

Abdelrahman Rezk

Web Developer

NLP & ML student

AOU University

في 3 week هنتدى بعد شغلنا على ال linear regression والى هو كان بيحاول يتوقع حاجه معينه بتختلف وملهاش قيمه او قيمه محدده بل هى توقع لقيمة ممكن تزيد او تقل على حسب ال features ولكنه فى الآخر نوع من انواع ال supervised learning واللى بيشتغل على حاجه زي توقع مثلا سعر شقه ما ... نبتدى بقا نتكلم عن تانى حاجه بعد ال linear regression وهى ال classification problem وديه ليها ال algorithm معروف اسمه logistic regression وده بيتشغل على الحاجات الى بيكون قيمه محدده يعنى انا معايا برضه features ولكن ال y-actual هى قيم محدده ممكن تكون binary classification بمعنى انى تكون قيم ال y كلها اما 0 او 1 او اكثر من كده ولكنها فى الآخر بتكون قيم محدده بظبط زى انى احاول اعمل classification for mail is it a spam or not لذلك ديه تعتبر binary classification problem. لازم اكون عارف انى ضرب الثنيتات فى ال xs فى الآخر المفروض يطلع قيمه واحده. اخلى بالى من موضوع ال data regularization وهو انى بظبط البيانات الى عندى او اعرف انى هاستخدمها ازاي وده ببساعدنى فيه شخص متخصص فى المجال الى انا بحاول اعمله مودل او ماشين ليرنج

محتاج تركز فى ال sigmoid function وفي القيم الى بتخرجها والرسمه بتاعتها

In all of these problems the variable that we're trying to predict is a variable y that we can think of as taking on two values either zero or one, either spam or not spam, fraudulent or not fraudulent, related malignant or benign.

Classification

To attempt classification, one method is to use linear regression and map all predictions greater than 0.5 as a 1 and all less than 0.5 as a 0. However, this method doesn't work well because classification is not actually a linear function.

The classification problem is just like the regression problem, except that the values we now want to predict take on only a small number of discrete values. For now, we will focus on the **binary classification problem** in which y can take on only two values, 0 and 1. (Most of what we say here will also generalize to the multiple-class case.) For instance, if we are trying to build a spam classifier for email, then $x^{(i)}$ may be some features of a piece of email, and y may be 1 if it is a piece of spam mail, and 0 otherwise. Hence, $y \in \{0, 1\}$. 0 is also called the negative class, and 1 the positive class, and they are sometimes also denoted by the symbols “-” and “+.” Given $x^{(i)}$, the corresponding $y^{(i)}$ is also called the label for the training example.

Hypothesis Representation

We could approach the classification problem ignoring the fact that y is discrete-valued, and use our old linear regression algorithm to try to predict y given x . However, it is easy to

construct examples where this method performs very poorly. Intuitively, it also doesn't make sense for $h_{\theta}(x)$ to take values larger than 1 or smaller than 0 when we know that $y \in \{0, 1\}$. To fix this, let's change the form for our hypotheses $h_{\theta}(x)$ to satisfy $0 \leq h_{\theta}(x) \leq 1$. This is accomplished by plugging $\theta^T x$ into the Logistic Function.

Our new form uses the "Sigmoid Function," also called the "Logistic Function":

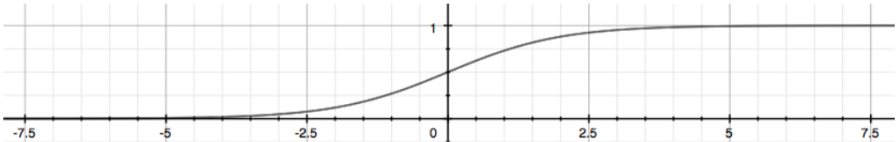
Our new form uses the "Sigmoid Function," also called the "Logistic Function":

$$h_{\theta}(x) = g(\theta^T x)$$

$$z = \theta^T x$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

The following image shows us what the sigmoid function looks like:



The function $g(z)$, shown here, maps any real number to the $(0, 1)$ interval, making it useful for transforming an arbitrary-valued function into a function better suited for classification.

