

Learn to Build Loss Function for Linear Regression & Logistic Regression from the Ground Up

Nhóm AIO_TimeSeries

Ngày 18 tháng 11 năm 2025

Toàn bộ nội dung Logistic Regression có thể thịnh hành 3 phần lớn, mỗi phần giải quyết một vấn đề riêng biệt nhau - từ trắc nghiệm phân loại, đến cách xóa dữ liệu không hợp lý và xác định độ tin cậy của mô hình. Trong bài học này, chúng ta sẽ tập trung vào cách xây dựng hàm mất mát cho cả hai mô hình: Linear Regression và Logistic Regression.

Phần 1: T Linear Regression và Logistic Regression

Giai quyết vấn đề phân loại: từ Linear Regression + MSE không phải là phù hợp cho Classification. Trong đó có Sigmoid, Bernoulli Distribution, Likelihood và Log-Likelihood có cách xử lý dữ liệu. Tuy nhiên, trong bài học này, chúng ta sẽ tập trung vào cách tính toán Binary Cross-Entropy và Negative Log-Likelihood mà không cần đến gradient descent.

Phần 2: Mạng Logistic Regression khi áp dụng VectorMatrix

Phát triển một toàn bộ công thức để áp dụng vector matrix:

- chuyển đổi dữ liệu thành ma trận X ,
- viết Loss, Gradient và backpropagation,
- giải thích tách rời dữ liệu thành hai lớp.

Phân tích giúp đỡ cho việc tối ưu hóa bằng cách sử dụng gradient descent.

Phần 3: Tí sao BCE là hessian Convex - Phản hồi bao gồm Hessian Matrix

Đúng kinh nghiệm Second Derivative mạng tách biệt bin sang nhiều bin. Giai phương trình tối ưu hóa:

- hiểu Hessian là gì và $f(x, y)$,
- áp dụng Hessian cho BCE gradient,
- phản hồi cu trực $H = X^T D X$,
- chứng minh Hessian luôn Positive Semi-Definite với $D = \text{diag}(h(1 - h))$.

Tuy nhiên, chứng minh BCE là một hessian convex trong khung gian tham số - lý do chính là khin Logistic Regression có thể có nhiều điểm cực tiểu.

Phần 1: T Linear Regression và Logistic Regression

Thay đổi cách giải quyết bài toán phân loại, **mc tiỏu** là một mô hình lớn nhất có thể giải quyết bài toán phân loại từ đầu, bao gồm cả Linear Regression và cả Logistic Regression.

1.1 Vn vi Linear Regression

Trong bji toon d oon giò c phiu vi mt bin u vio x vi mt bin u ra y , y tng ca Linear Regression lị tóm mt hım d oon dng ng thng $f(x)$ sao cho vi mi giò tr x_i mi, ta cú th d oon c y_i . óy lị mt gi thuyt rt n gin nhng hot ng hiu qu nh vio kh nng nm bt xu hng bc nht ca d liu.

K c khi d liu thc t dao ng mn, khú khp, xu hng chung (global trend) vn thng cú th c mū t bng mt ng thng. iu nịy phn ònh ũng bn cht toon hc: ng thng chòn lị bc u tiển trong khai trin Taylor, tc lị mc n gin nht mū t quan h gia còc bin.

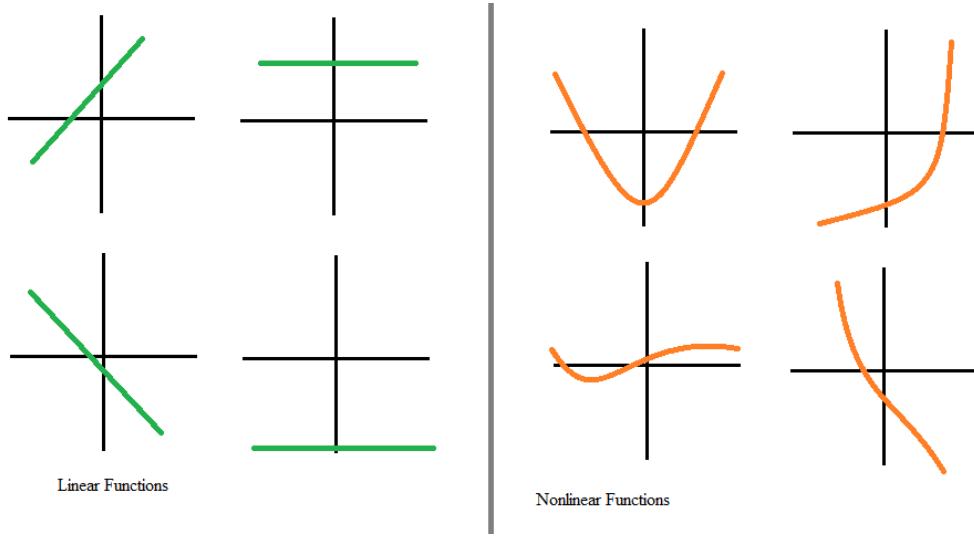


Figure 1: linear function vs nonlinear function

Linear Regression ch cn hai tham s mū t d liu: intercept θ_0 iu chnh cao ca ng thng vi slope θ_1 mū t dc tc tng gim ca d liu. Sai s ϵ_i phn ònh mc lch gia d oon vi giò tr thc t; nu toin b $\epsilon_i = 0$, tt c còc im d liu s nm trồn ũng ng thng. Chòn vơ n gin nh vy, Linear Regression cú th d dìng c ti u bng o hım thüng qua còc hım s bc hai nh Square Loss.

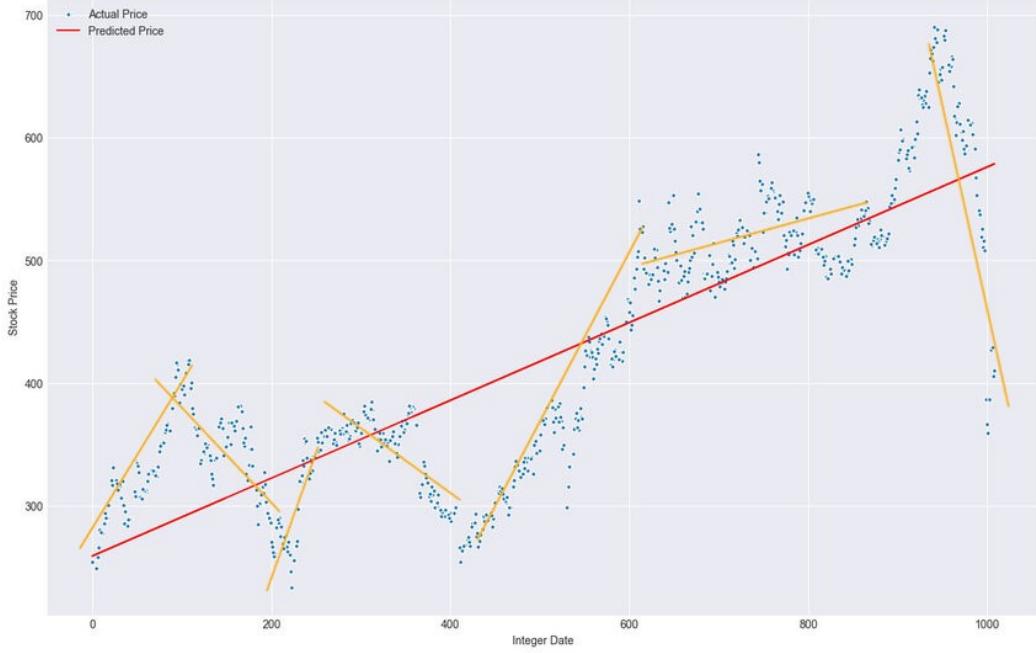


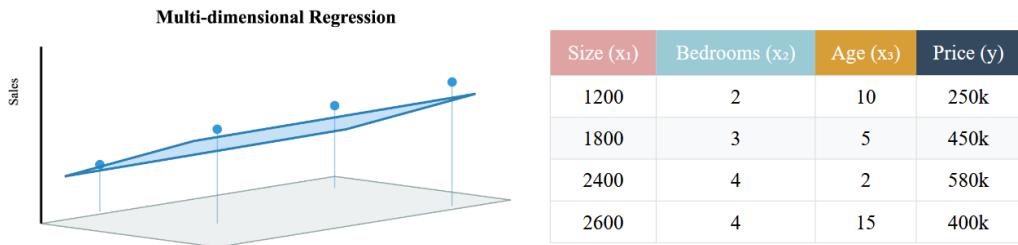
Figure 2: Linear Regression

Mc dứ mū hñh tuyn tñnh cù kh nng mñ t xu hñg mnñ nht trong d liu, bn cht ca Linear Regression cng mang theo mt gi nh quan trng: u ra ca mñ hñh lị mt giò tr the trong khong $(-\infty, +\infty)$. Do ú, khi d liu thc t cù bn cht khñng tuyn tñnh, cù gii hn t nhöñ (nh xòc sut ch nm trong $[0, 1]$), hoc khi bñi toòn lị phón loi nh phón 0/1, Linear Regression s to ra còc d oòn sai lch hoc thm chờ vñ ngha. Khñng ch vy, mñ hñh cùn ngm gi nh rng sai s ca d liu tuón theo phón phi Gaussian iu hoïn toïn khñng phú hp vi bñi toòn phón loi vn cù bn cht Bernoulli.

