Introduction to Data Science - Final Project

Roi Yozevitch & Idan Kaminetsky

הקדמה:

בעקבות ניסיון העבר ומגבלות הקורונה, הציון המסכם בקורס יתבסס ברובו על עבודת הגמר הנוכחית. מדובר בעבודה משמעותית הדורשת לא מעט מכם הסטודנטים.

את העבודה יש להגיש בתור מחברות Jupyter (סיומת ipynb) וקבצי PDF באתר GitHub ולצרף תיעוד משמעותי. נקודות יינתנו (או יוחסרו) על תיעוד מעולה (גרוע).

חשוב מאד

- נא לקרוא את **פרקים 1-4** בספר הקורס Hands-On Machine Learning with Scikit-Learn
- נא לעבור על הקורס של git באתר .Udacity . **זו מטלה חובה.** חלק מהציון הסופי תלוי במספרי ה-commits אשר תעשו.
- רשות - עברו על ***כל הקורס*** "מבוא למדעי הנתונים" באתר קמפוס (ולא רק על הפרקים העוסקים ברכישת נתונים).

הנחיות כלליות

1. לעבודה שני חלקים מרכזיים. החלק הראשון כולל שאלות תיאורטיות ושאלות תכנות פשוטות. במידה ועשיתם את שיעורי הבית, לא אמורה להיות לכם בעיה עם שאלות אלו. ההגשה של שאלות אלו היא **פרטנית** דרך GitHub. ז"א שכל סטודנט צריך להגיש בעצמו את הפתרונות של השאלות.
2. החלק השני (המרכזי) של העבודה מורכב מניתוח נתונים של data אותו אתם אמורים להשיג בעצמכם. את החלק הזה מגישים בזוגות. שימו לב לקובץ הרישום.

"בכל מערכת יחסים יש את הצד שאוהב והצד שמרשה שיאהבו אותו" [טולסטוי]

3. חשוב להסביר ששני סטודנטים יכולים לקבל ציון שונה על עבודה שהם הגישו ביחד. הסיבה היא שהעבודה כוללת מבחן סופי בע"פ (דרך zoom) אשר בו שני הסטודנטים יצטרכו להגן על העבודה.

ניסיון העבר מלמד שכאשר סטודנט אחד עושה את כל העבודה, ניתן לגלות את זה בצורה יחסית פשוטה בבחינה בעל-פה.

רכישת ה-data:

הדרך הטובה ביותר להשיג את ה-data היא על ידי תהליך של web scrapping. הקורס "מבוא למדעי הנתונים" באתר קמפוס מסביר במשך שני פרקים איך לעשות זאת. בנוסף, יש מדריכי למידה מעולים ברשת. בכל מקרה, מי שלא מצליח, יכול להוריד קובץ נתונים מאחד האתרים המרכזיים ברשת (לא מאתר Kaggle!) ולעבוד עליו. אתר מצוין בעברית הינו מאגר הנתונים הממשלתי ויש לו מקבילות גם באנגלית. ה-data יכול לעשות גם בשאלת סיווג (קלסיפיקציה) וגם בשאלת חיזוי.

*שימו לב, מי שיוריד נתונים **שלא דרך** web scrapping, הציון **המקסימלי** שהוא יוכל לקבל על העבודה הינו 90.*

חלק א- שאלות תיאורטיות ותכנות פשוט

הסתברות, חוק בייס:

1.

א. בערך 1/125 מהלידות זה תאומים לא זהים ו-1/300 מהלידות זה תאומים זהים. לאלביס היה אח תאום שמת בלידה. מה ההסתברות שאלביס היה תאום זהה? (ניתן להניח שההסתברות להולדת בן ובת שווה ל-1/2).

ב. יש שתי קערות של עוגיות. בקערה 1 יש 10 עוגיות שקדים ו-30 עוגיות שוקולד. בקערה 2 יש 20 עוגיות שקדים ו-20 עוגיות שוקולד. אריק בחר קערה **באקראי** ובחר ממנה עוגיה **באקראי**. העוגיה שנבחרה היא שוקולד. מה ההסתברות שאריק בחר את קערה 1?

2. בשנת 1995 חברת M&M הוסיפה את הצבע כחול. לפני השנה הזו, התפלגות הצבעים בשקית M&M היתה נראית כך:

30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan

החל משנת 1995, ההתפלגות נראית כך:

24% Blue , 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown.

לחבר שלכם יש 2 שקיות M&M, אחת משנת 1994 ואחת משנת 1996 והוא לא מוכן לגלות לכם איזו שקית שייכת לאיזו שנה. אבל הוא נותן לכם סוכריה אחת מכל שקית. סוכריה אחת היא צהובה ואחת היא ירוקה. מה הסיכוי שהסוכריה הצהובה הגיעה מהשקית של 1994?