$h_{\theta}(x)$ will give us the **probability** that our output is 1. For example, $h_{\theta}(x) = 0.7$ gives us a probability of 70% that our output is 1. Our probability that our prediction is 0 is just the complement of our probability that it is 1 (e.g. if probability that it is 1 is 70%, then the probability that it is 0 is 30%).

$$h_{\theta}(x) = P(y = 1 | x; \theta) = 1 - P(y = 0 | x; \theta)$$

$$P(y = 0 | x; \theta) + P(y = 1 | x; \theta) = 1$$

<https://www.coursera.org/learn/machine-learning/supplement/N8qsm/decision-boundary>

<https://www.coursera.org/learn/machine-learning/supplement/bgEt4/cost-function>

انا بعد ما اعمل training data لل data بتاعى وجبت قيم ال theta المناسبه الى هي بتطلع predication تمام وكمان ضبط ال cost function ببيتدى ناخذ قيم ال theta ديه او نحفظها عشان لما نروح نعملها على test set بنشتغل بالقيم بتاعت الثيتات الى انا جببتها من ال training data

<https://www.coursera.org/learn/machine-learning/supplement/0hpMl/simplified-cost-function-and-gradient-descent>

<https://www.coursera.org/learn/machine-learning/supplement/cmjlC/advanced-optimization>

<https://www.coursera.org/learn/machine-learning/supplement/HuE6M/multiclass-classification-one-vs-all>

<https://www.coursera.org/learn/machine-learning/supplement/VTe37/the-problem-of-overfitting>

<https://www.coursera.org/learn/machine-learning/supplement/1tJlY/cost-function>

ملخص

ال classification هو نوع من انواع ال supervised learning الى بيكون مهتم اكثر بالحاجات الى ليها قيم محددة تسمى discrete values والى هي عبارته عن مدخلات ومخرجات محددة غالبا ما بتكون binary classification عنى المخرجات اما حاجه من الاثنين ممكن نقول 0 و 1 وعن طريق ده بيتدى اشتغل وافصل الاثنين عن بعض عن طريق ال logistic regression.

Ln is standard for logarithm natural number

ال natural number هو $e = 2.718$

عن طريق ال sigmoid function قدر اخلى قيم التوقعات عندى دايم بين ال 0 و 1 وده بيساعدنى انى اعمل classification وبناء عليها بشتغل على فكره ال probability انى لو كان مثلا الرقم 7 فيقريبه مثلا ل 1 وهكذا

Activities Firefox Web Browser Mon 11:32 PM

D 03.pdf - Mozilla Firefox

Genral Tasks | (3) Hesham As | Google Transl | My Drive - Goo | Week 3 Classi | Lecture Slides | _964b8d77dc0ee | D 03.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford mahcine learning/week 4/

7 of 12 Automatic Zoom

أساسيات التصنيف

المعادلة :

- المقصود بالـ $\theta^T x$ معادلة حاصل ضرب الثبتات في اللكسات مثلما فعلنا في التنبؤ
- قد تكون معادلة من الدرجة الاولى او الثانية او أكثر , حسب كمية التعقيد في المسألة

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

بعد كده ببندى احسب ال cost function الى عن طريقها هاقدر اعرف الدنيا كويسه عندى ولا لا وده عن طريق المعادلة المجمعه.

اخلى بالى من انى معادله التوقع الى فوق ديه هى الى بعد اما اعمل training للداتا هابتدى استخدمها عشان اتوقع اى داتا جديده بعد اما دربتها وجبت قيم الثبتات الى انا عايزها.

