,*

*לסיכום נתון: .*

*חשבו את ההסתברויות:*

1. *לפחות פגם אחד.*
2. *יש פגם ואין פגם .*
3. *אין פגם בכלל.*
4. *יש בדיוק פגם אחד.*

*פתרונות:*

1. *נחשב את .*
2. *נחשב את ,*

*ניתן גם לחשב את .*

1. *נחשב את כלומר את , נקבל .*
2. *נחשב את .*

**תרגיל:**

מושכים מחבילה 3 קלפים, חשבו את ההסתברות:

1. כל הקלפים הם לב:

בלי חזרות , עם חזרות: .

1. *אף קלף אינו לב: .*
2. *כל קלפים הם אסים: .*

**שיעור שני –** *המשך מבוא לסטטיסטיקה והסתברות:*

**הסתברות מותנת:**

*התסברות מונתנת הוא הסתברות של מאורע כלשהו , ונשאלת השאלה האם התנון של מאורע משנה את את האינפורמציה ואז ההסתברות של משתנה בהתאם.*

**חוק בייס - :**

חוק בייס או נוסחת בייס הוא תוצאה בתורת ההסתברות המאפשרת לחשב הסתברות מותנת של מאורע, הנוסחה:

הסתברות של בההתן אינה בהכרח שווה להסתברות של בהנתן ולכן אינה סימטרית.

**דוגמה:**

הסתברות לקבל בקוביה 6: , הסתברות לקבל בקוביה מספר זוג: ,

אם אומרים לי שהתוצאה יצאה זוגית אז ההסתברות שקבל 6 הוא , כי מרחב המדגם שלי השתנה רק למספרים זוגיים.

נשים לב כי ,

ניקח את ההסתברות המותנת של בהנתן לפי נוסחת בייס:

ניקח את ההסתברות המותנת של בהנתן לפי נוסחת בייס:

מכיוון שמתקיים נקבל:

*ולכן חוק בייס אומר גם:*

נשנה בחוק בייס את: ,

– השארה/היפוטזה – מה שאני מאמין בו.

– , עדויות – מה שקורה בפועל.

למשל: אם אני מאמין שאצליח בתואר (היפוטזה), אך אם נכשלתי במבחנים זה יכול לגרום לי לשנות את ההיפוטזה שלי.

* *הערה: .*

*חוק בייס הופך את ההסתברות למשהו סובייקטיבי ופחות אובייקטיבי, כלומר ההסתברות הופכת להיות ממה הסיכוי שזה יקרה לכמה אני מאמין שהדבר הזה יקרה והאם יקרה מאורע שישנה את האמונה שלי.*

***תרגיל:***

*יש מחלה באוכליסיה, על כל אזרחים יש 10 חולים, הבדיקה למחלה זו מדייקת בכ-,*

*נגדיר: – האם אני חולה, – האם הבדיקה חיובית.*

הסיכוי שהדיקה חיובית כאשר אני לא יודע אם אני חולה או בריא:

הסיכוי שהבדיקה חיובית כאשר אני חולה ,

הסיכוי שאני חולה ,

לפפי נוסחת בייס:

* מונחים חשובים בהסתברות:
  + אפריורי – מידע שאני יודע לפני הנסיון.
  + פוסט-אפריורי – מידע שאני יודע אחרי הניסיון.

**שיעור שלישי –** *המשך מבוא לסטטיסטיקה והסתברות:*

*כאשר אומר באופ פרופרוציונאלי – אם אגף ימין גדל אזי גם אגף שמאל גדל ולהפך בהתאם.*

* *הערה: , כיוון שהם משלימים אחד את השני.*

**חוק ההסתברות השלמה:**

**משתנה אקראי:**

לוקח תוצאה ומצמיד לה מספר אקראי .

**לדוגמה:**

עבור הטלת מטבע: – תוצאה של אינה קבועה.

כלומר משתנה רגיל מקבל ערך יחיד, משתנה אקראי יכול לקבל מספר משתנים.

1. בדיד: מקבל מספר ערכים מוגבל של מספרים שלמים .
2. רציף: מספר בין 0 ל-1 כלומר אינסוף ערכים.

