למידה חישובית

תרגיל בית מספר 1

רגרסיה לינארית

1. (15 נקודות) בתרגיל זה נלמד להשתמש במודל של רגרסיה לינארית על-ידי שימוש ב- Scikit-Learn, ספריה של למידה חישובית. עבור כל אחד מהתרגילים בהמשך (2-6) יש להשוות את התוצאות המתקבלות על-ידי מימוש של הפונקציות שיש לכתוב באמצעות Python ו- numpy לתוצאות המודל של LinearRegression של scikit-learn (ראו בהמשך תרגיל זה).

ראשית נבצע את יבוא הספריות הדרושות לתרגיל:

```
import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns; sns.set()#  data visualization library

from sklearn.linear_model import LinearRegression
```

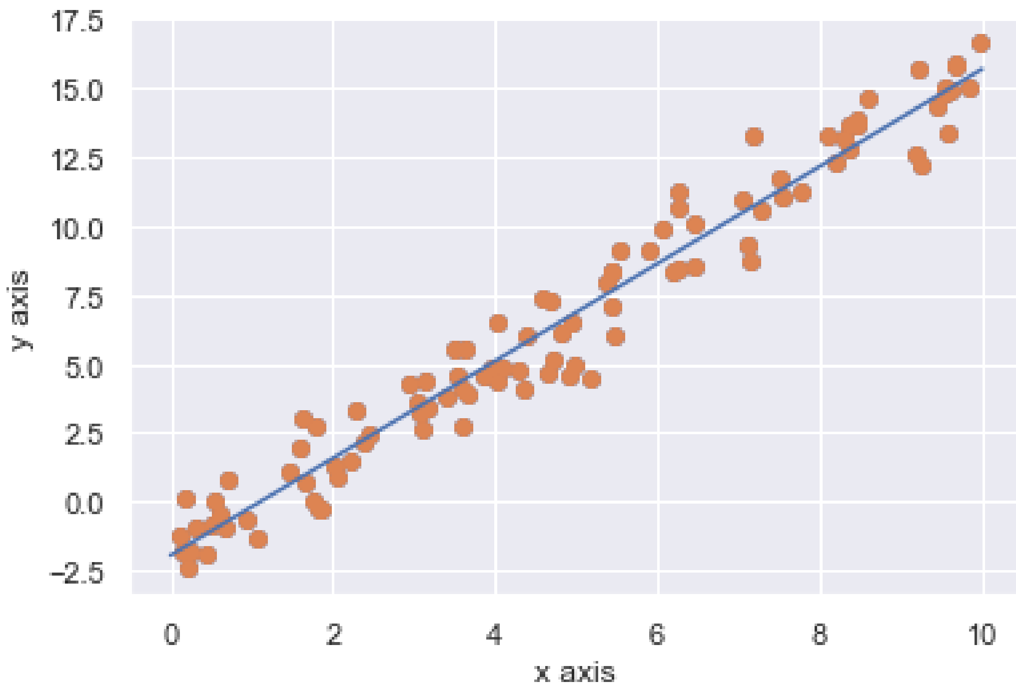
עתה נכין את הנתונים על-ידי דגימה אקראית של נקודות על הקטע [0,10].

```
# preparing the data

a1 = 1.8

a0 = -2

x = 10 * np.random.rand(100)

y = a0 + a1 * x  +  np.random.randn(100)

plt.scatter(x,y)
```

נשתמש במשערך רגרסיה לינארית של Scikit-Learn כדי להתאים מודל לנתונים (כלומר נשערך את a0 ו- a1.

```
model = LinearRegression(fit_intercept = True)

model.fit(x[:, np.newaxis], y)

xfit = np.linspace(0,10, 10000)

yfit = model.predict(xfit[:, np.newaxis])

plt.scatter(x, y)

plt.plot(xfit, yfit)
```

```
print("Model slope a1 = ", model.coef_[0])

print("Model intercept a0 = ", model.intercept_)
```



חזרו על תרגיל זה עבור דגימה אקראית של 500 נקודות על ציר ה- x בין 0 ל- 35, ויצרו נתוני y כך ש- y=ax+b עם b=5,a=2.7 והרעש האקראי הנוסף לדגימות y מתפלג גאוסית עם תוחלת 0 ושונות 25.

