Abdelrahman Rezk

Web Developer

NLP & ML student

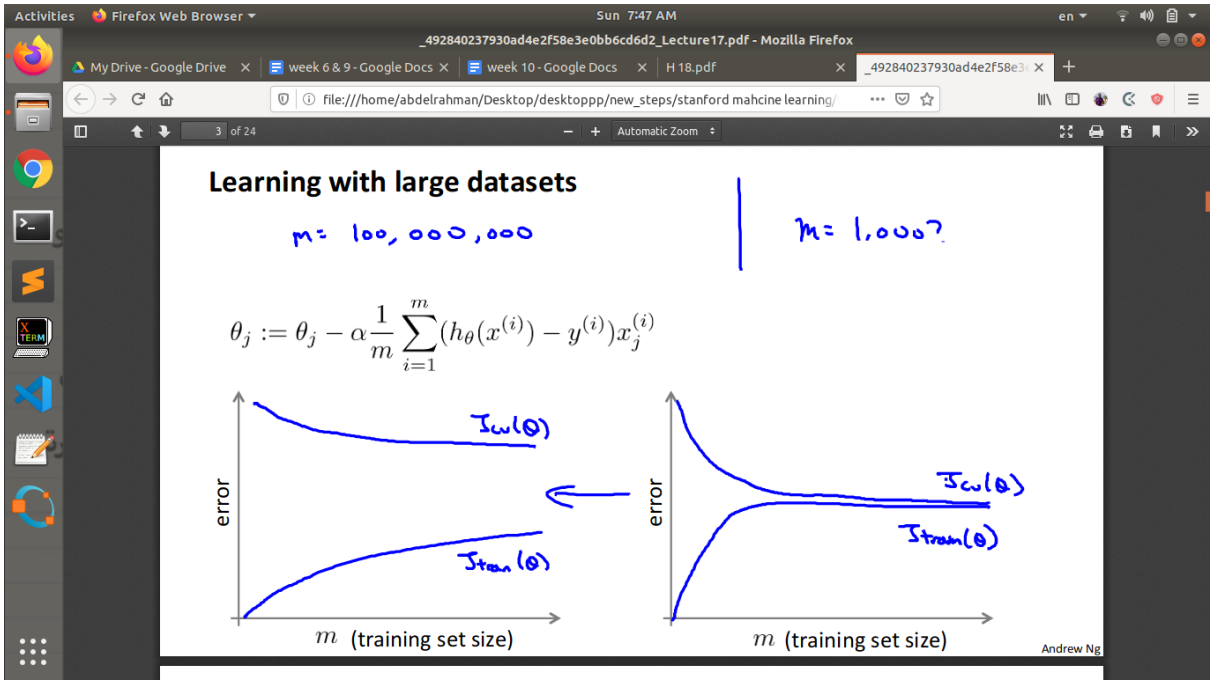
AOU University

Big Data
"It's not who has the best algorithm that wins, It's who has the most data".It is given by Andrew Ng

من المهم جدا قبل أى تفكير فى ال algorithm او اى حاجة تانية هو التفكير ال data نفسها لأن ال data هى الى بيتبنى عليها كل حاجه بعد كده وال data بتفرق جدا مع كبر حجمها وعدد ال features الى فيها وده بيختلف هل زيادة ال data مفيد بعد وقت معين ولا لا يعنى بعد استخدام algorithm ما على ال data لى عندى ممكن ساعتها اعرف زيادة عدد ال examples فى ال data هايكون مفيد ولا لا لكن قبل أى شىء المهم اول حاجة بنفكر فيها هي وجود ال data.

بتقابلنا مشكلة تانية الا وهي بعد وجود ال data بكميات ضخمة وهائلة كيفية التعامل مع ال data دي بيتم ازاى وانا هنا بتكلم عن ملايين الصفوف من ال data مش مجرد 1000 صف بنتعامل معاهم فهنا اى algorithm مثلا لو قلنا ال linear regression فهنا العملية بتاعت ضرب ال thetas فى xs وطرحها من ال actual value الى هى ال y وبعد ده كله اجمعهم فأنا هنا لو بتكلم فى مليون مثال فقط العملية بتكون مكلفه جدا سواء على ال machine او ك cost او ك time ما بالك بقا لو ساعتها محتاجين نعمل gradient descent عشان نبتدى نجيب القيم المثلى لل thetas الى بتقلل ال cost function لل global minimum فلو قلنا عدد الـ 100 terations فانتا هتضرب 100 فى عدد ال examples الى انت بتحاول تنطبق عليها المعادلة والكلام ده ممكن يتطلب اسابيع او اكثر عشان يتم عمله.