כאשר יש מרחב רציף אז הסיכוי לקבל ערך מסויים הוא 0 כי יש איסוף אפשרויות.

(פונקצית צפיפות הסתברות) – מתאר את ההסתברות לקבל ערך מסויים.

האינטגל של ,

**לדוגמה:**

כאשר מטילים מטבע 3 פעמים, מה הסיכויים שלא ייצא פלי?

מה הסיכוי שייצא פעם אחת?

בציור:

האפשרויות:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

– יצא פלי .

1. .
2. .

**תוחלת של משתנה אקראי - (הערך שמציפ םלקבל אחרי זמן).**

הסתברות של מה שייצא ב- הכי הרבה פעמים אחרי מספר פעמים (סימון: ).

**לדוגמה:**

ההסתברות שייצא סכום 7 בהטלת 2 קוביות:

**דוגמה:**

משחקים משחק, מוציאים קךף באקראי מהחפיסה, אם הקלף הוא צורה (מלך, מלכה, נסיך), אתה מקבל 5 ש"ח, אם הקלף הוא מספר אתה משלם 2 ש"ח.

האם שווה לשחק או שלא?

התוחלת של - , מכיוון שהתוצאה שלילית לא שווה לשחק.

**התפלגות אחידה בדידה:**

*התפלגות בה לכל האיברים בקבוצה סופית (בדידה) יש הסתברות שווה, כלומר לכל אחד מ- האיברים שיכולים להתקבל יש הסתברות שווה שיתקבלו שהיא .*

***התפלגות רציפה:***

*על טווח איסופי, רציף של ערכים כמו גובה, משקל, אורך חיי אדם וכו'.*

**שיעור רביעי –** *המשך מבוא לסטטיסטיקה והסתברות:*

**סטיית תקן:**

*מדד סטטיסטי לתיאור הפיזור של הנתונים המספריים סביב הממוצע שלהם, התלוי במרחק שלהם מן הממוצע, סטייית התקן נמדדת באותן היחידות כמו הנתונים עצמם, והוא שווה לשורש הריבועי של השונות ולכן היא חיובית, ושווה לאפס רק כאשר כל הנתונים שווים זה לזה, סימון (סיגמה קטנה).*

*יש להבחין בין סטיית התקן המחושבת לכל הנתונים לבין סטיית התקן המדגמית המחושבת על המדגם (תת-קבוצה), ומשמשת רק למדידה של סטיית התקן של התת-קבוצה.*

* *– הנתונים,*
* *– הממוצע.*

*לדוגמה:*

*נחשב את סטיית התקן בין המשכורות :*

**שונות:**

*תוחלת ריבועית – הסטיות מהתוחלת – נותן אינדוקציה על הפיזור והסיכון של פונקציית ההסתברות , באופן אינטואטיבי, השונות הוא הגודל החיובי התלוי במרחק (הריבועי), הממוצע של כל ערך ממוצע של כל הערכים, ערך שונות גבוה מעיד על פיזום רחב של משתנים, ערך נמוך מעיד על פיזור צר, שונות שווה לאפס אם כל הערכים שווים וזהים ומתרכזים בנקודה אחת.*

*סטיית התקן המדגמית:*

**התפלגות נורמלית:**

*בהתפלגות נורמלית ניתן להמיר את ציוני התקן לאחוזונים, אם ידועים לנו הממוצע וסטיית התקן – נוסחת הקו הידועה ולכן אפשר לחשב בדיוק את השטח תחת העקומה עד לציון מסויים באמצעות חישוב האינטגרל, שטח זה הוא בעצם האחוזון הציון.*

*בהתפלגות נורמלית הממוצע של משתנים בלתי תלויים בעלי אותה התפלגות, מתכנס בהתפלגות להתפלגות המואמליץ, לכן משתמשים בה המון כאשר לוקחים ממוצע של משתנים רבים כגון הממוצע של אנשים באוכלוסיה, ממדים פרמטרים שונים ועוד.*

