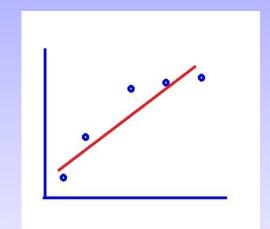
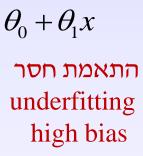
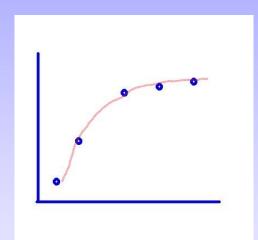
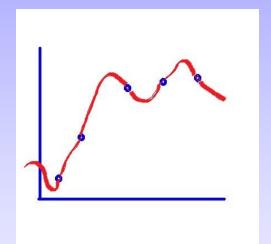
- overfitting בעיית התאמת היתר
 - :דוגמא רגרסיה לינארית







$$\theta_0 + \theta_1 x + \theta_2 x^2$$
 מודל מתאים right model

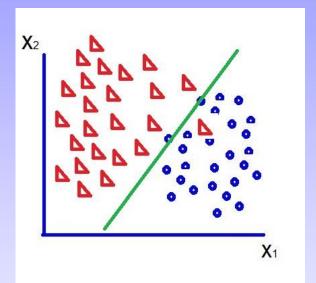


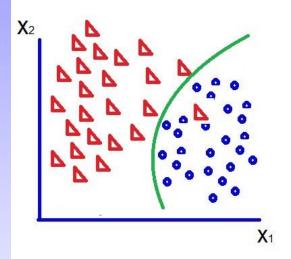
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \dots$$
התאמת יתר

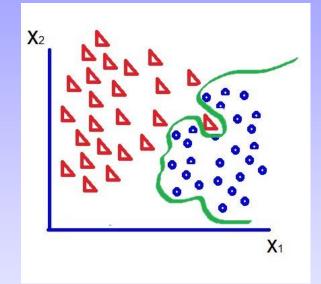
overfitting high variance

- אם יש יותר מדי תכונות היפותזת הלמידה או פונקציית החיזוי עשויה להתאים לקבוצת האימון בצורה מושלמת, כלומר D=(θ)=0, אבל יכולת ההכללה לדוגמאות חדשות היא נמוכה.
 - (generalization) הכללה •

דוגמא: רגרסיה לוגיסטית







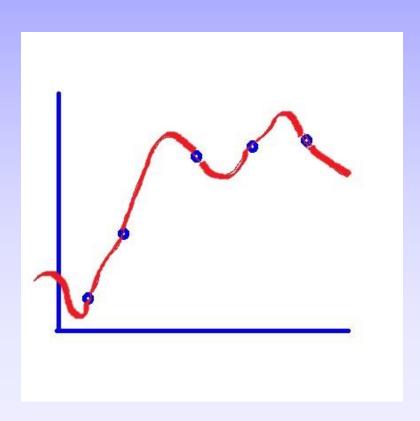
$$h_{\theta}(x) = g\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)$$

underfitting

$$h_{\theta}(x) = g \begin{pmatrix} \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \dots \\ \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 \end{pmatrix}$$

$$h_{\theta}(x) = g \begin{pmatrix} \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots \\ \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 \end{pmatrix}$$
 overfitting

- ? מה אפשר לעשות כדי להחליט מהי ההיפותזה הנכונה
- אם יש רק שתי תכונות, אפשר לצייר את עקום הרגרסיה ולהחליט.לדוגמא עבור הרגרסיה הליניארית:



אבל עבור מספר רב של תכונות (לדוגמא עבור בעיית הגייזר):

משך ההתפרצות – X_1

עצמת ההתפרצות – X_2

שטח הפיזור – X_3

 X_4 גובה ההתפרצות

 $- x_5$ מרחק ההתפרצות

הרבה יותר קשה לראות מהו העקום אותו צריך להעביר.