ציירו את הנתונים ואת הישר המתקבל באמצעות המודל כמו בדוגמה.

2. (10 נקודות) Piarce (1948) מדד את תדירות הצרצור של צרצרי קרקע (מספר תנודות כנפיים לשניה או פולסי קול לשניה). וכן את טמפ׳ הקרקע (ראו טבלה 1). מאחר וצרצרים הם בעלי חיים אקזותרמיים (בעלי דם קר) קיים בסיס להשערה כי הפעילות הפיזיולוגית שלהם תהיה תלוייה בטמפ׳ החיצונית, ולכן לכך קשר בין תדירות התנודות לבין הטמפ׳.

באופן כללי נמצא כי הצרצרים אינם משמיעים קול בטמפ׳ הנמוכה מ- 60 מעלות או גבוהה מ- 100 מעלות פרנהייט (15.5 ו-37 מעלות צלזיוס בהתאמה).

בתרגיל זה נניח כי קיים קשר לינארי בין התדירות לבין הטמפ׳.

א. ציירו את הנתונים באמצעות Python (ראו קובץ Cricket במחיצה Materials for ex. 1 - Linear Regression and Gradient Descent). 

צפו במודל בהקלטות הבאות הנמצאות ב- Linear regression :

1. Lecture 6 - Ex. 1.4 - computing the cost function, vectorization
2. Lecture 6 - vectorization of the cost function and of the gradient descent
3. Lecture 6 - python lab - plot_reg_line

ב. בסעיף זה נכתוב פונקציית Python לחישוב פונקציית המחיר עבור אלגוריתם ה- gradient descent. כתבו פונקציה בשם cost_computation שתחשב את המחיר לכל ערך q. הקלט לפונקציה וקטור הפרמטרים q, מטריצת הנתונים X וקטור המשתנה התלוי y, הפונקציה תחזיר את משתנה הפלט J.

ג. כתבו פונקציה למימוש אלגוריתם ה- Gradient Descent בשיטת batch.

ד. חשבו את הפרמטרים של הרגרסיה theta0, theta1 באמצעות הפונקציה שכתבתם, כאשר פונקציית ההיפותזה של תדירות הצרצור כתלות בטמפרטורה נתונה על-ידי $h_\theta(x) = \theta_0 + \theta_1 x_1$.

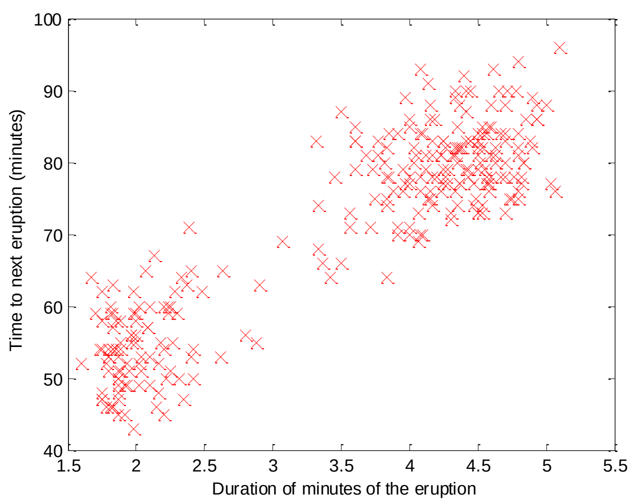
ה. ציירו את ישר הרגרסיה על דיאגרמת הפיזור של הנתונים.

ו. חשבו וציירו גרף של J כתלות במספר האיטרציה עבור ערכי α שונים, ובחרו ערך

ז. חשבו וציירו את הערך של פונקציית המחיר כתלות במספר האיטרציות (epochs). בדקו כמה epochs נדרשים להתכנסות עבור ערכי alpha שונים (בדקו לפחות 5 ערכי alpha שונים.

