Post 1

مش لازم اروح اعمل Grdient Descent علطول عشان اجيب ال Params الى بتعمل minimize for cost function
لا أنا ممكن عن طريق ال Normal Equation اروح اشتغل عادى جدا واجيب Direct ال Params ديه بدل ما اعمل Iterations انا مش عارف عددها ممكن يكون كام عشان اوصل لافضل قيم لل Params ديه ولكن ده دايما مش Applicable بحيث انى فى بعض الحالات ده مش هينفع زى مثلا لو عندى داتا خاصه بالصور وحجم الصور ده كبير الصورة الواحده ممكن تكون اكتر من 10000 بيكسل مثلا فى الحالة ديه بروح لل Grident Descent وده لان ال Complixity هنا بتكون قريبه من O(n3) لكن لو كان عندك عدد Features مناسب وفى نفس الوقت الميمورى بتاعتك تقدر تشيل ال Data الى عندك فى الحاله ديه ممكن تروح Dierct تجيب ال Params
.Theta_hat: is the value of theta that minimize the cost function very well
T: transpose means make convert rows to columns and columns to rows
.Is the inverse of the matrix :1-

https://www.linkedin.com/posts/abdelrahman-rezk_machinelearning-mathematics-activity-6800158010577612800-GunP

Post 2

٨٨ الى عنده معلومات زيادة ياريت ينورنا فى الكومنتات
ال Gradient Descent هو عبارة عن iterative process بيروح فيها يعمل Tweak لل model paramters الى بتقلل ال cost function
فى كل خطوه هو بيحاول يشوف ال cost function بالنسبة لل params الجديدة بعد ما عملتلها tweak وبعد كده بيحاول يشوف الخطوه الجاية المفروض ينحدر فى انهى اتجاه وده بيختلف على حسب ال slope وطبيعة ال function كمان بتفرق فى حاجة زى ال Mean Square Error هى اصلا عباره عن Convex shape فده بيساعد ال model يروح لل Global minium لان مفيش اصلا Local Minium لكن بيظهر عندى مشكلة فى حجم الخطوه الى بحاول اخدها عشان مخدش وقت كبير جدا عشان او اصل لل Global او انى يحصل diverge وهو انى انتهى فى اماكن بعيده عن ال Global وده الى احنا انسمية learning rate وفى الغالب احنا بنقلله كلما اتجهنا لل Global عشان حجم الخطوه يقل زى ما واضح فى ال graph كمان انى اخلى ال features تكون scaled بيساعد الموديل يوصل اسرع بعيدا عن لو فيه قيم كبيرة اثناء العمليات الى بتحصل ممكن يحصل عندكArithmetic overflow
الجاى هيكون عن ال Approaches المختلفة فى ال Gradient Descent إن شاء الله

https://www.linkedin.com/posts/abdelrahman-rezk_machinelearning-linearregression-activity-6800638120002932736-pNng

Post 3

ال sklearn بتوفر بعض ال model paramters الى هى مرتبطة بالموضوع ده ال learning rate هو المعروف ب eta وال tollerance وهو عامل زى علامه ال summation كده وده الفايدة بتاعته هو التحكم فى عدد ال iterations بمعنى انى بعد عدد معين الموديل بتاعك بيغير فى ال cost function حاجه لا تذكر تغير بسيط جدا لانه بيكون قرب جدا جدا من امثل حل وفى الحاله ديه انتا محتاج توقف ال iterations بتاعتك وهنا ال tolerance ده بيروح يشوف الفرق بين two gradient vectors بتاع iteration والى جايه او الى قبلها ويشوف الفرق لو كان اظغر من قيمة معينه الى احنا كتير بنسميها epsilone ايام ال problem solving الجميلة بيروح يوقف الموديل بدل ما يكون بيعمل iteration على
الفاضى

https://www.linkedin.com/posts/abdelrahman-rezk_machinelearning-linearregression-gradients-activity-6800876316737589248-Kq2L