1.2 Còch Linear Regression chn hñm Loss

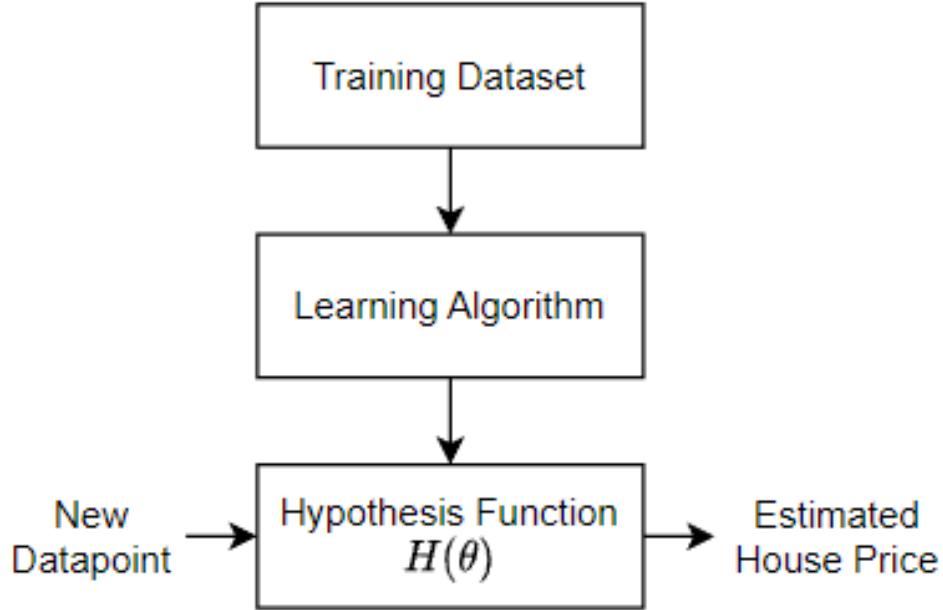
hiu rủ vic xóy dng hñm Loss cho Linear Regression, mñnh s dng vò d d oòn giò nhí vi ba c trng x_1, x_2, x_3 vị mt nhöñ y . Ging nh còch Andrew Ng mñ t trong phn thit k that toòn hc, Supervised Learning luññ i theo workflow: **Training Dataset Learning Algorithm Hypothesis Function**. Hypothesis $H(\theta)$ chñnh lị hñm d oòn mñ hñh hc c, vị nhim v ca Learning Algorithm lị tñm ra b tham s θ sao cho $H(\theta)$ d oòn gn ũng nht giò tr y thc t ca tng mu.

$$\text{Model: Price} = \theta_0 + \theta_1(\text{Size}) + \theta_2(\text{Bedrooms}) + \theta_3(\text{Age})$$



n gin húa, ta gom còc tham s thñnh vector $\theta = [\theta_0, \theta_1, \theta_2, \theta_3]^T$ vị mi im d liu thñnh vector $X_i = [1, x_1, x_2, x_3]^T$. Khi ú hñm d oòn ca Linear Regression tr thñnh

$$\hat{y}_i = h_{\theta}(X_i) = \theta^T X_i.$$



Mc tiőu ca hc mòy lì iu chnh dn còc giò tr trong θ sao cho d oòn \hat{y} ngiy cing gn giò tr tht y . iu niy dn chững ta n vic phi xóy dng mt h m o sai s hay h m Loss. òch t nhi nht o mc t ca mt d oòn lì xem mc ch nh lch $\hat{y} - y$ v i b nh phng n  trонh óm dng trit ti u nhau. Gom tojn b d liu li, ta thu c h m Loss quen thuc:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (\theta^T X_i - y_i)^2 \text{ hay } \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y_i)^2.$$

H m Loss niy cù dng bc hai theo θ , ngha lì mt cong ca n  lu n cù h nh parabol v i 1 dim cc tiu (global minimum), óy lì h m Loss Convex (Li). iu niy rt quan trng v i n  bo rng bt k phng phòp ti u njo cù t nh i xung dn cng s lu n hi t v  ũng nghim ti u. V i ti thiу húa t hay h m Loss niy, mognh s s dng Gradient Descent t m nghim ti u.

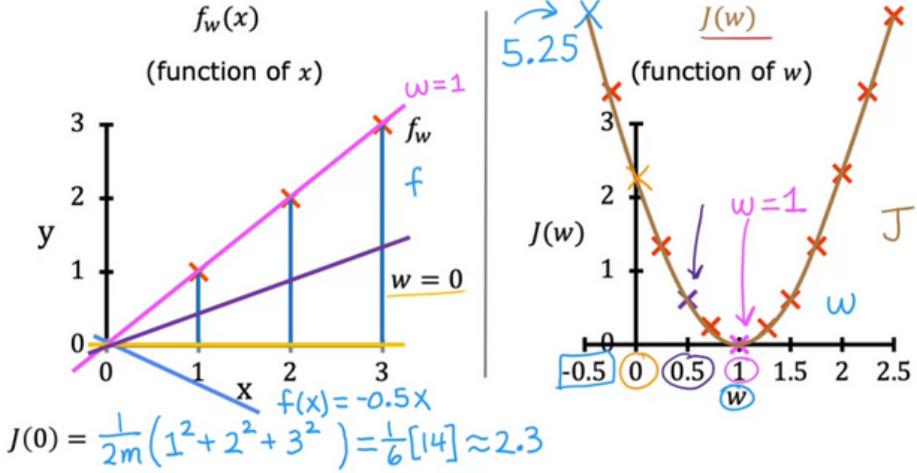


Figure 3: Minh họa hịnh Loss ca Linear Regression trong 2D ($w \approx \theta$): Vì $f_{w,b}(x) = wx + b$ có $b = 0$, Khi mogn th mi giờ tr w t -0.5 n 2.5 cho $f(x)$, mogn s thy hịnh Loss ca Linear Regression s lị mt hinh parabol.

Tuy nhiên, khi chuyen sang bii toòn Phón loi (Classification) vi Logistic Regression, mogn khung th s dung hịnh MSE niy na. Lý do lị khi kt hp vi hịnh Sigmoid ca Logistic Regression, hịnh Loss MSE tng th s có nhieu Local Minimum, óy lị 1 hịnh Loss Non-convex (Khung Li).