ח. מהי תדירות הצרצור הצפויה עבור טמפ׳ של 87 מעלות? ועבור 58 ו- 38 מעלות פרנהייט?

3. (10 נקודות) הגייזר הנאמן הוא גייזר הידרותרמי הנמצא בשמורת ילוסטון בוויומינג, ארה״ב, והוא אתר תיירות פופולרי. מקור השם הוא בסדירות המיוחסת להתפרצויות שלו. ידוע כי קיים קשר בין משך ההתפרצות הנוכחית לזמן עד ההתפרצות הבאה. עבור קבוצת נתונים המכילה 272 תצפיות, שכל אחת מייצגת התפרצות בודדת ומכילה שני משתנים המתאימים לזמן ההתפרצות בדקות, והזמן עד ההתפרצות הבאה בדקות. בתרגיל זה נבנה מודל של רגרסיה לינארית, באמצעותו נוכל לחזות את הזמן עד להתפרצות הבאה, אם ידוע משך ההתפרצות הנוכחית.



א. ציירו דיאגרמת פיזור של הזמן עד להתפרצות הבאה כפונקציה של משך ההתפרצות.

היתרון של שיטת ה- batch הוא בכך שבכל איטרציה נעשה שימוש בכל דוגמאות האימון, ואם צעד הלימוד הוא מספיק קטן קיימת התכנסות לנקודת מינימום. מצד שני אם קבוצת האימון מכילה דוגמאות רבות, כל איטרציה עשויה להימשך זמן רב לפני עדכון. השיטה השניה אותה למדנו, שיטת ה- on-line המכונה גם stochastic Gradient Descent פותרת בעיה זו על-ידי כך שנעשה עדכון לאחר כל דוגמת אימון. היתרון של שיטה זו הוא מהירות העדכון, והחיסרון הוא שכוון ההתקדמות עשוי להשתנות בכל דגימה, וכן קיימת הגעה

קרוב לנקודת המינימום. שיטת עדכון נוספת המקובלת מאוד באימון אלגוריתמי למידה היא שיטת ה- mini-batch Gradient Descent. בשיטה זו בכל איטרציה במקום לבחור את אחת הדוגמאות כמו ב- stochastic GD, או את כולן כמו ב- batch GD בוחרים באופן אקראי וללא החזרה קבוצות קטנות מתוך נתוני האימון. מחלקים את נתוני האימון באופן אקראי לתת-קבוצות שוות בגודלן כ בהתאם למספר הדוגמאות בקבוצת האימון. כל תת קבוצה נקראת mini batch, ומעדכנים את וקטור הפרמטרים theta בכל איטרציה לפי וקטורי התכונות ב- mini batch אחד. לאחר שעוברים על כל נתוני האימון (כל ה- mini batches), מעבר המכונה epoch אחד, חוזרים שוב על התהליך עבור כל נתוני האימון ב- epoch הבא, כאשר החלוקה לקבוצות mini-batch נעשית שוב באופן אקראי. ההיפרפרמטרים אותם צריך לקבוע כאן הם צעד הלימוד, גודל ה- mini batch ומספר ה- epochs.

ב. באמצעות אלגוריתם ה- gradient descent חשבו את מקדמי הרגרסיה הלינארית $q_0,\ q_1$ עבור המודל $h_q(x^{(i)}) = q_0 + q_1 x^{(i)}$.

הדרכה: כתבו script שייקרא main_faithful, וכן פונקציה שתיקרא gradient_descent בה תממשו את האלגוריתם (בשיטת mini-batch, עם $mb = 16$, כלומר 16 דוגמאות בכל mini-batch). משתני הקלט של הפונקציה הם X – מטריצת התכונות, בה כל שורה היא וקטור תכונות המייצג תצפית בודדת (עבור כל תצפית צריך להוסיף את $x_0^{(i)} = 1$), y – המשתנה התלוי, הזמן עד להתפרצות הבאה, q - וקטור הפרמטרים ההתחלתי, a - קצב הלימוד, ו- max_iter – מספר האיטרציות המקסימלי של אלגוריתם ה- gradient descent. משתני הפלט הם q - וקטור הפרמטרים, ו- J - המחיר לאחר ההתכנסות.