Post 4

ال Stochastic Gradient Descent هو عباره عن out-of-core model بمعنى انى من خلاله ققدر اشتغل على online learning model وده لانه سريع جدا بسبب انه بياخد one-instance فى كل مره وبيعمله feed للموديل بتاعك على عكس ال Batch الى كانت بتاخد كل الداتا وتعملها feed بس فى نفس الوقت كلمة stochastic ديه بتعنى انى بياخد random instance فى كل مرة وممكن يحصل overlapping فى نفس الوقت هنلاقى ال cost functions بتعلى وتوطى لكن بعد فترة معينه هنلاقيها بتنزل وده بسبب انى الموديل لسه مشفش داتا كتير وال cost function فى الحالة ديه بتكون غير منتظمة طبيعى للتعلم غير منتظم ولكن ده بيساعد الموديل انه يهرب من ال local minimum بس بيقرب لل global minimum ولكن مش احسن حاجه وده من خلال ال learning rate وانى اعمل learning schedule بمعنى انى فى كل ما ققرب من ال global قلل الخطوة بتاعتى ده بيساعد كتير انى اوصل لل global وممكن اتجنب موضوع ال randomness ده عن طريق انى اعمل shuffle وامشى فى ال process بتاعتى من اول instance لاخر واحد ورا بعض ولكن اخلى بالى من عملية ال shuffle بحيث ميكونش فيه class ما مثلا بيجى ورا بعض بشكل متكرر كتير جدا

(sgd_reg = SGDRegressor(max_iter=1000, tol=1e-3, penalty=None, eta0=0.1

كده انا ققدر اعمل train for SGD regressor ولكن من غير اى regularization ال tol ده بيوضح انى لما اوصل انى ال loss function بتقل بقيمة اقل من .001 اوقف الموديل بدل ما اعمل iterations على الفاضى وال eta هى ال learning rate

https://www.linkedin.com/posts/abdelrahman-rezk_machinelearning-regression-activity-6805212107353714688-Vpj9

Post 5

ال Mini-batch هو مرحلة وسط ما بين ال Batch وبين ال Stochastic لا منك هتاخد كل الداتا مرة واحدة فيكون بطيء جدا ولا منك هتاخد one-instance بحيث يكون سريع جدا وكمان فى ميزة وديه ليها علاقة بال hardware design مع ال matrix operations فمثلا ال train بيكون اسرع مع 32 instances او مضاعافات ال 32 زى بظبط ال ram فى نفس الوقت ال Mini-batch على عكس ال Stochastic يقرب اكتر لل global minimum فى وقت اقل وده لانى باخد جزء من الداتا وكل ما كان الجزء اكبر كان قربة سريعا لل global اكتر ولكن ده ممكن يخلية يقع فى local minimum وهنلاحظ هنا الفرق بين ال 3 وازاى ال Batch علطول راح لل minimum بنما الاتنين التانين حوالين ال global minimum
ولكن طبعا ال Batch بياخد وقت كبير جدا فى كل teration لانه بيرن على كل الداتا
والجدول ده ملخص لل linear model الى اكلمنا عليهم فى الكام بوست الى فاتوا ولسه مكملين بإذن الله
ال SVD standard for Singular value decomposition
الى بتحاول تلاقى inverse فى حالة انى ال singular matrix مع ال Normal equations

https://www.linkedin.com/posts/abdelrahman-rezk_machinelearning-linearregression-activity-6805430258016301057-ORMo