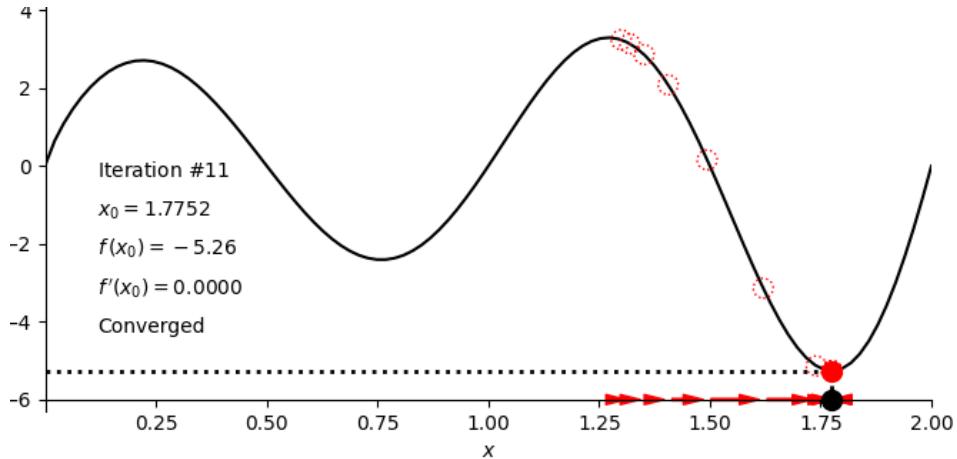


Figure 4: Minh họa hịnh Loss ca Linear Regression trong 2D

T óy, ta óp dung Gradient Descent tóm b tham s ti u. Nu bn cha quen vi ti u húa hoc quón toòn, có th hiu n gìn rong Gradient Descent ging nh ang ng trồn mt qu i (hịnh Loss) vì luñ nhon xem hng i xung dc nht lị hng njo. Mi bc cp nht θ chørnh lị mt bc chón nh i v phòa thp hn, vì vờ hịnh Loss lị hinh parabol, con ng i xung niy s luñ dn ti òy duy nh ca nút tc nghim ti u ca mū hinh.

$$\theta_j := \theta_j - \alpha \frac{\partial J}{\partial \theta_j}, \quad \frac{\partial J}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (\theta^T X_i - y_i) X_{ij}.$$

Vit li tojn b di dung vector hoò cho gn hn, ta có:

$$J(\theta) = \frac{1}{2m} (X\theta - Y)^T (X\theta - Y), \quad \nabla_{\theta} J = \frac{1}{m} X^T (X\theta - Y).$$

Nu ta khũng mun lp Gradient Descent mị mun nhỵ thng n nghim ti u thợ Linear Regression cùn cù nghim úng c suy ra t **NormalEquation**:

$$X^T X \theta = X^T Y \implies \boxed{\theta = (X^T X)^{-1} X^T Y}.$$

1.3 Gii thiу Logistic Regression

Trong phn trc, mñnh õ cung xóy dng tojn b Linear Regression t trc giòc, n thit k hım Loss, còch ti u bng Gradient Descent vñ nghim úng dng ma trn. Khi nhñn li workflow ú, ta thy Linear Regression vn hñnh rt tt cho còc biji toòn d oòn giò tr lññ tc, nhng li khũng phít hp khi biji toòn yñu cu d oòn theo xòc sut hoc phón loi nh phón (0 vñ 1). óy chònh lị im bt u ta phòt trin Logistic Regression.

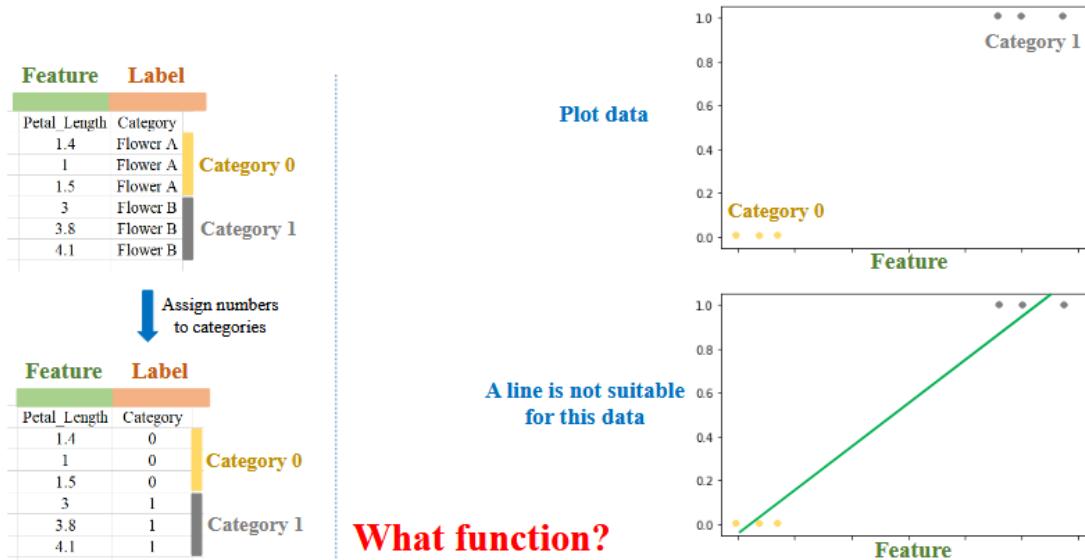


Figure 5: Nhc im ca Linear Regression vi giò tr khũng tuyn tònh

Trc khi i vñ chi tit, mñnh s tham kho còch Andrew Ng trñnh bijy quò trñnh tòi thit k mt mñ hñnh phón loi t Linear Regression, trong ú ũng gii thòch rng chøa khúa nm vic chn hım ònh x u ra sao cho phít hp bn cht xòc sut (**linear regression and gradient descent**).

Vi suy ngh nñy, ta vn gi phn lñi ca Linear Regression tc lị vn tònh mt i lng tuyn tònh:

$$z = \theta^T x,$$

nhng thay vñ dñng trc tip z lñm d oòn nh Linear Regression, Logistic Regression s a z i qua mt hım phi tuyn c bit ònh x giò tr thc $(-\infty, +\infty)$ v on $[0, 1]$ ú chònh lị hım Sigmoid:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}.$$

iu nñy khin mñ hñnh tr thñnh mt b d oòn xòc sut:

$$\hat{y} = P(y = 1 | x).$$

Trc giòc ca Sigmoid rt n gin: khi z cng ln thợ xòc sut cng tin gn 1, khi z cng óm thợ xòc sut tin gn 0, cùn vñng gn $z = 0$ lị vñng khũng chc chn, ni mñ hñnh phn ng nhỵ nht. Nh c tònh cong hñnh ch S nñy, Logistic Regression cù th mñ hñnh húa ranh gii phón loi tt hn Linear Regression.

Tóm lại, Logistic Regression chòn lị:

Linear Regression + hịm Sigmoid.

Phn tip theo s tp trung tr li cóu hi: Nu u ra gi ō lị xòc sut, vy ta cn xóy dng hịm Loss mi nh th njo vic ti u húa vn cù ý ngha?

1.4 Xóy dng hịm Loss cho Logistic Regression t Suy Lun

1.4.1 Hịnh tronh tóm hịm Loss: th nghim vơ sao MSE khũng hp cho Logistic Regressions

Trc khi la chn mt hịm Loss mi cho Logistic Regression, mnh mun cho bn thy rǔ rịng hn vơ sao **Mean Squared Error (MSE)**vn hot ng rt tt trong Linear Regressionli **khũng phứ hp** khi kt hp vi hịm Sigmoid. Thay vơ ch núi rng MSE lịm Loss tr nổn non-convex (nhiu im cc tiu, khú hi t), chđng ta s kim chng iu niy t ba gúc nhơn: hịnh vi ca nú khi thc nghim, phn ng ca Loss theo tng im d liu, vị cui cting lị o hịm.

(1) Quan sòt t th nghim thc

Hôy xem th iu gợ xy ra khi ta hun luyn Logistic Regression nhng vn dúng Loss lị MSE. hnh di, mū hñnh ō c gng d oòn phón loi 0 vị 1, nhng ta nhanh chung thy rng hịm Loss gn nh dng li rt sm vị khũng gim c na. Accuracy ūi khi t mc tm c, nhng mū hñnh khũng hc sóu hn, vị sai s vn cao dứ chy rt lóu.