סverfitting תכונות רבות ומעט דוגמאות ⇒ייתכן שיתקבל

- מענה לבעיית התאמת היתר:
 - 1. הפחתת מספר התכונות
- א. החלטה באופן ידני איזה תכונות לשמור.
- ב. הפעלת אלגוריתם אוטומטי לבחירת התכונות

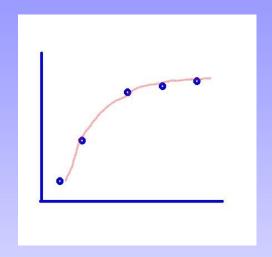
חיסרון – הפחתת מספר התכונות עשויה לפגוע בביצועי המסווג, התכונות עשויות להכיל מידע חשוב לחיזוי (לדוגמא משך ההתפרצות בדוגמת הגייזר הנאמן או שטח הבית בדוגמת הבתים).

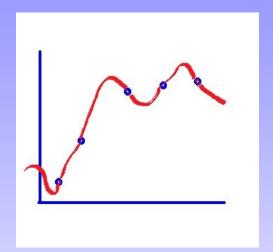
:(Regularization) רגולריזציה

שומרים על כל התכונות, אבל מורידים את ערכי הפרמטרים θ.

היתרון – שיטה זו עובדת היטב כאשר יש מספר רב של תכונות, וכל אחת מהן תורמת לחיזוי y.

• ראינו בדוגמאות הקודמות (רגרסיה לינארית):





$$\theta_0 + \theta_1 x + \theta_2 x^2$$

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \dots$$

- בחירה בהיפותזה השמאלית התאמה די טובה.
 - בחירה בהיפותזה הימנית התאמת יתר.
- נניח כי "מענישים" על-ידי הוספת גורם קנס לפונקציית המחיר כך $heta_3, heta_4$ יהיה מאוד קטן. $heta_3, heta_4$

• נתבונן בפונקציית המחיר:

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)})^{2}$$

: θ_3, θ_4 -נוסיף גורם עונש ל

$$\min_{\theta} \left(\frac{1}{2m} \sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)})^2 + 10^6 \theta_3^2 + 10^6 \theta_4^2 \right)$$

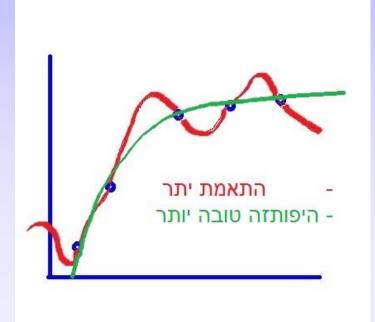
?מהו ערכם של $heta_{\!\scriptscriptstyle 3}, heta_{\!\scriptscriptstyle 4}$ לאחר מזעור פונקציית המחיר

$$\theta_3 \approx 0$$
, $\theta_4 \approx 0$

לאחר מזעור פונקציית המחיר: •

נקבל ביטוי קרוב לריבועי (קוודרטי)

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$



- ערכים קטנים עבור כל הפרמטרים מביאים להיפותזה פשוטה יותר ופונקציה חלקה יותר.
- overfitting ההיפותזה פחות נוטה ל

אפשר להראות שערכים קטנים עבור כל ערכי הפרמטרים מביאה $\theta_0, \theta_1, \dots, \theta_n$ להיפותזה פשוטה ופונקציה חלקה יותר

$$X_0, X_1, \ldots, X_{99}$$

$$\theta_0, \theta_1, \ldots, \theta_{99}$$

- י דוגמא: מחירי הבתים:
- איך נדע איזה פרמטרים צריך לבחור ואת איזה לכווץ?
 - הפתרון: כווץ כל הפרמטרים.

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

.(לא מקובל לכלול את $\,\,\,\,\,\,\,\,\,\,\,\,\,\,\,\,\,\,$. ההבדל מאוד קטן אם כוללים או לא) -

• הפתרון: כווץ כל הפרמטרים.

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

- ($heta_0$ את "לא מקובל לכלול את $heta_0$. (למרות שזו לא שגיאה "להעניש" גם את $heta_0$
 - ההבדל מאוד קטן אם כוללים או לא).

