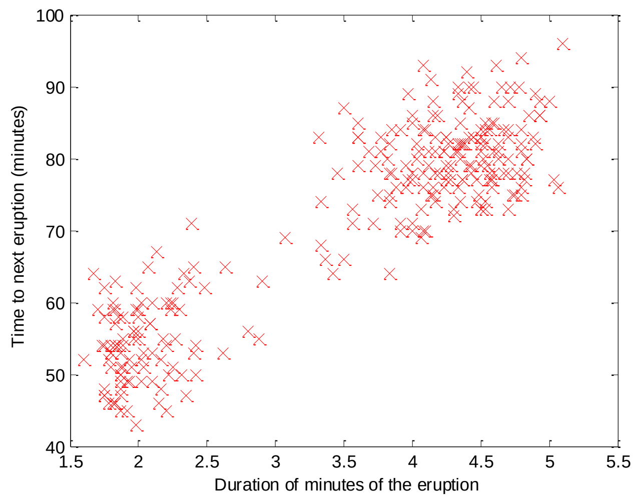
למידה חישובית וזיהוי תבניות

תרגיל בית מספר 1

1. (20 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 מדדנו את משך ההתפרצות הנוכחית.



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

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. 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.

כדי לבחון האם האלגוריתם ממומש באופן נכון ציירו את J כתלות במספר האיטרציה.

הנתונים נמצאים באתר הקורס (faithful.txt).

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

ג. חשבו את הזמן הצפוי עד להתפרצות הבאה אם משך ההתפרצות הנוכחית הוא 1.5 דקות, 3.0 דקות ו- 5 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. יצרו סריג של ערכי q באופן הבא:

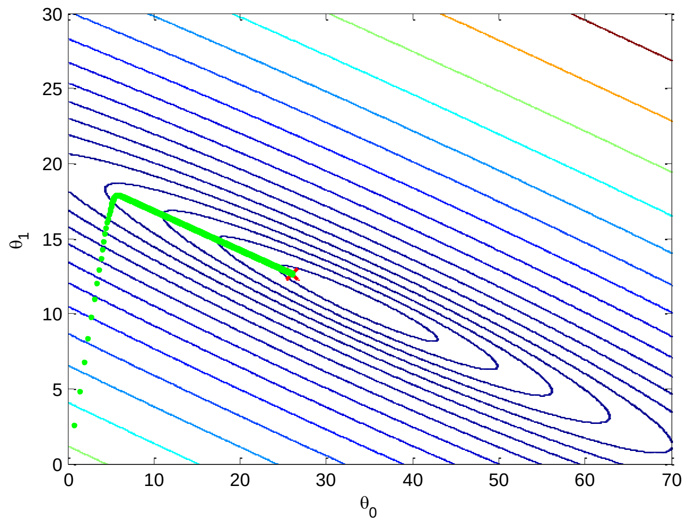
```
% Grid for contour plot of gradient descent
theta0=linspace(0, 30, 500);
theta1=linspace(0, 20, 500);

J =zeros(length(theta0),length(theta1));

% a matrix of J values for each theta
for i = 1:length(theta0)
    for j = 1:length(theta1)
      thetamatrix = [theta0(i); theta1(j)];
      J(i,j) = computeCost(X, y, thetamatrix);
    end
end

% contour plot using a logarithmic scale
contour(theta0, theta1, J, logspace(-2, 7, 35))
xlabel('\theta_0'); ylabel('\theta_1');
hold on;
plot(theta(1), theta(2), 'bx', 'MarkerSize', 5, 'LineWidth', 2);
% theta(1) and theta(2) are the values computed by gradient descent
```

ציירו על ציור ה- contour plot (על-ידי שימוש בפקודת hold on) את האבולוציה של ערכי q המחושבים על-ידי אלגוריתם ה- gradient descent. הציור שיתקבל צריך להיות דומה לציור הבא:



ו. בחנו את קצב הלמידה על-ידי שימוש בערכי a שונים.