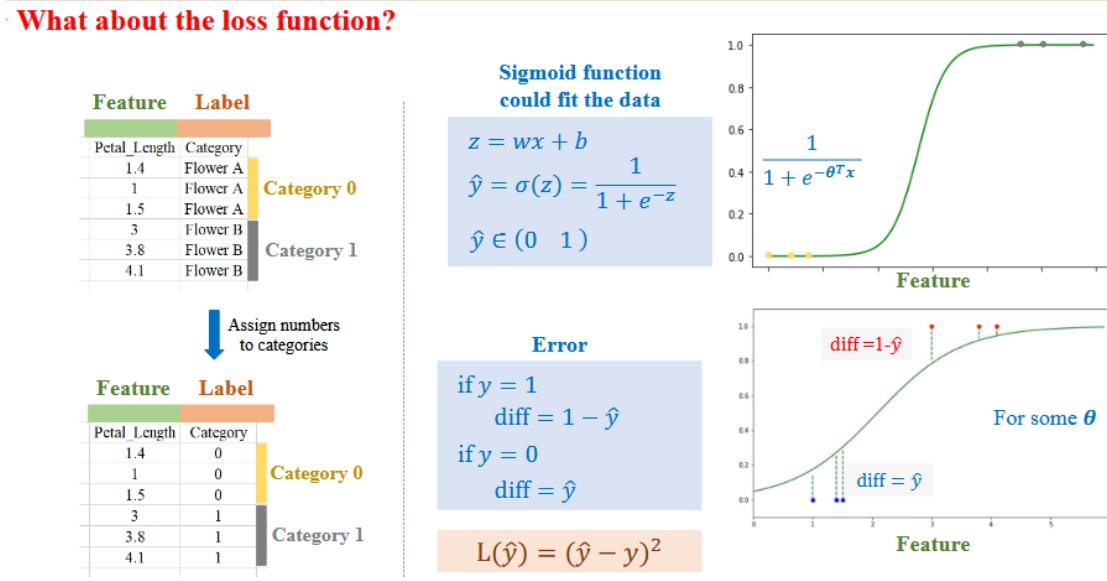


Figure 6: Hun luyn Logistic Regression vi MSE: Loss dng quò sm, mū hñnh khũng hc tip

iu rǔ rịng óy lị mū hñnh ri vio vúng mt phng ca Lossni gradient gn nh bng 0nổn khũng th tip tc cp nht tham s. óy lị hu qu trc tip ca vic Sigmoid búp ŷ v gn 0 hoc 1, lịm cho $(\hat{y} - y)^2$ tr nổn gn nh phng khi o hịm nh.

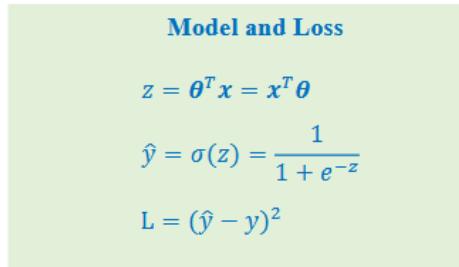
(2) Quan sòt sai s theo tng im d liu

hñnh tip theo ta thy còch MSE tng tóc vi tng im d oòn. Vi mt θ nht nh, Sigmoid cù vúng rt phng hai u (on gn 0 vị gn 1). Khi tòanh MSE, vúng phng niy lịm cho sai s thay i rt ờt ngay c khi mū hñnh d oòn sai nghiòm trng. iu ú ngha lị mū hñnh khũng c pht mnh hc tip.

Logistic Regression-MSE

❖ Result

Done?



$$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y})$$

Feature	Label
Petal_Length	Category
1.4	0
1	0
1.5	0
3	1
3.8	1
4.1	1

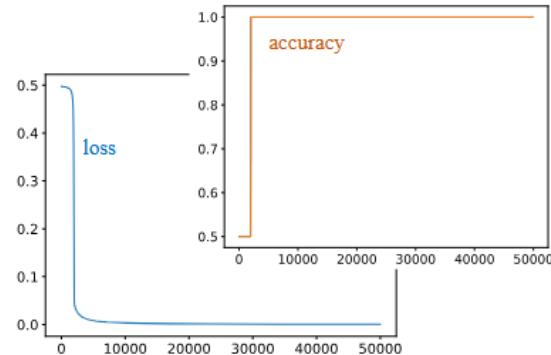


Figure 7: Sigmoid lịm mt nhỵ ca MSE ti hai víng 0 vị 1, mū hñh khñg c pht mnh

Khi d liu lị phón loi, ta cn mt hịm Loss phn ng mnh khi d oòn sai, c bit khi mū hñh sai nhng li t tin. Nhng MSE hojn tojn khñg lịm iu niy.

(3) Phón tòch o hịm ca MSE trong Logistic Regression

óy lị phn quan trng nht thy vn nm óu. Ví Logistic Regression, ta có:

$$z = \boldsymbol{\theta}^T \mathbf{x}, \quad \hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad L = (\hat{y} - y)^2.$$

Ly o hịm theo tng tham s θ_i :

$$\frac{\partial L}{\partial \theta_i} = 2(\hat{y} - y) \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta_i} = 2(\hat{y} - y) \hat{y}(1 - \hat{y}) x_i.$$

iu quan trng nm nhón t:

$$\hat{y}(1 - \hat{y})$$

óy chòn lị o hịm ca Sigmoid, ví nú luñ rt nh khi \hat{y} gn 0 hoc gn 1. Khi nhón vịo cñng thc o hịm tng, gradient tr nññ vñ cúng nhgóy ra hin tng:

$$\text{gradient} \approx 0 \Rightarrow \text{hc rt chm hoc khñg hc c.}$$

Hñh di óy mū t y còc bc o hịm vị thịnh phn gradient góy ra vn .

Model and Loss	Derivative
$z = \boldsymbol{\theta}^T \mathbf{x} = \mathbf{x}^T \boldsymbol{\theta}$	$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta_i}$
$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$	$\frac{\partial \hat{y}}{\partial z} = \hat{y}(1 - \hat{y})$
$L = (\hat{y} - y)^2$	$\frac{\partial L}{\partial \hat{y}} = 2(\hat{y} - y)$
	$\frac{\partial z}{\partial \theta_i} = x_i$

$$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y})$$

Figure 8: o hím ca MSE trong Logistic Regression: gradient b nhón vi $\hat{y}(1 - \hat{y})$ khin cp nht gn bng 0

Kt lun t ba quan sòt trồn:

- Khi Sigmoid búp u ra v 0 hoc 1, o hím Sigmoid $\hat{y}(1 - \hat{y})$ gn bng 0.
- Khi nhón vi $(\hat{y} - y)$ trong cung thc ca MSE, gradient cing b lịm nh thóm.
- Kt qu: Loss MSE tr nổn phng, khú ti u, vị mū hñh khñng th hi t tt.
- Hu qu nghiòm trng nht: **toịn b hím Loss tr thịnh non-convex** cú nhieu cc tiu vị khú hi t..

Chờnh vơ vy, MSE khñng phi lị hím Loss phí hp cho Logistic Regression. **Vy mt hím Loss ũng cho phón loi cn cú c im gó?**. tr li, ta tr li bn cht ca Classification:

- Nu mū hñh d oòn ũng vị t tin, Loss phi rt nh.
- Nu mū hñh d oòn sai nhng li t tin, Loss phi tng tht nhanh.
- Hím phi cú dng trn (ie. hím cú th o hím vñ tn, 1 ng cong ko b gõy), liñn tc, d ti u bng Gradient Descent.

Vi cùc iu kin trồn, ta th quan sòt cùc dng hím quen thuc vị xem phn ng ca chñng khi u vño tin gn 0 hoc 1.

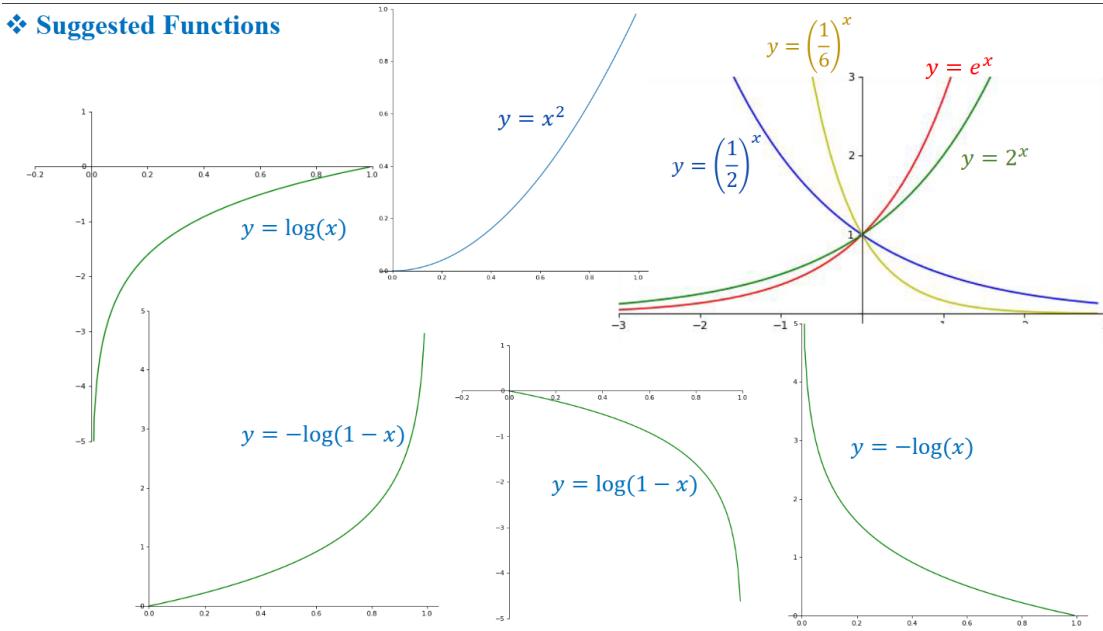


Figure 9: Quan sòt còc dng hím: ch còc hím logarithm tng rt mnh khi \hat{y} tin v 0. T hñnh nïy, ta nhn thy rng hai ng viøn t nhiøn cho bïi toòn phón loi lì:

