Introduction to Data Science - 2020 Semester A

Final Project

Roi Yozevitch

הקדמה:

בעקבות ניסיון העבר ומגבלות הקורונה, הציון המסכם בקורס יתבסס ברובו על עבודת הגמר הנוכחית. מדובר בעבודה משמעותית הדורשת השקעה רבה.

את העבודה יש להגיש בתור 3 מחברות Jupyter (סיומת ipynb) ולצרף תיעוד **משמעותי**. נקודות יינתנו (או יוחסרו) על תיעוד מעולה (גרוע). מחברת 1 תעסוק בשאלות הסתברות (10 תרגילים), ובכל תא תענו על שאלה אחת בדיוק בצורת הערה, מחברת 2 תעסוק בשאלות התכנות (7 תרגילים) גם כן לפי סדר השאלות כאשר בכל תא תענו על שאלה אחת בדיוק, מחברת 3 תעסוק בפרוייקט העיקרי שבחרתם אותו גם תצטרכו להציג לנו בפגישה בזום במהלך פגישה בת 10 דקות, פירוט יתר על משימה זו בהגדרת השאלה. אין להעזר\ להעתיק מסטודנטים אחרים, בדיקת ההעתקות תתבצע בהחלט.

יש להגיש את כל המחברות כולל התיעוד באנגלית

עליכם להגיש מסמך טקסט לתיבת ההגשה ששמו הוא מספר הת.ז שלכם למשל(12345678 .txt), ובפנים קישור לגיטהאב ולרפוסיטורי המתאים עם שלושת המחברות בגרסתו המעודכנת. יש לוודא שהקישור אכן מוביל אתכם לשם, טעויות מעין אלו יובילו להורדת נקודות משמעותית.

1. נא לקרוא את **פרקים 1-4** בספר הקורס Hands-On Machine Learning with Scikit-Learn
2. נא לעבור על הקורס של git באתר .udacity . **זו מטלה חובה.** בנוסף, יש לשנות את המחברות בגיטהב שלכם דרך commits שונים. יש להראות בבדיקת המטלה כרונולוגיקה של הפרוייקט ושיפור מתמיד.

מחברת 1 - הסתברות, חוק בייס (40 נקודות):

1.

א. בערך 1/125 מהלידות זה תאומים לא זהים ו-1/300 מהלידות זה תאומים זהים. לאליס היה אח תאום שמת בלידה. מה ההסתברות שאליס 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 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 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?

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?

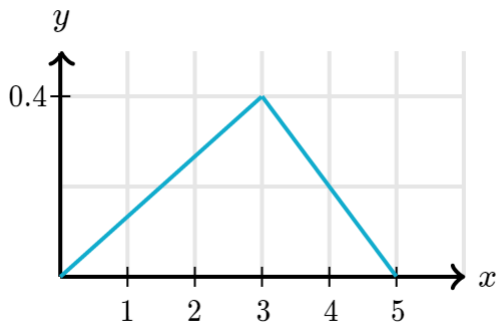
3. A division of a company has 200 employees, 40%, percent of which are male. Each month, the company randomly selects 8 of these employees to have lunch with the CEO.

   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),

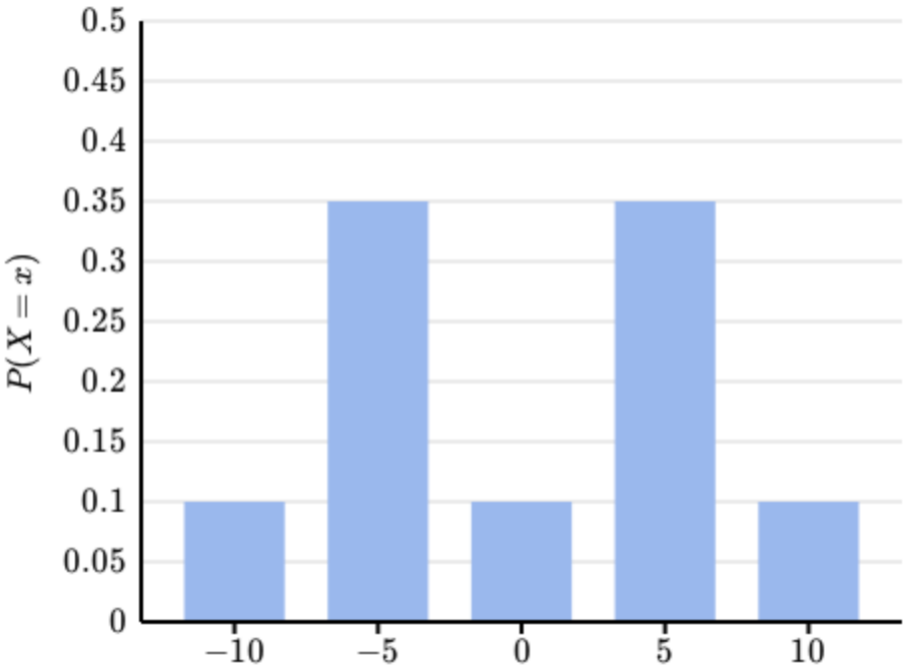
5. Given the following distribution, what is P(x>3)?