ومن الحاجات الى ممكن تساعد فى كده هو انى اعمل shuffle لل data الى عندى ثم ابتدى انى استخدم جزء منهم و قسمه بال cross validation ارسم ال cost function واشوف ساعتها هل جزء البيانات ده كافى ولا لا انى اعمل model كويس يعمل fitting for data.

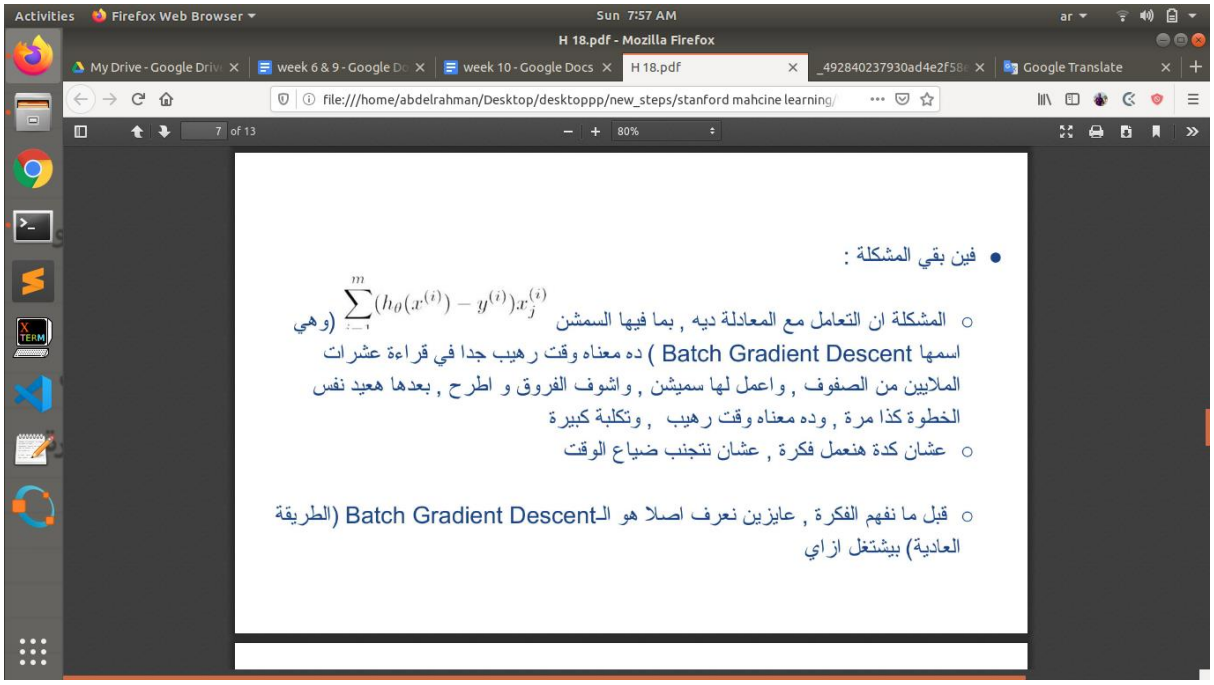


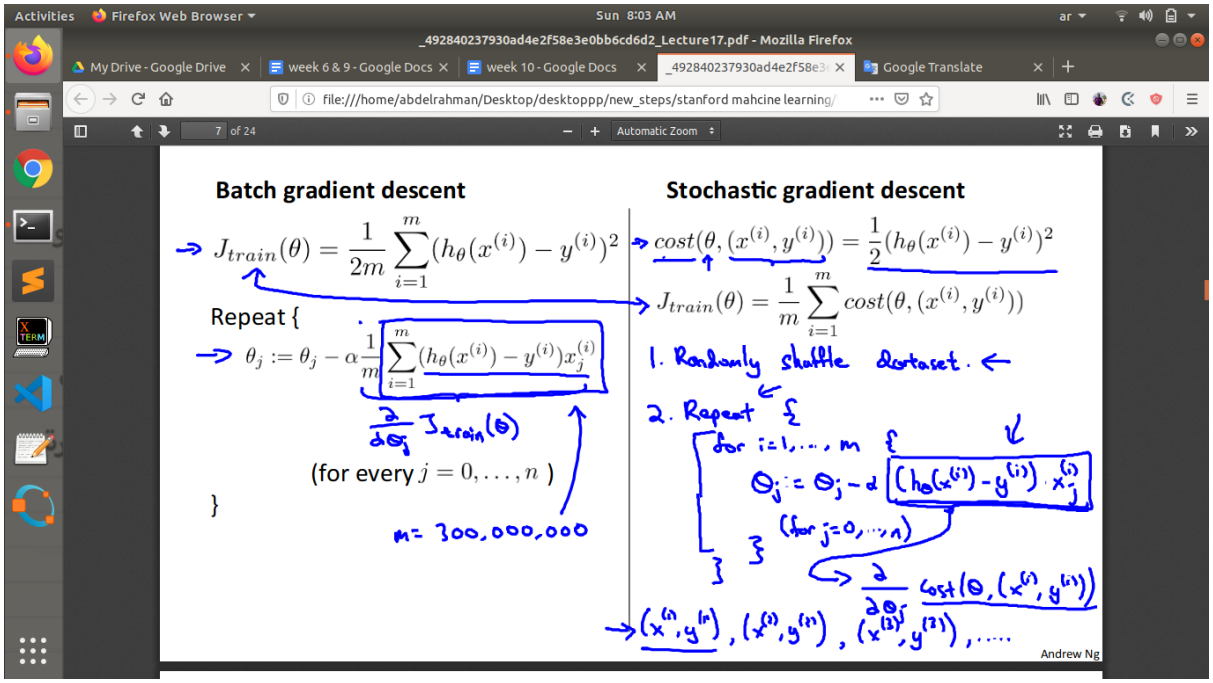
Stochastic gradient descent

المشكلة الى بتبدأ تظهر عندى مع وجود بيانات ضخمة هو إني بحاول قلل قيمة ال cost function عن طريق ال thetas parameters الى بحاول انى اشوف القيمة ال optimal ليها عشان قدر استخدمها فى الاخر مع اى data تانية انا مشفوتهاش وبعتبر انى قيم ال thetas ديه بتكون المثلى وانها هى بتقلل قيمة ال cost function الى عندى لل global minimum.

ولكن تبتدي تظهر مشكله معايا انى عشان قلل قيم ال thetas ديه محتاج انى اعمل gradient descent الى بيكون فيه تقليل ال thetas ديه مكلف جدا لانى مبقاش عارف بظبط انى محتاج اعمل كام iteration عشان قدر اوصل للقيم المثلى لل thetas ومع كل iteration انا بعمل عملية على بيانات كبيرة جدا وده بيكون مكلف جدا فى كذا حاجة سواء وقت فلوس او على ال machine نفسها.