1. $-\log(\hat{y})$ khi $y = 1$
2. $-\log(1 - \hat{y})$ khi $y = 0$

C hai u bñng n khi mñ hñnh sai mt còch t tinñng th chñng ta cn, vï trïi ngc hoïn toïn vi MSE vn pht rt nh.

Thay vñ vit theo kiu **if/else**, ta ch cn kt hp hai hím bng biu thc gn ging:

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}).$$

óy chñnh lì **Binary Cross-Entropy (BCE)** hím Loss chun cho Logistic Regression. Khñng ch phú hp trc giòc, BCE cùn khp hoïn ho vi bn cht xòc sut khi mñ hñnh hoò nhñn y nh mt bin Bernoulli, vï cung thc trñn chñnh lì negative log-likelihood.

Khi kt hp vi Logistic Regression:

$$z = \theta^T x, \quad \hat{y} = \sigma(z),$$

gradient tr nññ cc k gn:

$$\frac{\partial L}{\partial \theta_i} = x_i(\hat{y} - y),$$

tng t Linear Regression nhng mang y òng ngha xòc sut.

Construct loss

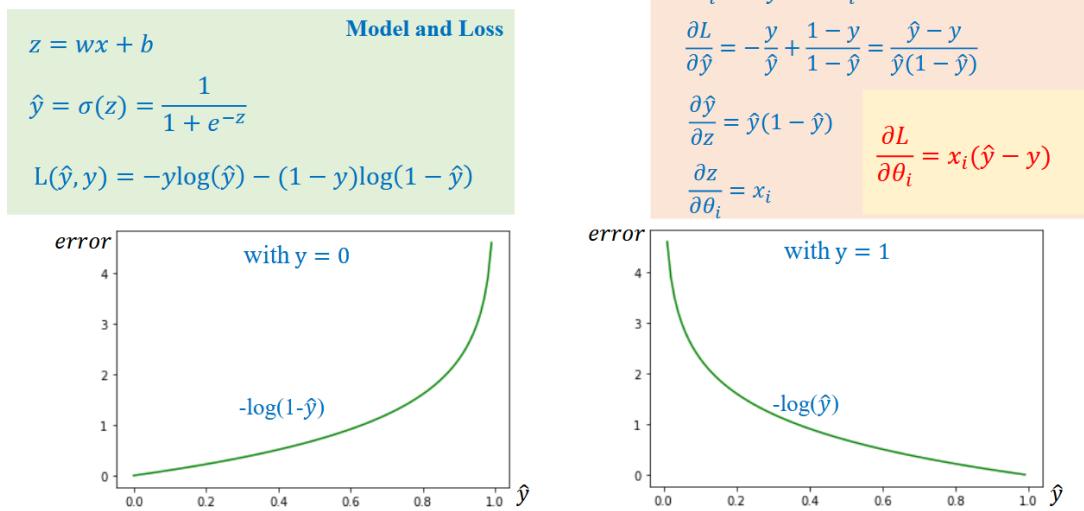


Figure 10: Pipeline Logistic Regression + BCE: n gin vị cht ch

T trc giòc n cũng thc, BCE òp ng tt c nhu cu ca bjì toòn phón loi: pht mnh khi sai, n nh khi ũng, o hịm n gin vị hc rt nhanh. iu quan trng nht lị khi quan sòt mt Loss ca BCE, ta thy nú cú dng bòt ũp khũng gian tham smt du hiu mnh m ca **Convexity**.

Convexity chòn lị chøa khúa giüp Logistic Regression hi t n nh vị tròn by local minima. Vø vy, trc khi chng minh BCE convex ví so sòn BCE vi MSE, ta cn xóy dng trc giòc v cong ca mt hịm thñng qua o hịm bc hai óy chòn lị ch ca phn tip theo.

1.4.2 Convexity lị go? Trc giòc ca o hịm bc hai

Trc khi chng minh hịm Loss ca Logistic Regression lị convex, monh mun cung bn quay li trc giòc toòn hc nn tng; **o hịm bc hai cho ta bit Rate of change of "the rate of change"** hay nui còch khòc, nú cho ta bit cong (curvature) ca mt hịm. Còch hñnh tng vị d hiu nht nm c iu niy lị nhơn vio chuyn ng ca mt chic xe.

Khi ta mū t chuyn ng ca xe theo thi gian t , ta cù ba i lng quen thuc: quõng ng $x(t)$, vn tc $v(t)$ ví gia tc $a(t)$. óy chòn lị vò d trc tip mị bji ging ca DeepLearning.AI õ dñng gii thòch o hịm bc mt ví bc hai. Vn tc lị o hịm bc mt ca quõng ng, $v = \frac{dx}{dt}$, cùn gia tc lị o hịm bc hai (ie. tc thay i ca vn tc, nghe rt hp lị ũng khũng :)), $a = \frac{dv}{dt} = \frac{d^2x}{dt^2}$. Nu vn tc ang tng thø gia tc dng, nu vn tc gim thø gia tc óm, cùn nu vn tc khũng i thø gia tc bng 0. iu niy giüp ta cù trc giòc: o hịm bc hai cho ta bit hịm ang cong lñn hay cong xung.

	x	Distance
	v	Velocity $\frac{dx}{dt}$
	a	Acceleration $\frac{dv}{dt} = \frac{d^2x}{dt^2}$

Figure 11: x , v , a : Khong còch Vn tc Gia tc

Hình di óy mū t rủ hn: th phòa trồn lị quõng ng theo thi gian, mi im c mū t bng 1 ng thng, th di lị vn tc (o hịm bc mt). giao on u, vn tc liõn tc tng nõn gia tc dng; on gia, xe chy u nõn gia tc bng 0; on sau, xe gim tc nõn gia tc óm.