Post 6

أنا كنت فاكر إنى لما الداتا تكون Non-linear وانا اروح اعمل polynomial features كده انا بعمل complex model بس الحقيقة هو انى الداتا نفسها هى الى complex وعن طريق انى بحاول ازود عدد ال features ده ف كانى بعمل combination بين ال features وبعدها وده بيساعد الموديل انه يقدر يلاقى علاقة بين ال features وبعضها لانها بتبقا مرتبطة ببعض عن طريق pattern أوضح زى ال second degree مثلا
لو خدنا المثال البسيط الى فى الصورة ده هنلاقى مثلا انى y او ال function of y لما نديها ال x بتروح تضربه فى رقم وكمان تعمله تربيع وتضربة فى رقم تانى + ال intercept وبجمع عليها random value
فلما انا اروح اعمل polynomial features من ال x الى هى كانى بروح اخلى ال x تشبة معادلة ال y ف الموديل ساعتها يقدر يربط الاتنين ببعض ويعمل map from x to y بطريقة افضل و فى الاخر ممكن استخدم linear model عادى جدا

https://www.linkedin.com/posts/abdelrahman-rezk_linearregression-models-machinelearning-activity-6805896015653220352-eu1l

Post 7

مهم جدا إنى اعرف هل الموديل بتاعى بي overfit ولا شغال كويس ولا اصلا underfit

ببساطة ال over fitting هو فى حالة انك تكون بتحقق score عالى على نفس الداتا الى عملت عليها train بس لما جيت تشوف الموديل على داتا جديدة او ما يسمى validation set لقيت انى ال score بتاعك قل كتير.
ال over fitting العكس ليه هو انك تكون فى مشكلة زى ال under fitting وهو انك اصلا ال score على الداتا الى عملت عليها train قليل جدا.
والحل هنا بيختلف من انك تزود مثلا الداتا سيت بتاعتك فى حالة ال over fitting ف الموديل يقدر يعمل Generalization او عن طريق استخدام ال Regularization على عكس حاجه زى ال under fitting انتا اصلا على نفس الداتا الى عملت عليها train مطلع score قليل ف هنا زيادة الداتا مش هيؤدى لشىء فبكون محتاج حاجة تقدر بيها اعمل detect ل pattern اكتر لان غالبا الداتا بتاعتك فيها complex pattern والموديل بتاعك بسيط جدا زى انى تكون الداتا Non-linear وانتا شغال ب linear model وممكن يكون المشكلة عندك فى الداتا نفسها وهى انى فيها outliers كتيرة او missing values كتير وهنا محتاج تقضى وقت اكتر فى انك تظبط الداتا.
فهنا انا بقدر اعرف المشكلة عندى فين ولكن عن طريق ال plots الموضوع بيبقا اوضح وهو الى احنا بنسمية ال Learning Curve لان ممكن الارقام نفسها تكون حاجه Abstract ولكنها فى الاخر بتدل برضه على شىء على عكس ال Graph العين بتقدر انها تلقط الأشياء سريعا وتلاقى تفسير للارقام ديه وهنا فى الجراف بيوضح الفرق بين المشكلتين وبين الحل الأمثل وهو فى poly degree = 2 ولكن ده لان الداتا اصلا معمولها genrate عن طريق quadratic equation لكن فى الغالب انا مبكونش عارف الداتا ديه اتكونت ازاى .