השתמשו בערך a של 0.01, והגבילו את מספר האיטרציות המקסימלי ל- 2000.

הנתונים נמצאים ב- faithful.txt,באתר הקורס במודל בתיקיה Materials for ex. 1 - Linear Regression and Gradient Descent).

ציירו את ישר הרגרסיה הלינארית על גרף דיאגרמת הפיזור של הנתונים.

ג. חשבו את הזמן הצפוי עד להתפרצות הבאה אם משך ההתפרצות הנוכחית הוא 2.1 דקות, 3.5 דקות ו- 5.2 דקות.

ד. כדי לבחון את אלגוריתם ה- gradient descent, כתבו פונקציה בשם cost_computation שתחשב את המחיר עבור כל ערך q. הפונקציה תקבל בכניסה את וקטור הפרמטרים q, את מטריצת הנתונים X ואת וקטור המשתנה התלוי y, ותייצר את משתנה הפלט J.

ה. בחנו את קצב הלמידה על-ידי שימוש בערכי a שונים ומצאו ערך המביא להתכנסות של פונקציית המחיר J.

4. (20 נקודות) בתרגיל זה נבצע חיזוי של מחירי בתים באמצעות רגרסיה מרובת משתנים. חברת הנדל"ן "בתים" אספה קבוצה של 100 נתוני בתים בשכונת נוף הים בעיר, ומציעה למוכרים או קונים חיזוי של מחיר הבית באמצעות מודל של רגרסיה לינארית הנבנה על בסיס נתוני קבוצת האימון. כדי להימנע מהכניסה לכל בית, הסוכן שביצע את המדידות מדד רק את אורך חזית הבית. הנתונים לתרגיל נמצאים בקבצי npy בתיקיית exercise 1 data במודל (TA_Xhouses.npy – אורך חזית הבית, TA_yprice.npy – מחיר הבית)

כדי לטעון את הנתונים ל- python יש להעתיק את הקובץ לתיקיית העבודה, ולאחר מכן להשתמש בפקודה הבאה:

```
np.load('TA_Xhouses.npy')
np.load('TA_yprice.npy')
```

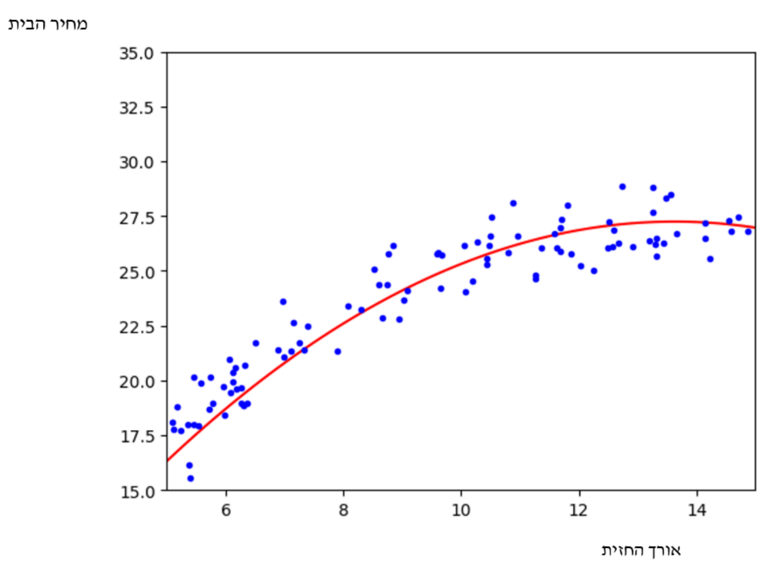
א. ציירו את דיאגרמת הפיזור של נתוני האימון (מחיר הבית במאות אלפי ₪ כתלות באורך חזית הבית במטרים).

ב. התאימו לקבוצת נתוני האימון מודל רגרסיה לינארית על-ידי 1. אלגוריתם ה- Gradient Descent, 2. חישוב באמצעות LinearRegression כפי שנעשה בשאלה 1. וציירו את ישר הרגרסיה על נתוני האימון.