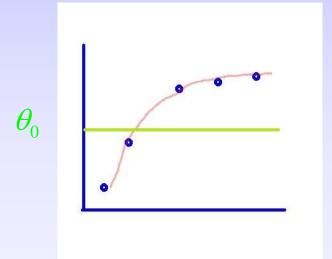
$$\min_{\theta} \left(\frac{1}{2m} \left(\sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right) \right)$$

- :פרמטר הרגולציה פשרה בין שתי מטרות שונות λ
 - 1) התאמה לנתוני האימון
 - .מניעת התאמת יתר על-ידי הקטנת הפרמטרים (2
 - .שולטת על הפשרה λ •
- הרגולריזציה מאפשרת להשתמש בכל התכונות מבלי לגרום להתאמת יתר.
 - אפשר לכלול פולינום מסדר גבוה.
- איך נראית הפונקציה הנוצרת? לא קוודרטית, אך עדיין חלקה ומתאימה לנתונים. •

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

 $\lambda = 10^{10}$ מה יקרה אם נבחר λ מאוד גבוה? מה יקרה אם נבחר לניח ערך זה עבור המודל:

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5$$

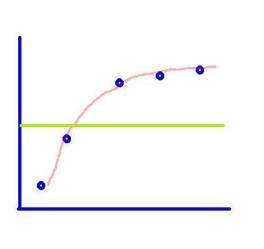


 $\lambda = 10^{10}$?מה יקרה אם נבחר λ מאוד גבוה

$$\theta_1 \approx 0$$
, $\theta_2 \approx 0$, $\theta_3 \approx 0$, $\theta_4 \approx 0$, $\theta_5 \approx 0$

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5$$

הישר בשיפוע 0 יוצר התאמת חסר, לפיכך דרושה בחירה טובה יותר של פרמטר הרגולציה λ



(Regularized Linear Regression)

נכליל את אלגוריתם ה- gd כך שיכלול גם ביטוי לרגולריזציה

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

Regularization term

רוצים למצוא פרמטרים למזעור פונקציית המחיר.

(Regularized Linear Regression)

$$\begin{split} \frac{\partial}{\partial \theta_{j}} J(\theta) &= \frac{\partial}{\partial \theta_{j}} \left(\frac{1}{2m} \left(\sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)})^{2} + \lambda \sum_{k=1}^{n} \theta_{k}^{2} \right) \right) = \\ &= \frac{1}{2m} \left(\sum_{i=1}^{m} 2(h_{\theta}(x^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_{j}} h_{\theta} \left(x^{(i)} \right) + 2\lambda \theta_{j} \right) = \\ &= \frac{1}{m} \left(\sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)}) x_{j}^{(i)} \right) + \frac{\lambda}{m} \theta_{j} = \\ &= -\frac{1}{m} \left(\sum_{i=1}^{m} (y^{(i)} - h_{\theta}(x^{(i)})) x_{j}^{(i)} \right) + \frac{\lambda}{m} \theta_{j} = \end{split}$$

(Unregularized Linear Regression)

Gradient Descent : ללא רגולריזציה

Repeat (until convergence)

$$\theta_{j} := \theta_{j} - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_{\theta}(x^{(i)}) - y^{(i)}) x_{j}^{(i)} \quad j = 0, 1, 2, ..., n$$
}

(unregularized Linear Regression)

עתה נתייחס ל- θ_0 לחוד (מאחר ובגורם הרגולריזציה הוא לא מופיע)

Repeat (until convergence)

$$\theta_0 := \theta_0 + \alpha \frac{1}{m} \left(\sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)})) x_0^{(i)} \right)$$

$$\boldsymbol{\theta_j} \coloneqq \boldsymbol{\theta_j} + \alpha \frac{1}{m} \left(\sum_{i=1}^m (y^{(i)} - h_{\theta}(x^{(i)})) x_j^{(i)} \right) \quad j = \emptyset, 1, 2, \dots, n$$

הסימן של α השתנה כי הפכנו את הסדר עבור ההפרש בין y(i) לפונקציית ההיפותזה

(Regularized Linear Regression)

נוסיף את גורם הרגולריזציה:

Repeat (until convergence)

$$\theta_{0} := \theta_{0} + \alpha \frac{1}{m} \left(\sum_{i=1}^{m} (y^{(i)} - h_{\theta}(x^{(i)})) x_{0}^{(i)} \right)$$

$$\theta_{j} := \theta_{j} + \alpha \left(\frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - h_{\theta}(x^{(i)})) x_{j}^{(i)} - \frac{\lambda}{m} \theta_{j} \right) \quad j = 1, 2, ..., n$$

$$rac{\partial}{\partial heta_i} J(heta)$$

(Regularized Linear Regression)