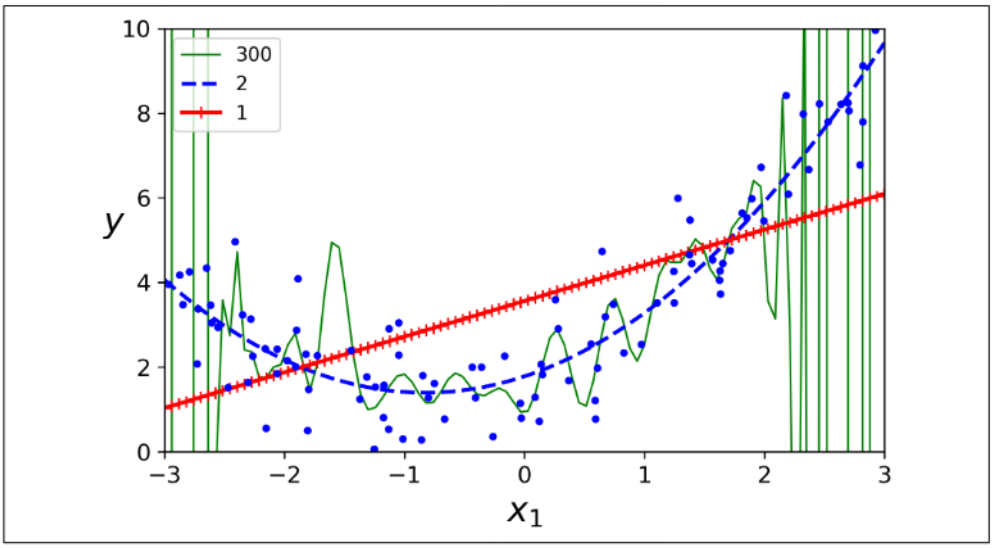

Figure 4-14. High-degree Polynomial Regression

Post 8

عشان نقدر نعالج مشكلة ال over fitting فيه اكتر من طريقة لكنها بتختلف بإختلاف ال model الى انا مستخدمه ، مثلا لو كنت مستخدم polynomial degree عالية تقليلها ممكن يعالج مشكلة ال over fitting لانك هتكون قللك ال constrain بتاعت الموديل ، والى بيخلى ال model يعمل over fitt هو انه بيكون فيه High Variance فى الداتا بتاعتك بمعنى تنوع كتير ، والموديل هنا وهو بيعمل train بيحاول انه يعمل tweak للمتغيرات على اساس كل ال patterns او الاختلافات الى بيشوفها فى الداتا ، رغم انى بعض الامثلة ممكن تكون لا تمثل اى شىء مجرد noise او وجودها فى الداتا بيمثل فى الاخر مجرد instance واحد فقط مجرد row فى الداتا ، وهنا انا بحتاج انى قلل ال constrain بتاعت الموديل درجة الحرية بتعبير تانى فى تعاملة مع الداتا عشان قدر اعمل generalization ولما الموديل يشوف داتا جديدة يقدر انه يطلع نتائج كويسة ، لكن ساعات ال plain regression model نفسه بيعمل over fitting الى هو يعتبر ابسط موديل عندى ، وفى حالة زى كده بحتاج انى اما ازود عدد ال training data او انى اعمل ما يسمى ب regularization وال regularization ده نفسه بيختلف فى عندى Ridge Regression

وفي lasso regression وغيرهم وهو بيختلف فى الجزء الى بضيفة لل cost function عشان اعمل regulaization.

Post 9

ال regularization ببساطة هو إنى بعد ما ال model عمل tweak للمتغيرات هروح انا اضيف عليها حاجة بسيطة تخليها متكنش حساسة لكل instance على حدا عشان ميحصلش over fitting ، والجزء الى بضيفة ده بيختلف ، فى البوست ده هنشوف واحد من ال Regulaizatio term وهو ال Ridge Regression او الى نسمع عنه ال L2 Norm وهو انى اضيف الجزء الى فى الصورة بعد علامة + لل gradient vector of cost function بتاعى ، وده بيحصل فقط فى ال training وساعت اصلا ال cost function الى بستخدمها فى ال training بستخدم evaluation method تانية غيرها فى ال testing ، وهنا بقا ال l2 ده وعن طريق hyper parameter تانى هو lambda او زى ما هو فى sklearn اسمه alpha وهنا زاد عندى متغير تانى محتاج اظبطة بجانب ال learning rate الى بيتحكم فى ال step بتاعت الموديل فى ال learning ، و alpha هنا لو كان بصفر هنلاقى انى الترم كله بعد ال + بقا سفر وكانى معملتش اى شىء وعبارة عن linear model الى عادى الى بستخدمه ، ولو كان قيمة كبيرة جدا هلاقى انى ال weights الاخيرة بقت قريبة جدا من 0 وبدل ما كنت فى مشكلة over fitting بقيت فى مشكلة تانية وهى ال under fitting زى ما هو واضح من ال graph هنلاقى ال weights بترجع لورا تانى وبتقرب من الصفر ، ال ridge regression كمان قدر اطبقة مع ال closed form ال Normal equation الى بتجبلى ال weights النهائية من خطوة واحده وتقدر تستخدم الاتنين عن طريق اما Ridge model او عن طريق SGDRegressor من sklearn. غير ال Ridge فى lasso regression وده بيحاول يخلى ال weights المرتبطة مع ال features الاقل اهمية تكون باصفار وبكدة يعمل features selection لل features المهمة وفى regularization وسط بينهم وهو ال Elastic Net وممكن قدر من خلاله اعمل Ridge فقط او Lasso فقط او انى يكون mix بينهم عن طريق l1_ratio parameter ، كمان فيه طريقة تانية بدل ما اضيف new hyper parameter وهى ال Early stopping انى اوقف ال training لما ال cost function توصل ال minimum مع ال validation set وده لو عملت plot هلاقيها بتوصل لنقطة وبعد كده بتبتدى تعلى تانى والفرق بين cost on training and validation ابتدى يكبر فى الحالة ديه ابتدى اوقف ال trainig .