وهنا ببتدى افكر فى طريقة تانية للتعامل مع ال data وهى Stochastic gradient descent التدرج العشوائى وهو إنى ببتدى بدل ما كنت بغير قيم ال thetas بعد كل teration وعمل مجموع لفرق كل قيم ال predictive من ال actual وبعد كده نضربهم فى عدد الامثلة اللى عندى والى ممكن يوصل لاكتر من مليون. هنا بقا ببتدى افكر فى طريقة زى الى فوق ديه وانى هغير قيم ال theta بناء على كل مثلا لوحده. بس لازم اكون اولا عامل shuffle لل data لى عندى وتكون مرتبة بشكل عشوائي عشان قدر اشتغل بالطريق دى وانى قيم ال theta لما هاتتغير مش هتكون بتتغير بناء على رؤية جزء معين من ال data لا هى هاتتغير بعد كل مثال لحد أما اوصل للقيم المثلى ليها.





Mini--- batch gradient descent

فيه طريقة تانية بتدى من خلالها أشتغل مع ال data هى ال mini-batch هى ببساطة انى بمسك ال data الى عندى مقسمة لأجزاء معينة مثلا فى range 100-2 فلو كان عندى مثلا ال data مليون صح وعايز قسمهم لاربع اجزاء فقسمهم ربع مليون لكل جزء وبعدها ببتدى أشتغل بطريقة الـ batch gradient descent.