Activities Firefox Web Browser Mon 11:48 PM

D 04.pdf - Mozilla Firefox

Genral Task | (3) Hesham | Google Tran | My Drive - C | Week 3 Cla | Lecture Slid | _964b8d77dc0 | D 04.pdf | D 05.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford mahcine learning/week 4/

7 of 8 Automatic Zoom Clockify

المعادلة المستخدمة للتصنيف

المعادلة المجمعة :

- يتم تجميع المعادلتين معا , فى كلتا حالتى y , حينما تكون بصفر ويواحد , بالصيغة التالية :

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

- حينما تكون y تساوي صفر , يختفى الجزء الاول من المعادلة و تصير هكذا :

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

- حينما تكون y تساوي 1 , يختفى الجزء الثانى من المعادلة و تكون هكذا :

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

بعد كده ببندى اجيب قيم الثبتات الى بتعمل fit للمعادله عن طريق ال graident descent نفس المعادله يعتبر ولكن التعبير عنها مختلف فيستخدم فى القيمه المتوقعة ال sigmoid function

Activities Firefox Web Browser Tue 12:03 AM

D 05.pdf - Mozilla Firefox

Genral Tasks | (3) Hesham As | Google Transl | My Drive - Go | Week 3 Classi | Lecture Slides | _964b8d77dc0ee | D 05.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford machine learning/week 4/

3 of 12 Automatic Zoom

مثال عملي للتصنيف

ايجاد الثيتا :

- كي لا ننسى , الثيتا هي معاملات الإكسات , التي يتم ايجاجها , وذلك للحصول على معادلة الكيرف الاكثر ملائمة best fit curve
- يتم ايجاد الثيتا بهذه بالمعادلة , حيث :
- الرمز z يشير لرقم الثيتا المطلوبة ($0, 1, 2, \dots$) , أي انه سيتم تكرارها لكل الثيتات المطلوبة
- الفا هي معامل يشير لمقدار خطوة الحساب (مثلما فعلنا في التنبؤ)
- المعادلة بعدد الفا هي تفاضل المعادلة الكبيرة السابقة (مش لازم تعرف اثباتها)
- يتم وضع اكس 0 لثيتا صفر و اكس 1 لثيتا 1 و هكذا

$$\theta_j := \theta_j - \alpha \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

ولكنها بعد اما عملت اشتقاق ليها بقت بالمنظر ده

Activities Firefox Web Browser Mon 11:51 PM

D 05.pdf - Mozilla Firefox

Genral Tasks | (3) Hesham As | Google Transl | My Drive - Go | Week 3 Classi | Lecture Slides | _964b8d77dc0ee | D 05.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford machine learning/week 4/

4 of 12 Automatic Zoom

مثال عملي للتصنيف

المعادلة كمصفوفات :

- يجب ان نقوم بتحويلها لمصفوفة للتعامل معها باي لغة برمجة , فستكون المعادلة كالتالى , حيث :
- ثيتا هنا هي مصفوفة عمود واحد , وفيها صفوف بعدد الثيتات ($n+1 \times 1$)
- الفا و m هي نفس الرموز السابق ذكرها
- اكس ترانسبوز , هي تدوير مصفوفة اكس الكبيرة , كانت ($m \times n+1$) , وستصير ($n+1 \times m$)
- اكس فى ثيتا , المقصود بها ضرب مصفوفة اكسات فى مصفوفة ثيتات
- الواي هنا هي مصفوفة قيم واي (اصفار وواحد) يتم طرحها منها ($m \times 1$)

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta) - \vec{y})$$

اخلى بالى من حاجتين مهمين هما ال over fitting and under fitting .
 واحاول اعمل regularization للداتا بالاستعانة بمختصين فى الحاجه الى انا شغال فيها لان ممكن يكون في بعض ال features الى مش مهمه والاستخدمها انا فتضيع الدنيا.
 ال regularization بيكون عن طريق واحد او اكثر من الحاجات ديه

Activities Firefox Web Browser Tue 5:59 AM

D 08.pdf - Mozilla Firefox

Genral Tasks | (3) Hesham As | Google Transl | My Drive - Go | Week 3 Classif | Lecture Slides | _964b8d77dc0ee | D 08.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford mahcine learning/week 4/

3 of 8 Automatic Zoom

تنعيم البيانات Regularization

المفهوم :

- عملية تنقية البيانات بهدف التخلص من الضبط الزائد OF
- يتم عبر 3 طرق هامة قد نقوم باحداها فقط او بثنين فقط او بالثلاثة معا :
 - * انتقاء البيانات
 - * تغيير المعاملات
 - * اضافة (لمدا)