3. (20 נקודות) הלכת לדוקטור בעקבות ציפורן חודרנית. הדוקטור בחר בך **באקראי** לבצע בדיקת דם הבודקת שפעת חזירים. ידוע סטטיסטית ששפעת זו פוגעת ב-1 מתוך 10,000 אנשים באוכלוסייה. הבדיקה מדויקת ב-99 אחוז במובן שההסתברות ל false positive היא 1%. הווה אומר שהבדיקה סיווגה בטעות אדם בריא כאדם חולה היא 1 אחוז. ההסתברות ל- false negative היא 0 – אין סיכוי שהבדיקה תגיד על אדם החולה בשפעת חזירים שהוא בריא. בבדיקה יצאת חיובי (יש לך שפעת).

   א. מה ההסתברות שיש לך שפעת חזירים?

   ב. נניח שחזרת מתאילנד לאחרונה ואתה יודע ש-1 מתוך 200 אנשים שחזרו לאחרונה מתאילנד, חזרו עם שפעת חזירים. בהינתן אותה סיטואציה כמו בשאלה א, מה ההסתברות (המתוקנת) שיש לך שפעת חזירים?

4. בערך 1/300 מהלידות היא של תאומים זהים ו 1/125 מהלידות היא של תאומים לא זהים. לנסיך צ'ארלס היה אח תאום שמת בלידה. מה ההסתברות שהיה לו אחת תאום זהה? (תאומים זהים חייבים להיות בני אותו המין)

קריאה/צפייה מומלצת:

מבוא מקסים לחוק בייס מאת אליעזר יודובסקי (מחבר הספר הארי פוטר והרציונליות)

Random Variables:

1. Roi is playing a dice game with Yael.

Roi will roll 2 six-sided dice, and if the sum of the dice is divisible by 3, he will win 6$. If the sum is not divisible by 3, he will lose 3$.

What is Roi's expected value of playing this game?

2. Sharon has challenged Alex to a round of Marker Mixup. Marker Mixup is a game where there is a bag of 5 red markers numbered 1 through 5, and another bag with 5 green markers numbered 6 through 10.

Alex will grab 1 marker from each bag, and if the 2 markers add up to more than 12, he will win 5$, 5. If the sum is exactly 12, he will break even, and If the sum is less than 12, he will lose 6$.

What is Alex's expected value of playing Marker Mixup?

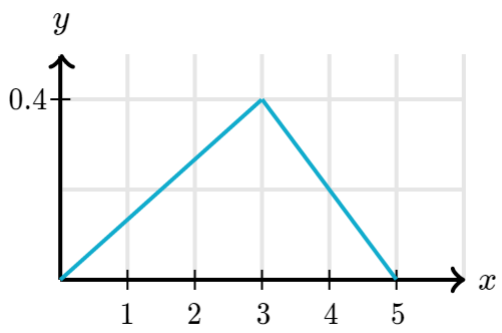
3. A division of a company has 200 employees, 40%, percent of which are male. Each month, the company randomly selects 8 of these employees to have lunch with the CEO.

What are the mean and standard deviation of the number of males selected each month?

4. Different dealers may sell the same car for different prices. The sale prices for a particular car are normally distributed with a mean and standard deviation of 26,000$ and 2,000$, respectively. Suppose we select one of these cars at random. Let X = the sale price (in thousands of dollars) for the selected car.

   Find P(26<X<30),

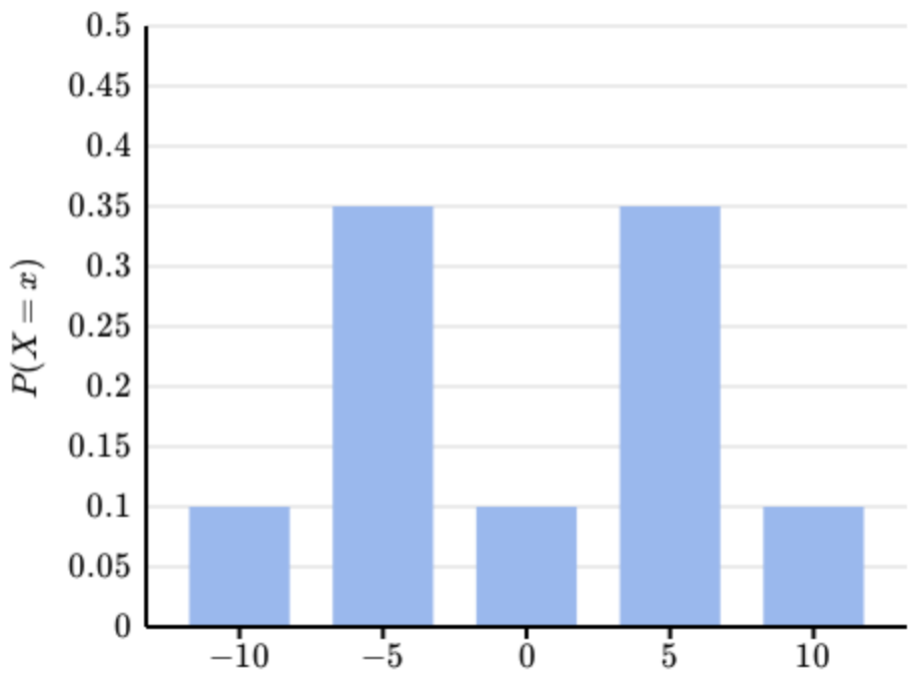
5. Given the following distribution, what is P(x>3)?