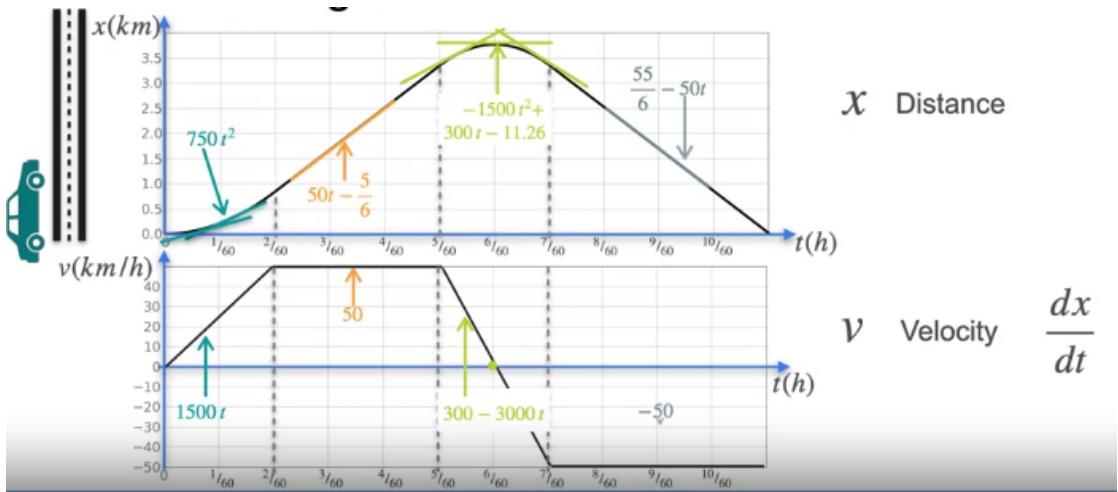


Figure 12: o hịm bc mt: Vn tc thay i theo thi gian

Understanding Second Derivative

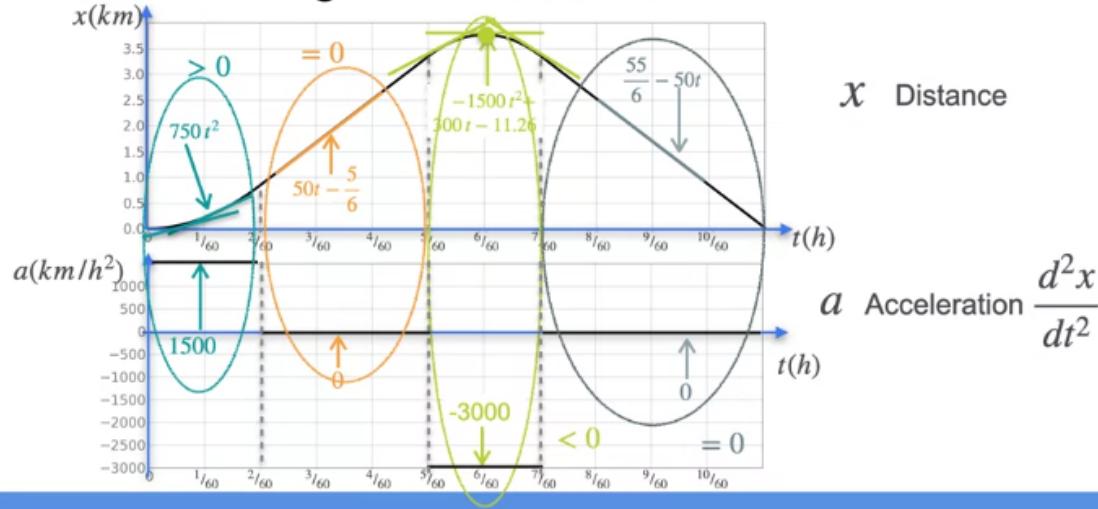


Figure 13: Hình di lị o hịm bc hai: Gia tc tc thay i ca tc thay i (tc thay i ca vn tc)

Khi nhơn vịo th quõng ng, ta cú th phón tòch cong da vịo du ca o hịm bc hai: nu $\frac{d^2x}{dt^2} > 0$, th cong lõn (concave up / convex); nu $\frac{d^2x}{dt^2} < 0$, th cong xung (concave down); cùn nu bng 0 thợ ta cn thõm thñng tin vñ th thng hoc ch ang i ch cong.

Understanding Second Derivative

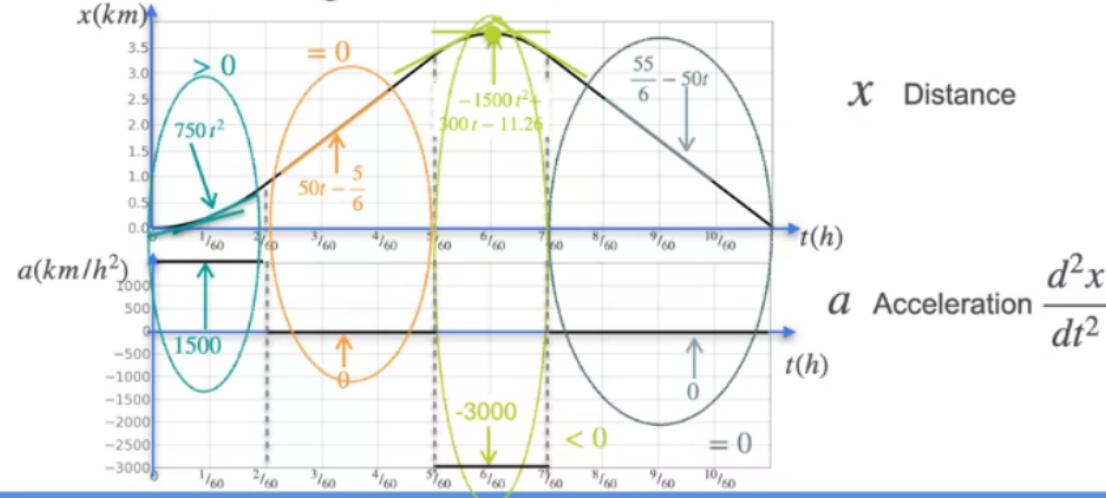
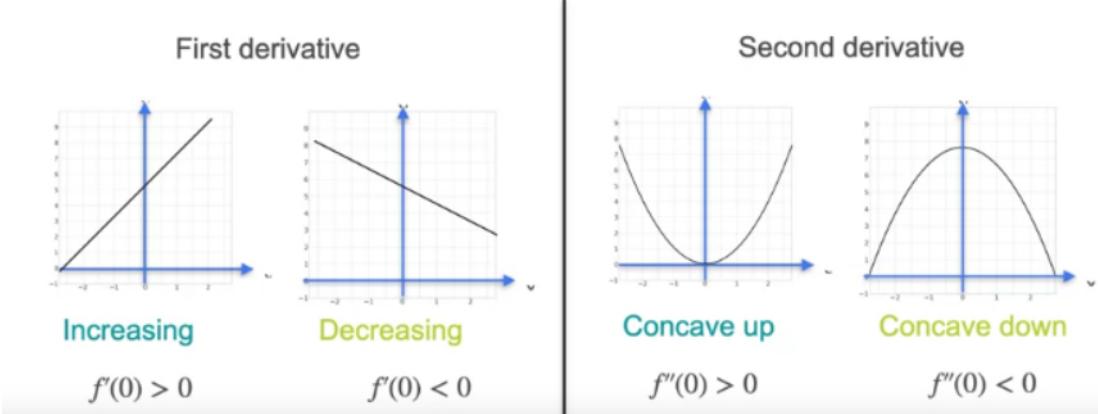


Figure 14: Ý ngha du ca o hịm bc hai vñ cong ca hịm

iu niý liõn kt trc tip vi ti u húa: ti còc im mị o hịm bc mt bng 0, **o hịm bc hai cho ta bit im ú lị cc tiu hay cc i**. Nu o hịm bc hai dng, ú lị cc tiu (th cong lõn). Nu óm, ú lị cc i (th cong xung). Nu bng 0, ta khñng kt lun c. óy chñnh lị lỳ do o hịm bc hai quan trng trong phón tòch convexity.

T gúc nhơn trc giòc: hịm convex lị hịm luñn cong lõn, ngha lị o hịm bc hai ca nú luñn khñng óm. Bn cú th hñnh dung rng nu th luñn cong lõn ging hñnh mt còi bòt ũp ngc, thợ dñr bn ng bt k im niø, hng i xung dc nht cng luñn dn bn v cñng mt öy duy nht. óy lị nn tng ca ti u húa: hịm convex m bo khñng cùnhiu cc tr cc b; Gradient Descent s khñng bao gi b mc kt.

Curvature



The 2nd Derivative tell at how the "rate of change" (1st Derivative) is changing.

Summary

- The **first derivative** gives the **rate of change or slope** of the function.
- The **second derivative** gives information about the **concavity or curvature**.
- For a **quadratic function**, the **second derivative is constant**, which gives rise to a **parabolic shape**.
- Higher-order functions have **changing second derivatives**, resulting in more complex and varied curves, not just parabolas.

Figure 15: Tóm tắt trắc giác о hÌm bÌc mt vÌ bÌc hai

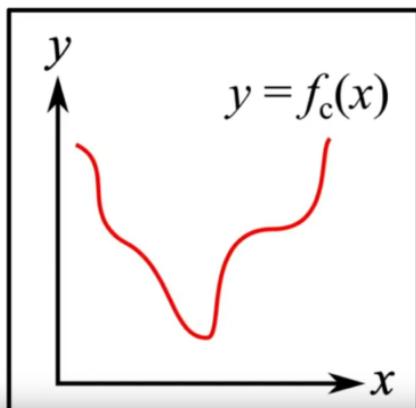
Khi chuyen t vÒ d chic xe sang bÌi toÙn hc mÙy, ta cÙ th nhÓm convexity nh mt s m rng ca nhng gÓ mÙnh va thy: nu o hÌm bÌc hai ca mt hÌm s luÙn khÙng óm trØn toÙn min, thØ hÌm ú lÌ convex. i vi hÌm nHiu bin, khÙi nim nÌy c tng quÙt bng **ma trÙn Hessian**. Mt hÌm nHiu bin c gi lÌ convex nu Hessian luÙn lÌ ma trÙn positive semi-definite. iu nÌy ng ngha mi eigenvalue ca Hessian u khÙng óm, chØnh xÙc nh logic ca o hÌm bÌc hai trong trng hp mt bin.