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?



מחברת 2-תרגילי תכנות (20 נקודות):

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

מחברת 3- Machine Learning(40 נקודות)

בשאלה זו המטרה היא לבחור 2 DATASET (רצוי מ-KAGGLE).

DATASET אחד אמור לעסוק בסיווג (CLASSIFICATION) והשני בחיזוי (REGRESSION).

מומלץ מאוד להתבונן במחברות אחרות שמפוזרות באינטרנט, אך בשום אופן לא להעתיק. תתבצע בדיקה מעמיקה כנגד העתקות אל מול מחברות רנדומליות שייבחרו.

המטרה היא לייצר מחברת עם הצגה של ה-DATA, בחירת המאפיינים הרלוונטים ובניית המודל.

עבור בעיית הקלסיפיקציה, בחרו גם את מודל KNN **וגם לפחות מודל אשר לא למדנו עליו תצטרכו ללמוד בעצמכם**. בדקו את הדיוק שלכם (Confusion Matrix ,ROC) והראו לכל אורך המחברת איך אתם משפרים את התוצאות. עליכם להציג ויזואליזציה **ברורה ומסבירה היטב** את הנתונים שאתם מבקשים להראות, עבור כל מחברת לפחות 3 גרפים מסבירים היטב. שימו לב, אין להשתמש בדאטה המוגדר "לימודי" כמו "אירוסים" "טיפים" "סרטים" וכיוצ"ב. תזכורת, תהליך למידת מכונה הוא כדלקמן: 1. התאמת הדאטה לעבודה. 2. ניקיון הדאטה. 3. חלוקת הרשומות לקבוצת אימון וקבוצת מבחן. 4. אימון המודל והצגת ביצועי המודל (הערכת מודל). יש להשתמש במשתני dummies אם צריך, יש להמיר משתני object לערכים נומריים.

עליכם להציג תובנות מעניינות מהמידע. עליכם לדעת להסביר היטב מדוע מחקתם פיצ'רים מסויימים, כאשר ההמלצה באופן כללי היא לא לעשות זאת, ובכל מקרה, בכל תהליך עיבוד דאטה משמעותי חייב להתלוות הסבר המניח את הדעת. יש להבין בצורה מושלמת האם המודל שלכם טוב או לא ולמה לא. יש להציג את מטריצת השגיאות בסוף עבור בעיית הקלסיפקציה, וכדאי להשתמש במפת חום עבור בעיית סיווג של שלושה או יותר משתנים דיכוטומיים.

בונוס: עבור המיקום שלכם יחסית לתחרות של KAGGLE.

בהצלחה לכולם

אוניברסיטת
אריאל
בשומרון

Additional Sources

(if you don't understand something or you wish to deepen your understanding)

Pandas and NumPy:

- https://www.udemy.com/course/the-pandas-bootcamp/
- https://www.youtube.com/watch?v=yzIMircGU5I&list=PL5-da3qGB5ICCsgW1MxlZ0Hq8LL5U3u9y&ab_channel=DataSchool
- https://bitbucket.org/hrojas/learn-pandas/src/master/

Probability and Random Variables:

- https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library
- https://www.youtube.com/watch?v=qBigTkBLU6g&list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9&ab_channel=StatQuestwithJoshStarmer
- https://arbital.com/p/bayes_rule/?l=1zq

Sci-Kit Learn:

- https://www.youtube.com/watch?v=elojMnjn4kk&list=PL5-da3qGB5ICeMbQuqbbCOQWcS6OYBr5A