6. A company has 500 employees, and 60% of them have children. Suppose that we randomly select 4 of these employees.

   What is the probability that exactly 3 of the 4 employees selected have children?

7. Look at the next Graph. What is the expected value of X?



קריאה מומלצת:

מאמר קצר באתר Medium

מאמר נרחב יותר בנושא.

תרגילי תכנות ו-pandas

1. כתוב תוכנית בשפת פייתון אשר מקבלת מספר בבסיס 10 ומדפיסה את המספר בכל הבסיסים האחרים (בסיס 2, בסיס 8 ובסיס 16).
2. לבעייה זו מצב DATASET של סרטים.

```
import pandas as pd
cast = pd.read_csv('data/cast.csv')
cast.head()
```

2]:

	title	year	name	type	character	n
0	Suuri illusioni	1985	Homo $	actor	Guests	22.0
1	Gangsta Rap: The Glockumentary	2007	Too $hort	actor	Himself	NaN
2	Menace II Society	1993	Too $hort	actor	Lew-Loc	27.0
3	Porndogs: The Adventures of Sadie	2009	Too $hort	actor	Bosco	3.0
4	Stop Pepper Palmer	2014	Too $hort	actor	Himself	NaN

טענו את ה-DATASET למחברת וענו על השאלות הבאות:

1. How many movies have the title "Hamlet"?
2. List all of the "Treasure Island" movies from earliest to most recent.
3. How many roles were credited in the silent 1921 version of Hamlet?
4. Use groupby() to plot the number of "Hamlet" films made each decade
5. How many leading (n=1) roles were available to actors, and how many to actresses, in each year of the 1950s?
6. List the 10 actors/actresses that have the most leading roles (n=1) since the 1990's.
7. List, in order by year, each of the films in which Frank Oz has played more than 1 role

צפייה מומלצת

סרטון מקסים של טריקים ב-pandas

חלק ב – מדעי הנתונים

במטלה זוגית זו (כל זוג סטודנטים מגיש), יש צורך למצוא נתונים ברשת, לרכוש אותם (מומלץ דרך ספריית beautifulsoup) ואז לנסות לייצר מודל של ה-data, (איזו עמודה אנחנו רוצים לנחש). ניתן לקחת כל מידע טבלאי (לא תמונות, לא וידאו, לא אודיו) וניתן לבחור בעיית סיווג או חיזוי.

בנוסף, במידה ובחרתם בעיית חיזוי, חפשו ב-Kaggle בעיית סיווג, הורידו את ה-data ופתרו אותה.

במידה ובחרתם בעיית סיווג, חפשו ב-Kaggle בעיית חיזוי, הורידו את ה-data ופתרו אותה.

המטרה היא להכיר לכם את שני סוגי הבעיות והשיטות להתמודד איתן.

נקודות חשובות:

1. האם ה-data שלכם הינו מסוג סיווג או חיזוי?
2. מה פונקציית המטרה שלכם (איזו עמודה אתם מנסים לנחש)?
3. **איך אתם יכולים להבטיח שאין לכם data leakage?**
4. **מה אומר מודל 0? מה אומר מודל קצת יותר מתוחכם?**
5. **האם יש לכם רעיון אפריורי לקשר בין המשתנים? מה אמורה (לפי דעתכם) להיות פונקציית המטרה?**
6. איך ויזואליזציה יכולה לעזור לכם להבין את ה-data? במידה והחלטתם לצרף תרשים לעבודה שלכם, *מה אתם לומדים ממנו?* (חשוב מאד!)
7. לפי מה בחרתם את המאפיינים של המודל? האם הורדתם מאפיינים לא רלוונטים? לפי מה בחרתם?
8. מה המודל המתאים ביותר לבעיה שלכם? אבקש להשתמש גם במודלים שלמדנו (knn, linear regression) וגם במודלים אותם לא למדנו (לפחות מודל אחד). שימו לב שכאשר אתם עובדים עם מודל מסוים, בדקו את התיעוד הרלוונטי עליו בספריית sklearn. נסו להבין את הפרמטרים השונים.
9. מה מודל השגיאה שלכם? שימו לב שעבור בעיות מסוגים שונים, קיימים מודלי שגיאה שונים (למשל, דיוק בבעיות סיווג מול MSE בבעיות חיזוי). האם דיוק מספיק? במידה ולא, מה למדתם מעקומת AUC?
10. חשוב מאד לחלק את ה-data לסט אימון (train) וסט בחינה (test) ולבדוק על סט הבחינה רק כאשר סיימתם להתאים את הפרמטרים שלכם. עשו זאת גם אם אתם משתמשים ב--cross validation
11. האם יש לכם מאפיינם קטגוריאליים ב-data? איך אתם מטפלים בהם?

Additional Sources

כל חומר העזר (וסיכומי סטודנטים) יופיע בקישור הזה.