*כאשר = התוחלת, = סטיית התקן.*

**התפלגות ברנולי:**

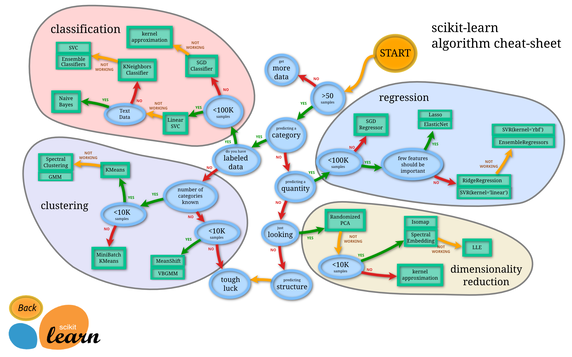
*מתארת התפלגות בדידה של משתנה אקראי המקל ערך 0 או 1, כלומר זו התאמה למערכות בהן יש שני מצבים – הצלחה או כישלון. במקרה זה מקובל לסמן הצלחה באות וכישלון בתור ההסתברות המשלימה , תוחלת של משתחנה אקראי המתפלג ברנולי הוא , שונות של משתנה אקראי הוא .*

**שיעור חמישי –** *מבוא ללמידת מכונה:*

**(ספרייה מדעית בפייתון):**

*ספריית נותנת אלגוריתמים ל-.*

* *– אנחנו צריכים מספיק דוגמאות כדי להתבסס עליהן.*
* *– אפשרויות ההחלטה של המודל (למשל: כלב או חתול).*
* *– מסווג אולי זה כלב או חתול, מסווג את הנתונים למחלקות שונות.*
* *– האם אנו יודעים לזהות ולמיין את סוגי הקלט השונים שלנו (דומגאות).*
* *– כשנדרש מספר מסויים למשל 3.5.*
* *– כאשר למודל אין מספר מסויים ידוע של ולכן אני חוזה אותו וזה בעיית – תוצאה שהיא מספרית.*



* *– לוקחת את המימד של הבעיה ומנסה להוריד אותה לבעיה קטנה יותר בעל מספר יותר מצומצם של מימדים.*

***הגדרות פורמליות:***

1. *– בהינתן מידע שהוא נתון לסיווג אנחנו רוצים לסווג אותו, לאחת מכמה קטגוריות, ידועות מראש – היוכלת לחזות מ-, מאמן את המכונה להבדיל בין המידע.*
2. *– כאשר רוצים לסווג מחלקות לפי נתונים מספריים, למשל בה ליידי ביטוי מחירים (נדל"ן, בורסה וכו').*
3. *– לנסות לעשות סוג של כאשר אין כלומר סידור לפי ההגיון.*
4. *– כאשר ישנה כמות הפרמטרים עצומה שנותנת לי תוצאה גבוהה מאוד המערכת בודקת האם ניתן לצמצם בפרמטים כדי להגיע בכל זאת לתוצאות טובות ולא לפגוע בכשירות המכונה.*

***– דוגמאות מסווגות מראש שנותנים למכונה:***

*לדוגמה – מביאים למכונה 150 תמונות של 3 סוגים שונים של פרחי ארוסי, אחרי שהמכונה תבדוק את כל הנתונים שקיבלנ מהתמונות הוא תדע את ההבדל בין כל הפרחים ותחשב את הממוצעים וכל מיני מידע אשר יותר מאוחר יעזור לה להחליט כאשר נביא לה פרח חדש מאחד הסוגים איזה סוג הפרח לפי הנתונים שאספה.*

***:***

*– החלק איתו מאמנים את המודל.*

*- החלק איתו בודקים את המודל.*

*– לראות האם יש משהו חשוד לא מתאים לשום דוגמה שהתבלה למודל.*

*- מודל שלא מתעדכן לאחר שסיים לבנות את המודל ומסתמך על כך שאם בעבר המודל עבר על כל הנתונים אז בדיעבד הנתונים בהווה לא השתנו.*

