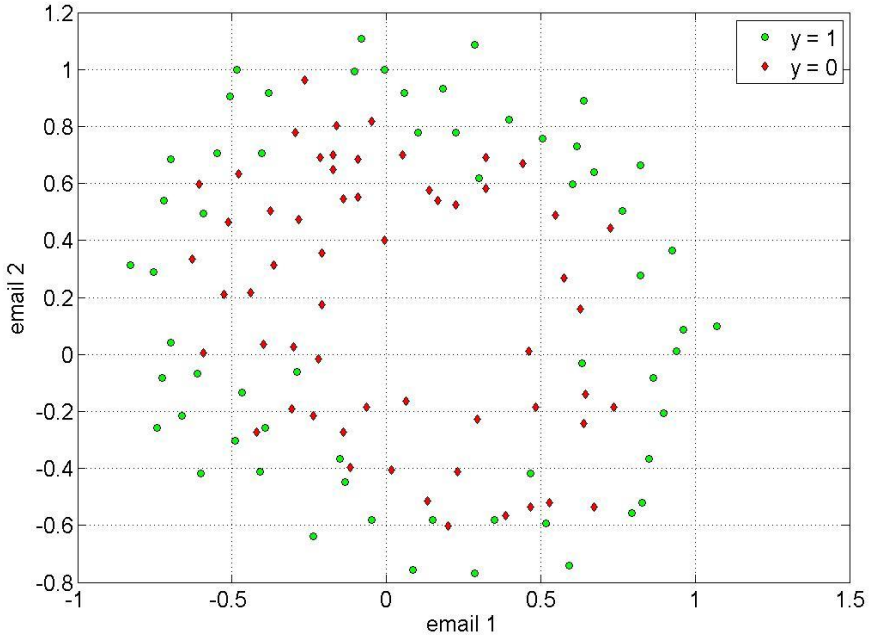
זיהוי תבניות ולמידה ממוחשבת

תרגיל כיתה מספר 5

1. כדי לבנות מערכת לומדת לזיהוי דואר spam נעשו שתי מדידות שונות, x1 ו – x2, באמצעותן רוצים לסווג את הדואר הנכנס לשרת e-mail למכתבים אמיתיים ול- spams.

לרשותכם קבוצת אימון של שתי דוגמאות, עבורן לכל אחת שתי תכונות, וידוע האם כל דוגמא היא מכתב או spam, כלומר הדוגמאות מתוייגות.

הנתונים מוצגים באמצעות דיאגרמת הפיזור הבאה:



א. הנתונים נמצאים במחיצת טענו את הנתונים של קבוצת האימון לתוך ה- Matlab על-ידי

```
data = load('email_data.txt');
 X = data(:, [1, 2]); y = data(:, 3);
```

יצרו דיאגרמת פיזור בדומה לציור למעלה, בו כל נקודה מיוצגת על-ידי שתי התכונות x1 ו- x2, ומסומנת על-ידי עיגול ירוק אם הדואר תקין y=1 ויהלום אדום עבור spam (y=0).

ב. השתמשו ברגרסיה לוגיסטית כדי להפריד בין שתי הקבוצות (דואר רגיל ודואר spam). מהי מסקנתכם?

ג. כדי להפריד בין דוגמאות האימון נשתמש ברגרסיה לוגיסטית עם פולינום מסדר גבוה יותר. ננסה לבחון מודל מסדר גבוה יותר כך שבמקום המודל הליניארי נתאים לנתונים מודל מסדר חמישי. נשתמש בפונקציה mapFeature כדי לעשות זאת:

```
X = mapFeature(X(:,1), X(:,2));
```

וקטור התכונות הוא עתה:

$$\text{mapFeature}(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1x_2 \\ x_2^2 \\ x_1^3 \\ \vdots \\ x_1x_2^5 \\ x_2^6 \end{bmatrix}$$

באמצעות פונקציית המחיר

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(x^{(i)})) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2.$$

וכן אלגוריתם ה- gd:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \qquad \text{for } j = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \left(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j \quad \text{for } j \geq 1$$

כתבו פונקציה שתבצע את פעולת המערכת הלומדת.