Activities Firefox Web Browser Tue 6:00 AM

D 08.pdf - Mozilla Firefox

Genral Tasks | (3) Hesham As | Google Transl | My Drive - Go | Week 3 Classif | Lecture Slides | _964b8d77dc0ee | D 08.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford mahcine learning/week 4/

7 of 8 Automatic Zoom

تنعيم البيانات Regularization

إضافة لمدا λ :

- واحيانا يسمى معامل التنعيم Regularization factor
- ونقوم بضربه في مجموع مربعات جميع الثبتات (باستثناء ثبثا صفر) , ثم اضافته لمعادلة الكوست

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

- باختيار قيمة معينة للمدا , و بدأ ادخال المعادلة المحدثة في الخوارزم , نري أن الـ OF سيخف
- اذا ظل الـ OF موجود , نغير قليلا من قيمة لمدا و نجرب مرة اخري

Activities Firefox Web Browser Tue 6:02 AM

D 08.pdf - Mozilla Firefox

Genral Tasks x (3) Hesham A x Google Transl x My Drive - Go x Week 3 Classi x Lecture Slides x _964b8d77dc0ee x D 08.pdf x

file:///home/abdelrahman/Desktop/new_steps/stanford mahcine learning/week 4/

8 of 8 Automatic Zoom

Regularization تنعيم البيانات

إضافة لمدا λ :

- الصيغة المعدلة لمعادلة الكوست ستكون

Repeat {

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

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

}

- وهي جزئين لان الثيتا صفر لا يتم ضربها في المدا , والثانية لباقي الثيتات

- يتم تكرار العملية للوصول لافضل قيمة للثيتات , ويتم تغيير قيمة المدا اذا ما ظل الـ OF موجود

ال classification في الحقيقة عامل زى ال regression عدا انى هو بيحاول يتوقع الحاجه ديه فى قيم صغيره ومنفصله ومحدده except that the values we now want to predict take on only a small number of discrete values.

لما بجى اشتغل انا هاتجاهل خالص انى اصلا قيم ال y قيم محدده وهانستخدم ال linear regression شان يعمل predict for y given x and parametrized by thetas ولكن انا بغير طريقه ال predication function الى كنت بستخدامها مع ال linear regression عن طريق ال logistic function or sigmoid function.

Activities Firefox Web Browser Mon 11:32 PM

D 03.pdf - Mozilla Firefox

Genral Tasks x (3) Hesham A x Google Transl x My Drive - Go x Week 3 Classi x Lecture Slides x _964b8d77dc0ee x D 03.pdf x

file:///home/abdelrahman/Desktop/new_steps/stanford mahcine learning/week 4/

7 of 12 Automatic Zoom

أساسيات التصنيف

المعادلة :

- المقصود بالـ $\theta^T x$ معادلة حاصل ضرب الثيتات فى الانكسات مثلما فعلنا فى التنبؤ

- قد تكون معادلة من الدرجة الاولى او الثانية او اكثر , حسب كمية التعقيد فى المسألة

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

وده خلانى انى احول function بتطلع القيم بطريقة عشوائية ل function هاتطلع قيم محصوره بين قيمتين محددين وهما 0 و 1 وبناء على ده هابتدى اشتغل كأنها مسألة probability عن طريق لو كانت اكبر من قيم معينة تبقا 1 مثلاً والعكس تبقا 0.

$h_{\theta}(x)$ will give us the **probability** that our output is 1. For example, $h_{\theta}(x)=0.7$ gives us a probability of 70% that our output is 1. Our probability that our prediction is 0 is just the complement of our probability that it is 1 (e.g. if probability that it is 1 is 70%, then the probability that it is 0 is 30%).