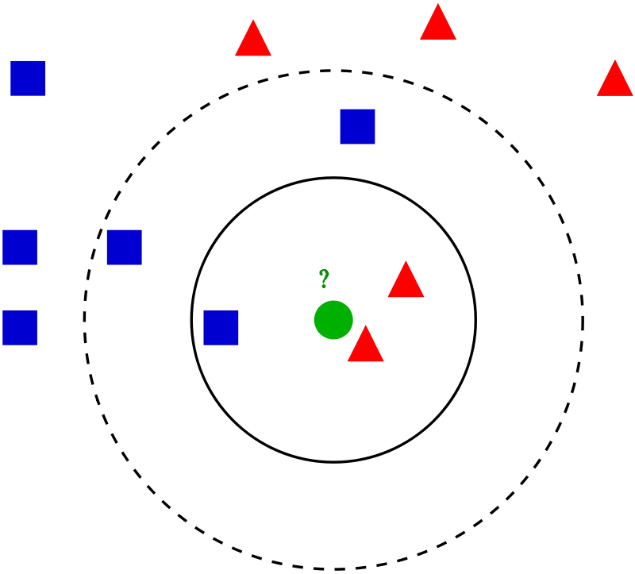
*– המודל יוצא מנקודת הנחה שמקרים ישנים נכונים אבל מעדכן את עצמו בהתאם למה שנכנס לו ולמה שמתקנים אותו.*

**אלגוריתם :**

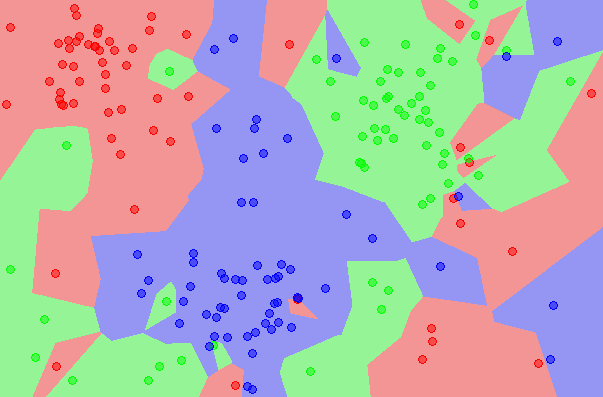
*נצא מנקודת הנחה שדברים דומים קרובים בנהים.*

*ממירים את הפיצ'רים – הפרמטרים לנקודות במרחב על מערכת צירים.*

*האלגוריתם בודק את הנקודות שקרובות לנקודה חדשה, מחשב לפי מרחקים גאומטרים, ולמה שהוא הכי קרוב הוא דומה לו.*

**

*מרחק מנהטן מחלק את כל השטחים לתחומים של סוגים שונים :*

**

*המשך – האלגוריתם מבוסס על זכרון קיים ונתונים שידועים מראש ללא פרמטרים שמתקבלים למודל.*

*פירוש: .*

**שיעור שישי –** *המשך מבוא ללמידת מכונה:*

**– למידת מודל:**

*בנית מודל מסויים המקושר לכל הנתונים שיש לי, גם אם יש סטיות בנתונים המודל יבנה לפי רוב הנתונים, ברגע שיש לנו מודל יש לנו םונקציה המקשרת בין רוב הנתונים וכבר לא צריך דוגמאות כי יש לנו את הפונקציה.*

*היתרונות:*

* *לא צריך לשמור את הדוגמאות – חוסך זיכרון.*
* *לא לקוח הרבה זמן ריצה.*

*חסרון:*

* *המודל לפעמים טועה אך ברוב המקרים צודק.*

**Memory / instance-based learning למידה עצלנית / למידה מונחת דוגמאות:**

*למשל זו למידה עצלנית, בגלל שעובדת רק לפי דוגמאות ולא לפי אג'נדה מסויימת, למדיה עצלנית מקבלת רק דומגאות מהזכרון וכל פעם שהיא עוברת על הדוגמא החדשה היא שומרת בזיכרון את הדוגמא ותוצואותיה ומשפרת את המודל*

*היתרון:*

* *דיוק גבוה.*

*חסרונות:*

* *זמן ריצה ארוך.*
* *דורש הרבה זיכרון.*

*אלגוריתם ב-:*

*פרמטרים:*

* *:*
  + *– לוקח את הרוב מתוך הדוגמאות.*
  + *– לוקח את המרחק מהדוגמה החדשה ונותן משקל למרחק.*

*ישנם שני פונקציות חשובות:*

* *– משתמש ב- כדי ללמוד ממנו וב- כדי להתאמן עליו.*
* *– לבדוק דוגמה חדשה (בדיקת ).*