2. (20 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');
```

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.

חזרו על התרגיל הקודם עבור הנתונים של מחירי הבתים (באמצעות gradient descent) וחשבו את הפרמטרים של מודל הרגרסיה הלינארית (מרובת המשתנים). מהו המימד של וקטור הפרמטרים q?

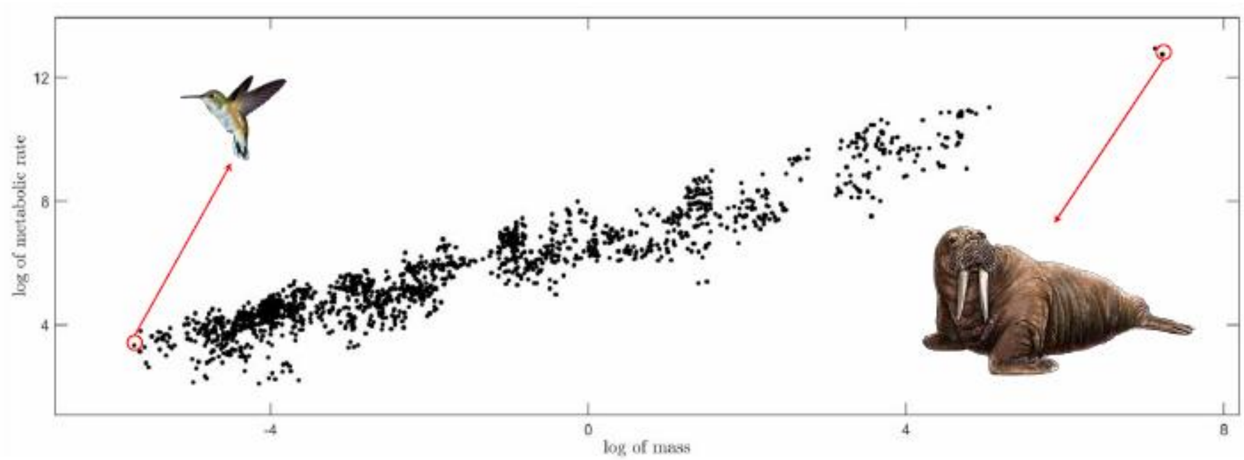
ב. מהו המחיר החזוי עבור בית ששטחו 1800 sf המכיל 5 חדרי שינה (לא לשכוח לנרמל את הנתונים עם ערכי הממוצעים וסטיות התקן לפי נתוני האימון)?

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

(אין צורך לנרמל את הנתונים)

3. (20 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:

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



א. טענו את הנתונים למשטח העבודה של ה- Matlab (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 10 ק״ג (כל קלוריה שווה ערך ל– 4.18 Joul).

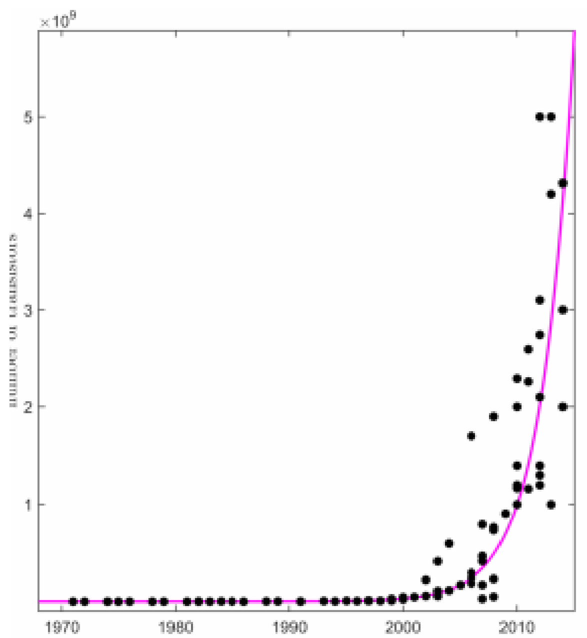
ה. מה משקלו של יונק ימי הצורך 1.63 kJoul ליום?

4. (20 נקודות) גורדון מור (Gordon Moore), אחד ממייסדי חברת אינטל, חזה במאמר משנת 1965 שמספר הטרנזיסטורים על מעגל משולב יוכפל בערך כל שנתיים. השערה זו, המכונה "חוק מור", הוכחה כמדוייקת במידה מספקת במשך כחמישה עשורים. מאחר וכוח המחשוב של מחשבים נמצא ביחס ישר למספר הטרנזיסטורים ב- CPU, חוק מור מספק מודל המאפשר לחזות את כוח המחשוב של מיקרופרוססורים עתידיים. באיור בהמשך אפשר לראות את כמות הטרנזיסטורים במספר מיקרופרוססורים החל מ- Intel 4004 משנת 1971 שהכיל 2300 טרנזיסטורים, ועד ל- Xeon E7 המכיל למעלה מ- 4.3 מיליארד טרנזיסטורים.

א. הציעו טרנספורמציה עבור קבוצת הנתונים של חוק מור הנמצאת בקובץ Moor_low.csv .
הדרכה: כדי ליצור קשר לינארי, יש צורך לבצע טרנספורמציה לנתוני הפלט (מספר הטרנזיסטורים) ולא לקלט.

ב. נסחו ומזערו פונקציית עלות עבור פרמטרים או משקלות מתאימים, והתאימו את המודל למרחב הנתונים לאחר הטרנספורמציה, ולמרחב הנתונים המקורי. צרפו גרפים מתאימים עבור כל אחד מהמקרים.

ג. מה מספר הטרנזיסטורים הצפוי לפי נתונים אלה עבור מעבדים שיוצרו ב- 2017?



5. (20 נקודות) עבור בעיית הרגרסיה הליניארית, נניח כי וקטור התכונות מכיל תכונה אחת בלבד (לדוגמא: שטח הבית בדוגמת מחירי הדירות, זמן ההתפרצות בדוגמת הגייזר הנאמן), כלומר ההיפותזה $h_\theta(x^{(i)})$ היא:

$$h_\theta(x^{(i)}) = \theta_0 + \theta_i x^{(i)} ,$$

וכן פונקציית המחיר :

$$J\left(\theta\right)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)})-y^{(i)}\right)^2=\frac{1}{2m}\sum_{i=1}^{m}\left(\theta_0+\theta_i x^{(i)}-y^{(i)}\right)^2$$

חשבו באופן אנליטי (על-ידי גזירת פונקציית המחיר) מהם θ_0 ו- θ_1 האופטימליים הממזערים את $J\left(\theta\right)$.