Decision Boundary

عشان نقدر اخلی قيم ال hypothesis function تكون فعليا discrete values هانتحتاج نقول انی لو كان ال hypothesis بتاعنا اكبر من حاجه معينه بيقا 1 والعكس بيقا صفر

$$h_{\theta}(x) \geq 0.5 \rightarrow y=1$$

$$h_{\theta}(x) < 0.5 \rightarrow y=0$$

$$g(z) \geq 0.5$$

$$\text{When } z \geq 0$$

Facts

$$z=0, e^0=1 \Rightarrow g(z)=\frac{1}{2}$$

$$z \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow g(z)=1$$

$$z \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow g(z)=0$$

So if our input to g is $\theta^T x$ then that means:

$$h_{\theta}(x) = g(\theta^T x) \geq 0.5$$

$$\text{when } \theta^T x \geq 0$$

عشان فی ال sigmoid function القيمه ديه بتكون مرفوعه لك اس سالب فيتصغر جدا وبيتبقا 1 + القيمه الصغيره

From these statements we can now say:

$$\theta^T x \geq 0 \Rightarrow y=1$$

$$\theta^T x < 0 \Rightarrow y=0$$

The **decision boundary** is the line that separates the area where $y = 0$ and where $y = 1$. It is created by our hypothesis function.

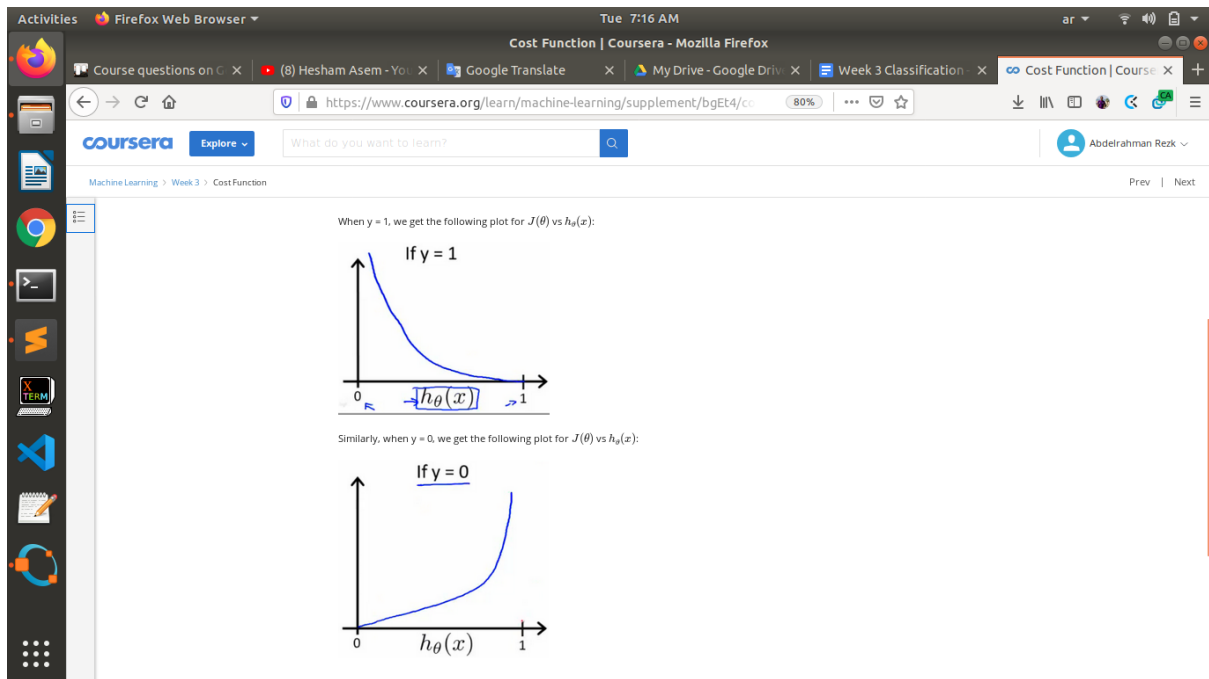
Cost Function

We cannot use the same cost function that we use for linear regression because the Logistic Function will cause the output to be wavy, causing many local optima. In other words, it will not be a convex function.

$$\text{Cost}(h_{\theta}(x), y) = 0 \text{ if } h_{\theta}(x) = y$$

$$\text{Cost}(h_{\theta}(x), y) \rightarrow \infty \text{ if } y=0 \text{ and } h_{\theta}(x) \rightarrow 1$$

$$\text{Cost}(h_{\theta}(x), y) \rightarrow \infty \text{ if } y=1 \text{ and } h_{\theta}(x) \rightarrow 0$$