:או באופן אקויולנטי

{ Repeat (until convergence)

$$\theta_0 := \theta_0 + \alpha \frac{1}{m} \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)})) x_0^{(i)}$$

$$\theta_{j} := \theta_{j} \left(1 - \alpha \frac{\lambda}{m} \right) + \alpha \frac{1}{m} \left(\sum_{i=1}^{m} (y^{(i)} - h_{\theta}(x^{(i)})) x_{j}^{(i)} \right) \quad j = 1, 2, \dots, n$$

(Regularized Linear Regression)

$$\left(1-lpha\,rac{\lambda}{m}
ight) < 1$$
 : קל לראות

ולכן החלק השמאלי של הביטוי מתכווץ לאחר כל איטרציה. מכווצים את הפרמטרים $\,\theta_{\rm i}\,$ מעט לאחר כל איטרציה

Repeat (until convergence)

$$\theta_0 := \theta_0 + \alpha \frac{1}{m} \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)})) x_0^{(i)}$$

$$\theta_{j} := \theta_{j} \left(1 - \alpha \frac{\lambda}{m} \right) + \alpha \left(\sum_{i=1}^{m} (y^{(i)} - h_{\theta}(x^{(i)})) x_{j}^{(i)} \right) \quad j = 1, 2, ..., n$$

החלק השני – כמו קודם ללא הרגולריזציה.

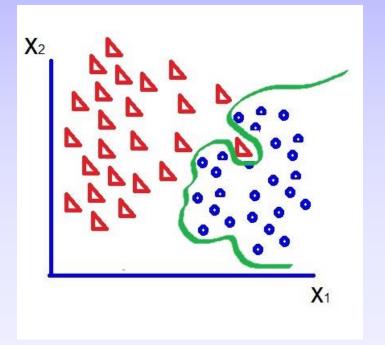
מניחים כי lpha קטן ו- m יחסית גדול alleq

(Regularized Logistic Regression)

ראינו כבר כי שימוש בפונקציית היפותזה מורכבת

: כמו

$$h_{\theta}(x) = g \begin{pmatrix} \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \dots \\ \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 \end{pmatrix}$$
: 122



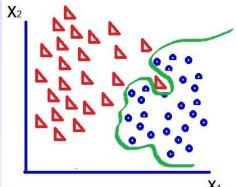
overfitting -עשוי לגרום ל

(Regularized Logistic Regression)

פונקציית המחיר:

$$J(\theta) = \left[\frac{1}{m} \left(\sum_{i=1}^{m} -y^{(i)} \log \left(h_{\theta}(x^{(i)}) \right) - \left(1 - y^{(i)} \right) \log \left(1 - h_{\theta}(x^{(i)}) \right) \right) \right]$$

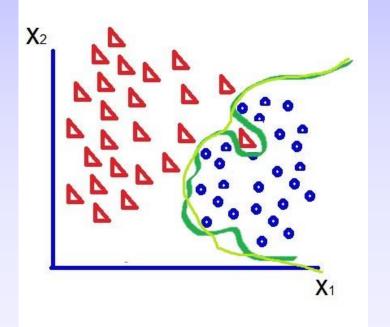
- שימוש בתכונות רבות (לאו דווקא פולינומיאליות) עשוי לגרום למצב
 של התאמת יתר.
 - כדי להימנע ממצב זה נוסיף איבר לפונקציית המחיר שיגרום לכך שהפרמטרים לא יהיו יותר מדי גדולים.



(Regularized Logistic Regression)

נוסיף איבר לפונקציית המחיר:

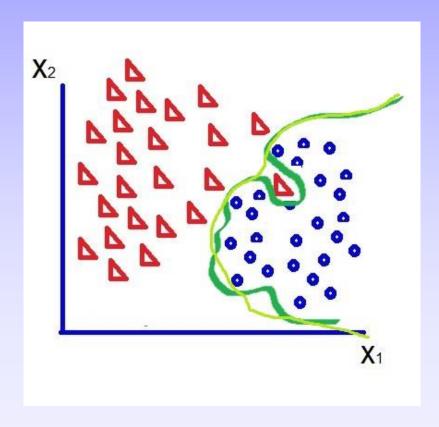
$$J(\theta) = -\left[\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))\right] + \frac{\lambda}{2m} \sum_{i=1}^{n} \theta_{i}^{2}$$



במקרה זה נמנע מהתאמת היתר גם אם יש הרבה מאוד תכונות, וגבול ההחלטה יהיה חלק יותר.