Post 10

معظم الشابتر كان الكلام بشكل عام ومتجة اكتر ناحية ال Linear Regression وهو انى اتوقع قيمة بتسمى continuous value يعنى فى الاخر مقدرش احصر القيمة ديه فى range معين ، لكن الجزء الاخر من ال Regression هو انى احاول اتوقع قيمة من مجموعة من القيم ، زى انى مثلا اتوقع هل الى فى الصور ده قطة ولا لا ، هنا ان يهمنى فقط انى الصورة قطة ولا لا(معنى كده اى صورة كانت اى حاجة تانية هتمثل بالنسبة ليا class واحد) وال class التانى هو القطة ، وده الى احنا بنسمية Binary Classification ، لكن ازاى انا هقدر اعمل map من القيم الى بيخرجها الموديل ل قيم فى الاخر عبارة عن 0 او 1 او قيمة بينهم وبناء على threshold معين زى انى لو طلع النسبة فوق 50% مثلا هقول انى ديه صورة قطة ولو اقل يبقا لا ، فهنا زى ال prediction الى كان بيحصل فى ال linear regression بظبط بس هخد ال output منه وادية ل function اخرى اسمها Sigmoid function وظيفتها انها تعمل map لل output ده يكون من 0 ل 1 ، بعد ما عملت ال prediction الى انا عايزه عن طريق ال Sigmoid function هبتدى احتاج حاجة تانية هو انى اعمل training يكون الهدف منه هو توقع كبير لل positive class لما تكون الصورة قطة يعنى اكبر من 50% ويكون توقع قليل لما يكون العكس ، وده بيحصل من خلال انى بعمل tweak لل paramters بتاعتى عشان تقدر تساعدنى فى التوقع ده .

Post 11

انا حبيت يكون الكلام عن ال cost function الخاصة بال Logistic Regression يكون فى بوست لوحده عشان نوضح فيه ايه الى بيحصل لو كان ال target ب 0 او 1 بناء على ال function الى معانا ، خلينا نطبقها لما يكون ال target بتاعنا الى هو y=1 ونشوف هنوصل لايه هنلاقى انى ال term الاخير بقا كله ب 0 لان (1-y) هتساوى 0 بما انا y=1 وبكده ده يوضح انى سواء كانت ال y=1 or 0 نص المعادلة بيطير ، طيب جميل دلوقت ال y=1 المفروض انى ده كده positive class ومعنى كده انى المفروض يكون ال output الخاص بال instance ده اكبر من ال threshold عشان نقدر نقول ده positiv ، فانا كده محتاج اشوف ال log(p_i) الى هو ال prediction بتاعى مضروب في ال negative الى بره وده مهم جدا جدا لان ال mapping من 0 ل 1 وبعد كده تاخد ال log هيدى negative value لذلك انا باخد negative كمان طيب معنى كده كل ما كان ال prediction بتاعى بيقرب من 1 ال error هيكون اقل ما يمكن وكل اما بيقرب ناحية ال 0 هيكون اكبر ما يمكن وده الى احنا عايزنه لان ال target فى ال instance ده هو 1 ، زى ما هو واضح فى المثال الى فى الصورة ، والعكس هيحصل لما تكون ال y=0 . بعد كده طبعا هنبتدى اننا نجيب التفاضل بتاع ال cost function عشان نعمل updates لل Weights .

```
In [38]: test = np.linspace(.01,1,10)
         print(test)
         test = [-np.log(i) for i in test]
         test

         [0.01 0.12 0.23 0.34 0.45 0.56 0.67 0.78 0.89 1.  ]

Out[38]: [4.605170185988091,
          2.120263536200091,
          1.4696759700589417,
          1.0788096613719298,
          0.7985076962177716,
          0.579818495252942,
          0.40047756659712525,
          0.2484613592984996,
          0.11653381625595151,
          -0.0]
```