ג. מה המחיר החזוי של בית עם אורך חזית של 15 מ׳ ושל 27 מ׳ לפי מודל הרגרסיה הלינארית?

ד. אנשי ה- ML של החברה מציעים להוסיף לנתוני האימון את ריבוע אורך החזית, וטוענים שמודל רגרסיה פולינומיאלי מסדר שני יכול להתאים טוב יותר לנתוני האימון ולבצע חיזוי מוצלח יותר עבור נתוני מבחן. כדי לבצע זאת הוסיפו תכונה נוספת – ריבוע התכונה של אורך החזית, העשוי לשקף את שטח הבית. כלומר עבור כל דוגמת בית, וקטור התכונות הוא מהצורה $X^{(i)} = [1 \quad X^{(i)} \quad (X^{(i)})^2]$ לדוגמא: [1 15 225] .
חזרו על סעיף ב׳ עבור מודל רגרסיה פולינומיאלי כנ״ל, והתאימו את העקום הריבועי לנתוני האימון (ראו ציור). כדי לבחון האם האלגוריתם ממומש וכן כדי למצוא ערך מתאים לצעד הלמידה α ציירו את J כתלות במספר האיטרציה עבור כל הרצה.
הדרכה: דוגמים את המשתנה הבלתי תלוי (אורך החזית) בין 5 ל- 15 ב- 500 נקודות, ומחשבים את y_predict עבור וקטור ה- theta הנלמד.

ה. מה יהיה המחיר החזוי עבור בית עם אורך חזית של 15 ו- 27 מ׳ באמצעות מודל הרגרסיה הפולינומיאלי? האם יש הבדל בין חיזוי זה לחיזוי באמצעות מודל רגרסיה לינארית?



כדאי לצפות ב- https://www.youtube.com/watch?v=sDv4f4s2SB8

5. (10 נקודות) עתה נשתמש ברגרסיה לינארית מרובה כדי לחשב את מחיר הבית כתלות בשטח של הבית ומספר חדרי השינה. הנתונים נמצאים בקובץ houses.txt כאשר העמודה הראשונה במטריצה data המתקבלת לאחר טעינת הקובץ על-ידי שימוש בפקודה :

```
data = load('houses.txt');
```

העמודה הראשונה מייצגת את שטח הבית, העמודה השניה את מספר חדרי השינה והעמודה השלישית את מחיר הבית באלפי דולארים.

א. קל לראות כי שטח הבית הוא בממוצע פי 1000 מהערך הממוצע של מספר החדרים. לפיכך קצב הלימוד עשוי להיות איטי. עלינו לבצע נירמול של המשתנים, כך שהערכים עבור כל תכונה יהיו עם ממוצע ושונות דומים. כדי לבצע זאת כתבו פונקציה data_normalization שתבצע נירמול של הנתונים. הפונקציה תקבל בכניסה את מטריצת הנתונים X, תחשב את הממוצע וסטיית התקן של כל עמודה, ותחזיר את הנתונים לאחר הפחתה של הממוצע וחלוקה בסטיית התקן. שמרו את נתוני הממוצעים וסטיות התקן.

כתבו סקריפט עבור תרגיל זה בדומה לתרגיל 3, התאימו מודל של רגרסיה לינארית עבור הנתונים של מחירי הבתים (באמצעות gradient descent) וחשבו את הפרמטרים של מודל הרגרסיה הלינארית (מרובת המשתנים). מהו המימד של וקטור הפרמטרים q ? הלמידה α ציירו את J כתלות במספר האיטרציה עבור כל הרצה.