(Regularized Logistic Regression)

• לסיכום: הרגולריזציה מאפשרת שימוש במספר רב של תכונות ומונעת התאמת יתר.



(Regularized Logistic Regression)

י נעדכן לחוד, ונוסיף Gradient Descent את θ_0 נעדכן לחוד, ונוסיף את גורם הרגולריזציה.

Repeat {

$$\theta_{j} \coloneqq \theta_{j} - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)} \right) x_{j}^{(i)}$$

(Regularized Logistic Regression)

: כלומר

Repeat {

$$\theta_0 \coloneqq \theta_0 - \alpha \left[\frac{1}{m} \sum_{i=1}^m \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_0^{(i)} \right]$$

$$\theta_{j} := \theta_{j} - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)} \right) x_{j}^{(i)} + \frac{\lambda}{m} \theta_{j} \right]$$

$$j =$$
 $(1, 2, \dots, n)$

}

(Regularized Logistic Regression)

Repeat { $\theta_0 := \theta_0 - \alpha \left[\frac{1}{m} \sum_{i=1}^m \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_0^{(i)} \right]$

$$\theta_{j} := \theta_{j} - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)} \right) x_{j}^{(i)} + \frac{\lambda}{m} \theta_{j} \right]$$

$$j =$$
 $(1, 2, \ldots, n)$

 $J(\theta)$ של θ_{j} של הנגזרת לפי הביטוי בסוגריים הוא הנגזרת לפי

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

(Regularized Logistic Regression)

Repeat {

$$\theta_{0} := \theta_{0} - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)} \right) x_{0}^{(i)} \right]$$

$$\theta_{j} := \theta_{j} - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)} \right) x_{j}^{(i)} + \frac{\lambda}{m} \theta_{j} \right]$$

$$j = \emptyset \quad 1, 2, \dots, n$$
}

 כלל זה נראה זהה לכלל העדכון עם הרגולריזציה אותו בצענו עבור רגרסיה ליניארית, אבל, אבל כמובן זה לא אותו כלל, כיון שפונקציית ההיפותיזה היא פונקציה לוגיסטית.

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

(Regularized Logistic Regression)

• נשתמש בפונקציית ה- Matlab לביצוע אופטימיזציה (מינימיזציה) ללא אילוצים:

fminunc

$$\min_{x} f(x)$$
 : f(x) של המינימום את מוצאת את המינימום של

כאשר (x) היא פונקציה המחזירה סקלר, ו- x הוא וקטור או f(x) מטריצה.

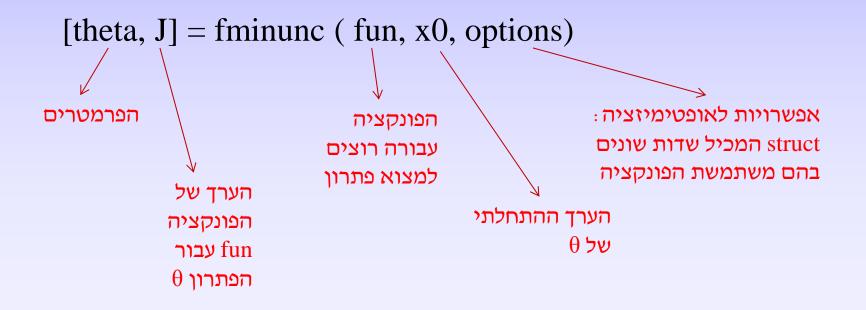
הפונקציה מבצעת אופטימיזציה לא ליניארית ללא אילוצים (unconstained nonlinear optimization).

הערה: אופטימיזציה עם אילוצים – הכוונה מציאה של אופטימום כאשר יש איזשהו ($||\theta||=1$ של הפרמטרים, לדוגמא $\theta>2$ או $\theta=0$

(Regularized Logistic Regression)

 $J(\theta)$ במקרה של רגרסיה לוגיסטית נרצה למזער את פונקציית המחיר - X, y עם הפרמטרים θ , המזעור יתבצע עם קבוצת דוגמאות האימון קבוצת דוגמאות קבועה.