Simplified Cost Function and Gradient Descent

We can compress our cost function's two conditional cases into one case:

Activities Firefox Web Browser Mon 11:48 PM

D 04.pdf - Mozilla Firefox

Genral Task | (3) Hesham | Google Tran | My Drive - C | Week 3 Cla | Lecture Slid | _964b8d77dc0 | D 04.pdf | D 05.pdf

file:///home/abdelrahman/Desktop/new_steps/stanford machine learning/week 4/

7 of 8 Automatic Zoom Clockify

المعادلة المستخدمة للتصنيف

المعادلة المجمعة :

- يتم تجميع المعادلتين معا , في كلتا حالتى y , حينما تكون بصفر وبواحد , بالصيغة التالية :

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

- حينما تكون y تساوي صفر , يختفى الجزء الاول من المعادلة و تصبح هكذا :

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

- حينما تكون y تساوي 1 , يختفى الجزء الثانى من المعادلة و تكون هكذا :

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

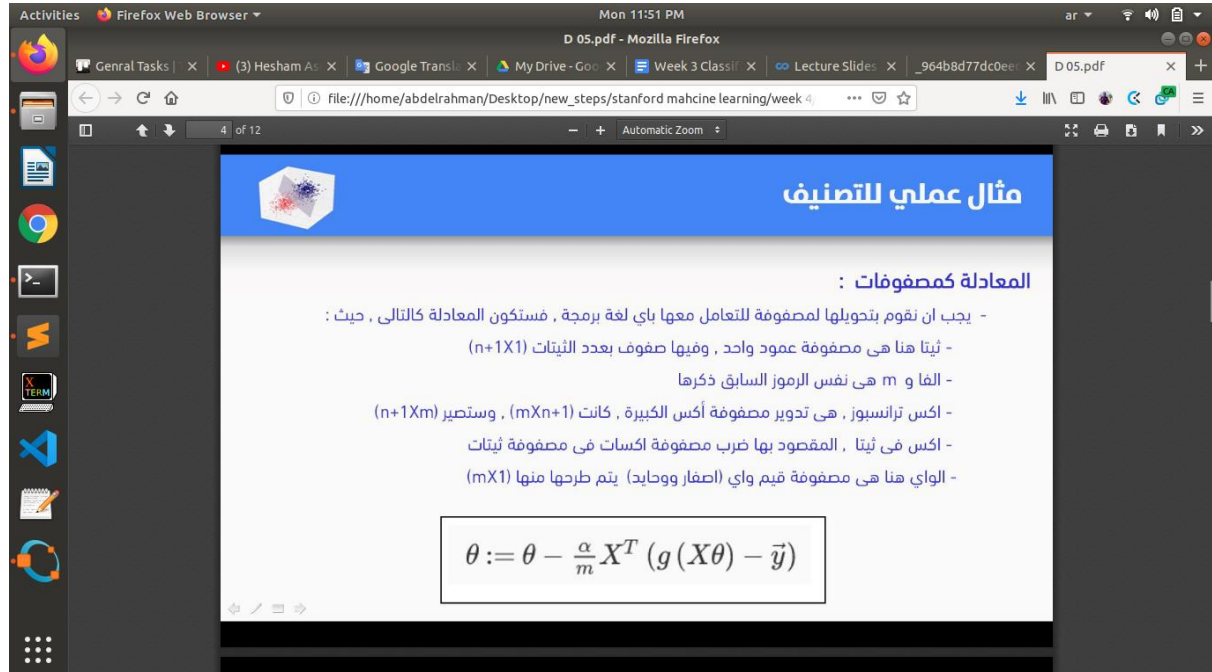
Instead of this form of gradient descent

Repeat{

$$\theta_j := \theta_j - \alpha / m \sum (h_{\theta}(x(i)) - y(i)) x_j(i)$$

}

We use A vectorized implementation of gradient descent is



Multiclass Classification: One-vs-all

<https://www.coursera.org/learn/machine-learning/supplement/HuE6M/multiclass-classification-one-vs-all>