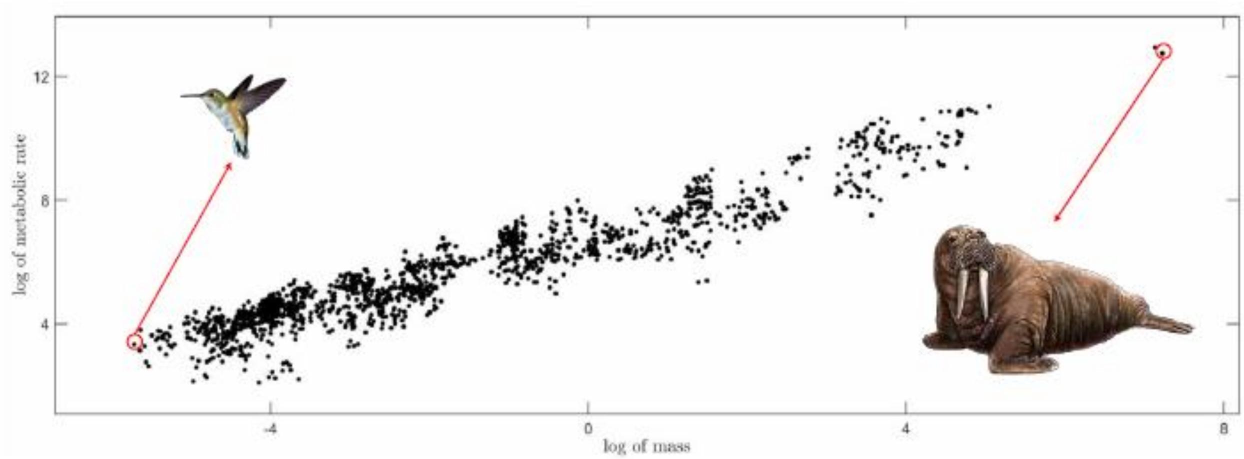
ב. מהו המחיר החזוי עבור בית ששטחו 1200 sf המכיל 5 חדרי שינה (לא לשכוח לנרמל את הנתונים עם ערכי הממוצעים וסטיות התקן לפי נתוני האימון)?

ג. חזרו על החישוב על-ידי שימוש במשוואות הנורמליות $q = (X^T X)^{-1} X^T y$

(אין צורך לנרמל את הנתונים)

6. (10 נקודות) הביולוג Max Kleiber אסף נתונים של מסת הגוף וכן של הקצב המטבולי של בעלי חיים רבים והבחין בקשר מעניין בין שני הערכים. לאחר סימון המשתנים כ- x_p ו- y_p עבור מסת הגוף בק״ג והקצב המטבולי ב- kJoul ליום בהתאמה, עבור כל בעל חיים, אם מפעילים לוגריתם טבעי על שני המשתנים מקבלים קשר לינארי ביניהם, כלומר :

$$\theta_0 + \log(x_p)\theta_1 \approx \log(y_p)$$



א. טענו את הנתונים (ראו במחיצה Materials for ex. 1 - Linear Regression and Gradient Descent). ציירו את דיאגרמת הפיזור של לוגריתם הקצב המטבולי כנגד הלוגריתם של מסת הגוף, כפי שמודגם באיור למעלה.

ב. התאימו מודל לינארי עבור הנתונים.

ג. השתמשו בפרמטרים האופטימליים המתקבלים מתוך הרגרסיה הלינארית ובתכונות הפונקציה הלוגריתמית כדי להביע את המשוואה הלא-לינארית הקושרת בין מסת הגוף x והקצב המטבולי y.

ד. השתמשו בישר הרגרסיה אותו התאמתם כדי לקבוע כמה קלוריות צורך יונק שמשקלו 250 ק״ג (כל קלוריה שוות ערך ל– 4.18 Joul).

ה. מה משקלו של יונק ימי הצורך 3.5 kJoul ליום?

7. (20 נקודות) תרגיל זה עוסק בשערוך סבירות מירבית (Maximum Likelihhod Estimation).

לפני הכנת התרגיל צפו בהקלטות הנמצאות ב- moodle ב - Linear Regression ובאופן מיוחד בשתי ההקלטות Maximum likelihood estimation – definition and examples, וב- Maximum likelihood estimation – linear regression.

נתונה פונקציית ההתפלגות הבאה:

$$p_X(x) = \theta \cdot e^{-\theta x}$$

כדי לשערך את הפרמטר θ מדדו m מדידות של המשתנה x: $x_1, x_2, \ \dots \ , x_m$.

מהו משערך הסבירות המירבית של θ?