: fminunc הסינטקס של הפונקציה



(Regularized Logistic Regression)

 $costF_log$ עבור רגרסיה לוגיסטית נשתמש בפונקציה הנקראת J ואת הגרדיאנט לכל θ .

options = optimset ('GradObj' , 'on', 'MaxIter' , 1000) ;

מספר מקסימלי של מודיע לפונקציה שהפונקציה
מספר מקסימלי של שלנו (costF_log) מחזירה
גם את המחיר וגם את

[theta, J]= fminunc (@(t) (costF_log(t,X,y)), theta0, options);

הגרדיאנט

דוגמא

```
% Initialize fitting parameters
initial_theta = zeros(size(X, 2), 1);
% Set regularization
lambda = 10;
% Set Options
options = optimset('GradObj', 'on', 'MaxIter', 5000);
% Optimize
[theta, J, exit_flag] = ...
  fminunc(@(t)(costFunctionReg(t, X, y, lambda)), initial_theta, options);
```

Anonymous Functions

What Are Anonymous Functions?

- •A function that is *not* stored in a file, but is associated with a variable whose data type is function_handle.
- •Anonymous functions can accept inputs and return outputs, just as standard functions.
- •However, they can contain only a single executable statement.

Example: create a handle to an anonymous function

$$sqr = @(x) x.^2;$$

- Variable sqr is a function handle.
- The @ operator creates the handle
- the parentheses () after the @ operator include the function input arguments.
- The anonymous function in this case accepts a single input x, and implicitly returns a single output, an array the same size as x that contains the squared values.

What Is a Function Handle?

 A function handle is a MATLAB® data type that stores an association to a function. Indirectly calling a function enables you to invoke the function regardless of where you call it from.

Typical uses of function handles include:

- Pass a function to another function (often called function functions). For example, passing a function to integration and optimization functions, such as integral and fzero.
- Specify callback functions. For example, a callback that responds to a UI event or interacts with data acquisition hardware.
- Construct handles to functions defined inline instead of stored in a program file (anonymous functions).
- Call local functions from outside the main function.

Creating Function Handles

- You can see if a variable, h, is a function handle using isa(h,'function_handle').
- To create a handle for a function, precede the function name with an @ sign. For example, if you have a function called myfunction, create a handle named f as follows:
- f = @myfunction;

Calling a function handle

- You call a function using a handle the same way you call the function directly.
- For example, suppose that you have a function named computeSquare, defined as:
- function y = computeSquare(x) y = x.^2; end
- Create a handle and call the function to compute the square of four.
- f = @computeSquare; a = 4; b = f(a)
- b = 16

 Find the square of a particular value (5) by passing the value to the function handle, just as you would pass an input argument to a standard function.

```
a = sqr(5)
>> a = 25
```

Example:

 find the integral of the sqr function from 0 to 1 by passing the function handle to the integral function:

```
q = integral(sqr, 0, 1);
```

- No need to create a variable in the workspace to store an anonymous function.
- Instead, you can create a temporary function handle within an expression, such as this call to the integral function:

```
>> q = integral(@(x) x.^2,0,1);
```

Variables in the Expression

- Function handles can store not only an expression, but also variables that the expression requires for evaluation.
- For example, create a function handle to an anonymous function that requires coefficients a, b, and c.

```
>> a = 1.3; b = .2; c = 30;
```

>> parabola = $@(x) a*x.^2 + b*x + c;$

- Functions with Multiple Inputs or Outputs
- Anonymous functions require that you explicitly specify the input arguments as you would for a standard function, separating multiple inputs with commas.
- Example: a function with two inputs, x and y:

```
>> myfunction = @(x,y) (x^2 + y^2 + x^*y); x = 1; y = 10; z = >> myfunction(x,y)
```

- However, you do not explicitly define output arguments when you create an anonymous function. If the expression in the function returns multiple outputs, then you can request them when you call the function. Enclose multiple output variables in square brackets.
- Example: the ndgrid function can return as many outputs as the number of input vectors.
- This anonymous function that calls ndgrid can also return multiple outputs:

```
>> c = 10; mygrid = @(x,y) ndgrid((-x:x/c:x),(-y:y/c:y));
>> [x,y] = mygrid(pi,2*pi);
```

 The output from mygrid can be used to create a mesh or surface plot:

```
>> z = \sin(x) + \cos(y);
```

>> mesh(x,y,z)