هنا انا بشتغل على فكره انها binary انى اخذ مثلا كلاس معين اسميه 0 والباقي كله اسمه 1 وافصل اول كلاس ثم التانى مثلا اخذ اسمه 1 واسمى الباقي كله 0 بلس الى انا عملته وافصل التانى وهكذا لحد n-classes

Activities Firefox Web Browser Tue 8:19 AM

Cost Function | Coursera - Mozilla Firefox

Course questions on Coursera | Google Translate | My Drive - Google Drive | Week 3 Classification | Cost Function | Coursera | Machine Learning - Linear Regression

https://www.coursera.org/learn/machine-learning/supplement/1tJlV/c... 140%

coursera

Machine Learning > Week 3 > Cost Function Prev | Next

We could also regularize all of our theta parameters in a single summation as:

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

The λ , or lambda, is the **regularization parameter**. It determines how much the costs of our theta parameters are inflated.

Using the above cost function with the extra summation, we can smooth the output of our hypothesis function to reduce overfitting. If lambda is chosen to be too large, it may smooth out the function too much and cause underfitting. Hence, what would happen if $\lambda = 0$ or is too small?

✓ Complete Go to next item

Regularized Linear Regression

Note: [8:43 - It is said that X is non-invertible if $m \leq n$. The correct statement should be that X is non-invertible if $m < n$, and may be non-invertible if $m = n$.

We can apply regularization to both linear regression and logistic regression. We will approach linear regression first.

Activities Firefox Web Browser Tue 8:37 AM

Regularized Linear Regression | Coursera - Mozilla Firefox

Course questions on Coursera | Google Translate | My Drive - Google Drive | Week 3 Classification | Regularized Linear Regression | Machine Learning - Linear Regression

https://www.coursera.org/learn/machine-learning/supplement/pKAsc/ 120%

coursera Explore What do you want to learn? Abdelrahman Rezk

Machine Learning > Week 3 > Regularized Linear Regression Prev | Next

Gradient Descent

We will modify our gradient descent function to separate out θ_0 from the rest of the parameters because we do not want to penalize θ_0 .

$$\begin{aligned} \text{Repeat } \{ \\ \theta_0 &:= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_0^{(i)} \\ \theta_j &:= \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \right] \quad j \in \{1, 2, \dots, n\} \\ \} \end{aligned}$$

The term $\frac{\lambda}{m} \theta_j$ performs our regularization. With some manipulation our update rule can also be represented as:

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m} \right) - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The first term in the above equation, $1 - \alpha \frac{\lambda}{m}$ will always be less than 1. Intuitively you can see it as reducing the value of θ_j by some amount on every update. Notice that the second term is now exactly the same as it was before.

Regularized Linear Regression | Coursera - Mozilla Firefox

Machine Learning > Week 3 > Regularized Linear Regression

Normal Equation

Now let's approach regularization using the alternate method of the non-iterative normal equation.

To add in regularization, the equation is the same as our original, except that we add another term inside the parentheses:

$$\theta = (X^T X + \lambda \cdot L)^{-1} X^T y$$

where $L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \\ & & & & 1 \end{bmatrix}$

L is a matrix with 0 at the top left and 1's down the diagonal, with 0's everywhere else. It should have dimension (n+1)×(n+1). Intuitively, this is the identity matrix (though we are not including x_0), multiplied with a single real number λ .

Recall that if $m < n$, then $X^T X$ is non-invertible. However, when we add the term $\lambda \cdot L$, then $X^T X + \lambda \cdot L$ becomes invertible.

Regularized Logistic Regression

We can regularize logistic regression in a similar way that we regularize linear regression. As a result, we can avoid overfitting.

Regularized Logistic Regression | Coursera - Mozilla Firefox

Machine Learning > Week 3 > Regularized Logistic Regression

Gradient descent

Repeat {

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

$$\theta_j := \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right] \leftarrow \frac{\partial}{\partial \theta_j} J(\theta)$$

}

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

Mark as completed