VØ vy, khi ta nÙi rng mt hÌm Loss lÌ convex, iu ú cÙ ngaha lÌ gradient luÙn dn mÙ hØnh v nghim ti u duy nht. óy s lÌ chØa khÙa gii thØch vØ sao Binary Cross-Entropy ca Logistic Regression lÌ hÌm Loss hc n nh vÌ khÙng gp vn local minima. Phn tip theo s chng minh trc tip iu nÌy bng cÙch phØn tØch o hÌm vÌ Hessian ca BCE.

1.4.3 Chng Minh Convex (lÌ) ca MSE vÌ BCE

Phn trc, ta thy rng Logistic Regression dÙng MSE khÙng nhng hc chm mÙ cÙn d ri vÌo cÙc vÙng gradient bng 0. Trc giÙc ban u ch cho thy MSE khÙng phÙt hp, nhng kt lun mt cÙch chc chn, ta cn nhØn vÌo hØnh dng ca mt Loss. óy lÌ lÙc khÙi nim **Convexity** tr nÙn quan trng: nu Loss lÌ convex (lÌ), Gradient Descent s luÙn hi t v nghim ti u duy nht; nhng nu Loss non-convex, mÙ hØnh cÙ th mc kt cÙc local minima.

NON-CONVEXITY ISSUES



Gradient descent doesn't guarantee to find the minimum point in logistic regression.

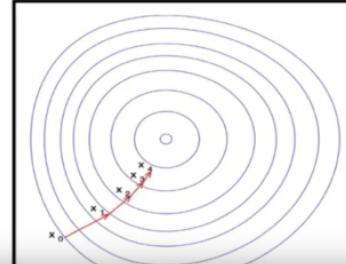


Figure 16: o hím bc hai ca MSE i du Non-Convex

Phn nịy tp trung viø ba im chørnh:

- (1) **MSE khi kt hp vi Sigmoid tr thịnh non-convex**

Hím Loss b bin dng thịnh dng “gn súng”, xut hin nhieu vúng phng vị òy cc b, khin Gradient Descent d mc kt.

- (2) **BCE to ra mt hím Loss convex**

cong ca BCE khi kt hp vi Sigmoid luñn khñng óm, b mt Loss cong mt vị cù duy nht mt cc tiu toin cc.

- (3) **Do ú BCE phứ hp hn MSE cho Logistic Regression**

BCE to khñng gian ti u húa n nh, cù gradient rủ rjing, khñng b bõo hùa nh MSE.

Chng Minh MSE lị Non-Convex trong Logistic Regression

Vì Logistic Regression ta cù:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad L = (\hat{y} - y)^2.$$

Ly o hím bc hai ca Loss theo tng tham s θ_i :

$$\frac{\partial^2 L}{\partial \theta_i^2} = 2x_i^2 \hat{y}(1 - \hat{y}) [-3\hat{y}^2 + 2\hat{y} - y + 2y\hat{y}] .$$

im quan trng nm biu thc ngoc vuñng. Khi \hat{y} chy trong khong (0, 1), giò tr trong ngoc cù th dng hoc óm túy vúng. Do ú o hím bc hai cù th óm **Loss cù vúng cong lồn vị cong xung xen k non-convex**

Mean Squared Error

Model and Loss	Derivative
$z = \theta^T x = x^T \theta$	$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta_i}$
$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$	$\frac{\partial \hat{y}}{\partial z} = \hat{y}(1 - \hat{y})$
$L = (\hat{y} - y)^2$	$\frac{\partial z}{\partial \theta_i} = x_i$
	$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y})$

$$\begin{aligned}\frac{\partial L}{\partial \theta_i} &= 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y}) = 2x_i(-\hat{y}^3 + \hat{y}^2 - y\hat{y} + y\hat{y}^2) \\ \frac{\partial^2 L}{\partial \theta_i^2} &= \frac{\partial}{\partial \theta_i} [2x_i(-\hat{y}^3 + \hat{y}^2 - y\hat{y} + y\hat{y}^2)] \\ &= 2x_i[-3\hat{y}^2x_i\hat{y}(1 - \hat{y}) + 2x_i\hat{y}\hat{y}(1 - \hat{y}) - yx_i\hat{y}(1 - \hat{y}) + 2x_iy\hat{y}\hat{y}(1 - \hat{y})] \\ &= 2x_i^2\hat{y}(1 - \hat{y})[-3\hat{y}^2 + 2\hat{y} - y + 2y\hat{y}]\end{aligned}$$

Figure 17: o him bc hai ca MSE i du Non-Convex.

Mean Squared Error

$$\frac{\partial^2 L}{\partial \theta_i^2} = 2x_i^2\hat{y}(1 - \hat{y})[-3\hat{y}^2 + 2\hat{y} - y + 2y\hat{y}]$$

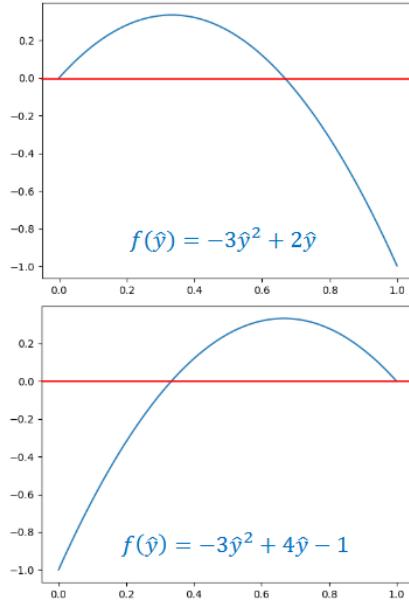
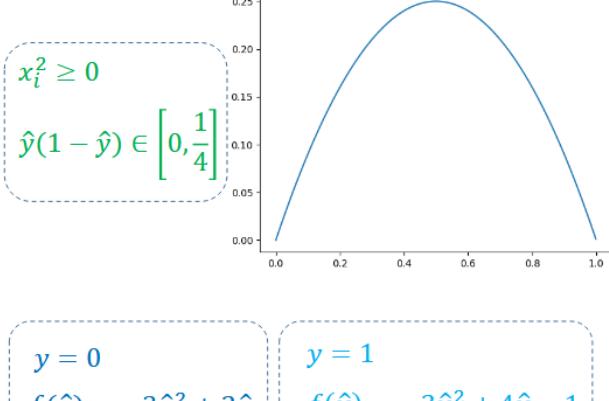


Figure 18: Mt s vting $\frac{\partial^2 L}{\partial \theta_i^2} < 0$ chng minh Loss khung li. Vy Logistic Regression + MSE to ra mt Loss Non-Convex

óy chòn lì lỳ do thc nghim MSE hc rt chm: khi Loss cù nhieu vting cong xung, Gradient Descent thng b kt hoc cp nht rt nh.

Ngc li, vi **Binary Cross-Entropy**:

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}), \quad \hat{y} = \sigma(z).$$

o him bc nht rt n gin:

$$\frac{\partial L}{\partial \theta_i} = x_i(\hat{y} - y).$$

Ly o hím bc hai:

$$\frac{\partial^2 L}{\partial \theta_i^2} = x_i^2 \hat{y}(1 - \hat{y}).$$

Võ:

$$x_i^2 \geq 0, \quad \hat{y}(1 - \hat{y}) \in [0, \frac{1}{4}],$$

nỗn o hím bc hai luñn khñng óm **hím Loss BCE luñn convex**

Binary Cross-entropy

Convex function

$$z = \boldsymbol{\theta}^T \mathbf{x}$$

Model and Loss

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$L = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

simplified version

$$\frac{\partial L}{\partial \theta_i} = x_i(\hat{y} - y)$$

$$\frac{\partial^2 L}{\partial \theta_i^2} = \frac{\partial}{\partial \theta_i} [x_i(\hat{y} - y)] = x_i^2(\hat{y} - \hat{y}^2) \geq 0$$

$$x_i^2 \geq 0 \quad \hat{y} - \hat{y}^2 \in [0, \frac{1}{4}]$$

Derivative
$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta_i}$
$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$
$\frac{\partial \hat{y}}{\partial z} = \hat{y}(1-\hat{y})$
$\frac{\partial z}{\partial \theta_i} = x_i$
$\frac{\partial L}{\partial \theta_i} = x_i(\hat{y} - y)$

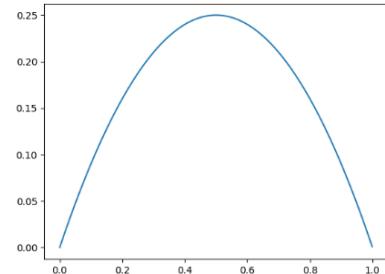


Figure 19: o hím bc hai BCE luñn dng mt Loss convex

iu niý khng nh BCE luñn cù **mt nghim ti u duy nht** vї Gradient Descent s hi t n nh.