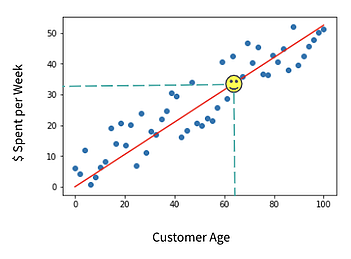
**תרגול שביעי :**

* *מידע נומריאלי (מספרי) :*
  + *מידע דיסקרטי (בדיד - מספרים שלמים) –*
  + *מידע רציף (אינסופי – מספרים ממשיים) -*
* *מידע קטגוריאלי :*
  + *כל המידע שיש במידע קטגוריאלי הוא ערך לא בר השוואה וממוין רק לפי קטגוריות, גם אם יש בו ערך שהוא מספרי, לא ניתן להשוותו עם עם ערכים אחרי*
* *מידע אורדינאלי ערבוב בין מידע נומריאלי למידע קטגוריאלי:*
  + *אפשר להשוות בין קטגוריות (למשל הסרט הנ"ל הוא טוב מהקודמו).*

**שיעור שביעי -** *המשך למידת מודל:*

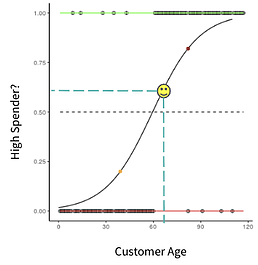
***:***

*בהנתן מידע, נמיר אותו לנקודות במערכת קרטזית ו- (בדומה לאלגוריתם ), נרצה ליצור פונקציה לינארית (או כמה פונקציות) אשר יוכלו לדמות מודל אשר יקבל מידע מסויים ועל פי הפונקציות שנבנו (על הדוגמאות) המודל יוכל לסווג מידע.*

**

***:***

כאשר יש לנו בעיות ריגרסיה ניתן לבנות את המודל על פונקציות שהם פולינומים ולא לינאריות



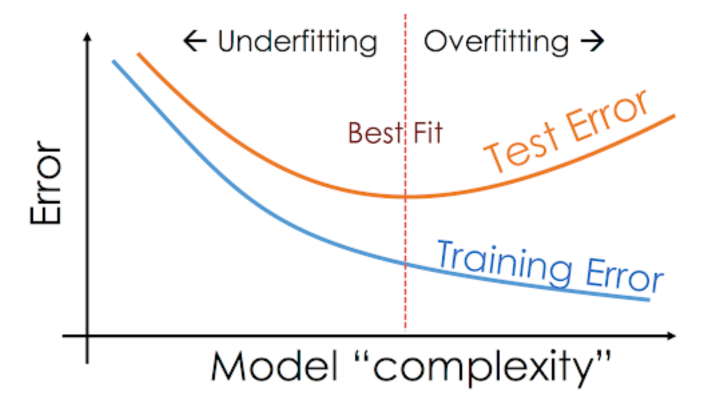
* **הערה:** מה ההבדל בין למידת מכונה ולמידה עמוקה ?:

בלמידת מכונה המדען נתונים (המומחה) הוא זה שבונה את המודל, מחליט איך הוא יעבוד ומגדיר את חוקיו בזמן שבלמידה עמוקה המכונה עצמה היא זאת שבונה את המודל.

**שיעור שמיני -** *המשך למידת מודל:*

***:***

*כאשר נבנה מודל יש לנו ישנם מספר בעיות אפשריות, או שהמודל פשוט מידי ואז המודל לא מספיק טוב בשביל הנתונים (), וזה נקרא , או שהמודל מורכב יותר מידי כך שמצד אחד יש פחות בעיות בלמידה על המודל כי הוא יותר מדוייק אך לאחר מכן בבדיקה על המודל יש מצבים שבהם המודל ישגה כיוון שהמודל מסובך מידי לנתונים החדשים, זה נקרא , ולכן נרצה כאשר אנחנו בונים מודל להגיע לנקודה האמצעית בין ה לבין ה, נקודה זאת נקראת :*

**

**שיעור תשיעי -** *המשך למידת מודל:*

:

: משיכת הנתונים לכיוון מסויים

: פיזור נתונים