Equation 4-17. Logistic Regression cost function (log loss)

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

Post 12

معظم ال classification problem بيكون multinomial يعنى بيكون مش فقط class-2 لا بيبقا اكتر من كده ممكن مثلا اتوقع رقم من 0 ل 9 فهنا عندى classes-10 وهكذا وده يختلف عن ال multi output الى فيه بتوقع اكتر من object مثلا زى الفيس بوك لما بيعمل tag لصحابك الى معاك فى الصورة ، ال multiclassifcation التعامل معاه ممكن يكون برضه من خلال ال Binary model ولكن هحتاج models كتير جدا عشان افرق مثلا 0 عن 1 و 0 عن 2 لحد 9 وبعد كده ال 1 عن ال 2 وبرضه نفس الكلام فهلاقى عندنا عدد كبير جدا من ال models وفى نفس الوقت كل ما ال Number of classes بيزيد ال models هتزيد وكل ما يكون عندى instance جديد لازم امررة على كل ال models ديه ، لكن ده مش practical ولذلك بنتعامل مع ال multi classification من خلال function اسمها Softmax وفيها انا الأول بشوف ال score بتاع كل class بالنسبة ل x instance اد ايه بعدين بعمل normalization عشان اوزع ال probabilty وتكون بين ال 0 و 1 من خلال ال softmax ، يعنى مثلا لو عندى 150 مثال فى الداتا وكل واحد معاه 2 features هنا انا هحتاج weights بس هتكون خاصة بكل class على حد يعنى بدل ما فى العادى كان عندى 2 * 1 عشان عندى 2 features لا هيبقا عندى بقا 2 * matrix 2 10 ليه عشان 2 features * 10-classes ومن خلال ال prediction fucntion العادية بتاعت ال linear regression هقدر اجيب ال score الخاصة بكل class بالنسبة لكل instance فى الداتا بمعنى

X = 150 * 2
Weights = 2*10
Prediction = X * weights = (150 * 2) . (2*10) = 150 * 10 which means 10-score for each instance.

بعد كده بقا ببتدى ادى الكلام ده لل softmax بالمعادلة الى فى الصورة عشان يكون فى الاخر مجموع ال scores بالنسبه لكل instance يكون 1 عن طريق انى بقسم score of each class بمجموع ال scores بتاعت ال instance ، ونفس الكلام هروح اشوف ال cross entropy function وبعد كده ال partial derivative بتاعها عشان اوصل فى الاخر انى اعمل updates لل weights.

Equation 4-19. Softmax score for class k

$$s_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta}^{(k)}$$

Equation 4-20. Softmax function

$$\hat{p}_k = \sigma(\mathbf{s}(\mathbf{x}))_k = \frac{\exp\left(s_k(\mathbf{x})\right)}{\sum_{j=1}^{K} \exp\left(s_j(\mathbf{x})\right)}$$

- K is the number of classes.
- $\mathbf{s}(\mathbf{x})$ is a vector containing the scores of each class for the instance $\mathbf{x}$.
- $\sigma(\mathbf{s}(\mathbf{x}))_k$ is the estimated probability that the instance $\mathbf{x}$ belongs to class k given the scores of each class for that instance.