1.4.4 So sánh MSE và BCE

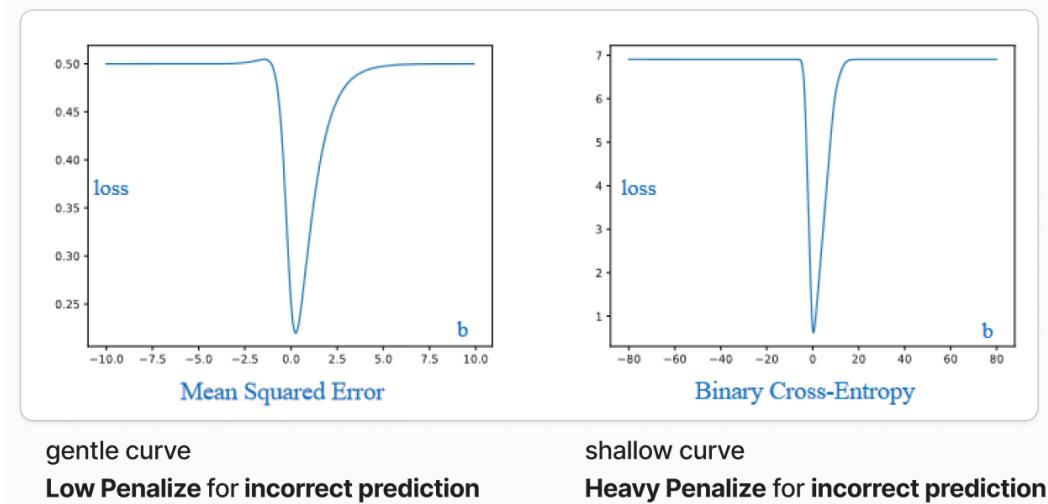


Figure 20: So sánh hım Loss MSE vı BCE

MSE: ng cong nh, di giò tr b chn

Loss ch dao ng trong khong hp ($\approx 0.25 \rightarrow 0.5$), cong thp vı nhieu vıng bng phng khi \hat{y} bō hùa. iu nıy khin gradient nh, mū hınh d ng yồn vı khú iu chnh tham s.

BCE: ng cong sóu, di giò tr m rng

Loss tng rt mnh khi d oòn sai t tin (vờ d $\hat{y} \approx 0$ nhng $y = 1$), to thịnh öy rng vı cong ln. Nh ú gradient rủ rıng, mū hınh hc n nh vı hi t ıng.

Xóy dng Hım Loss cho Logistic Regression t Xòc Sut Thng Kő

2.1 Bt u t vn : Vơ sao Logistic Regression khũng th da vıo MSE?

Trong bji toòn phón loi nh phón, u ra ch cù hai giò tr:

$$y \in \{0, 1\}.$$

Nu dúng MSE nh Linear Regression, mū hınh cù th d oòn giò tr ln hn 1 hoc nh hn 0, iu nıy khũng phứ hp vi y nghia xòc sut. Ngoi ra, MSE kt hp vi Sigmoid to ra b mt Loss non-convex, khin vic ti u húa bng Gradient Descent tr nőn khú khn.

Vờ th, ta cn mt còch tip cn ıng bn cht hn: *xóy dng hım Loss da trönn xòc sut thng kő*. lım c iu ú, ta s bt u t nhng khòi nim nn tng nht.

2.2 Ťn li nn tng: Logarithm vı Natural Logarithm

Trc khi i vıo xòc sut, ta cn hiu lỳ do vơ sao Logarithm (log) vı Natural Logarithm (ln) úng vai trù quan trng trong Logistic Regression.

(a) Logarithm lị go?

Logarithm tr li cóu hi:

$$\log_b(a) = c \quad \text{khi} \quad b^c = a.$$

Nói cách khác, log làm ra **phn s m** cn cù to ra mt s. óy lì lỳ do log c bit hu òch khi ta lìm vic vi d liu cù dng tng gím theo bi s, hoc khi còc xòc sut nh c nhón liён tip vi nhau.

(b) ng dng: Log chun hoò t l a v cứng mt thang o

Xét vò d trc quan trong hinh 21. Hai giò tr:

$$8 = 2^3, \quad \frac{1}{8} = 2^{-3}$$

cách nhau 8 ln so vi 1, nhng trồn trc s thñg thng, khong cách $\frac{1}{8} \leftrightarrow 1$ vị 1 $\leftrightarrow 8$ khñng i xng.

Tuy nhiên, nu đúng log c s 2:

$$\log_2(8) = 3, \quad \log_2(1/8) = -3,$$

tho hai giò tr nịt tr nññ i xng. iu nịt cho thy log bin t l (ratio) thịnh khong cách tuyn tòn. Vic chun húa nịt cc k quan trng khi còc tr s quò nh hoc quò ln khin vic ti u vi gradient khú khn.

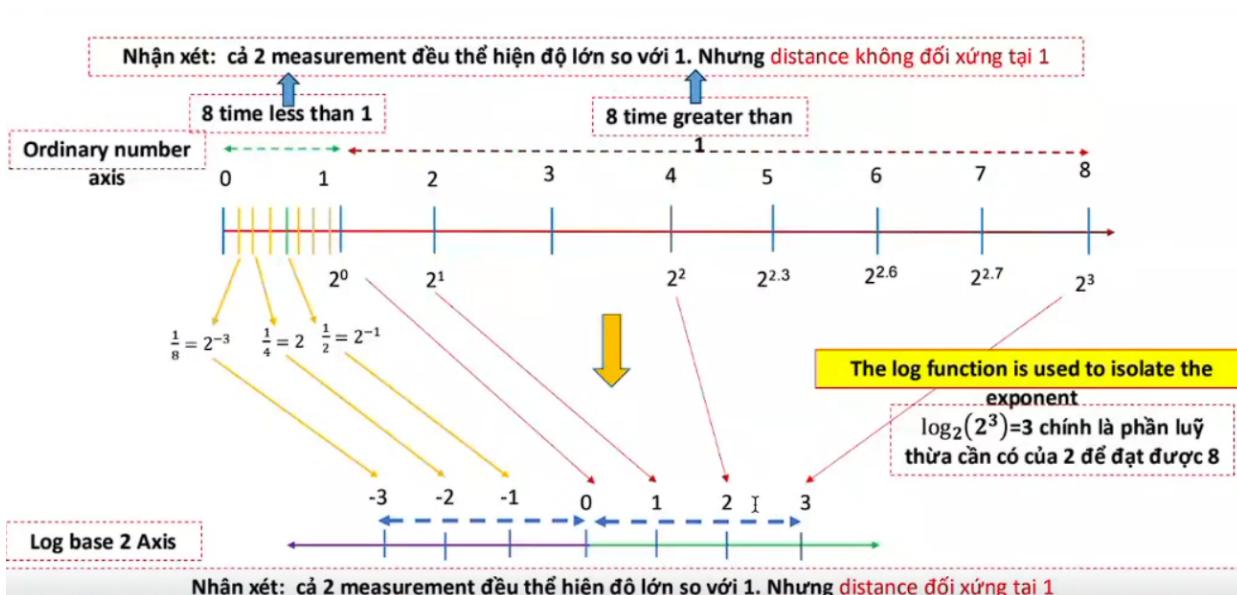


Figure 21: Log bin t l ly tha thịnh khong cách tuyn tòn i xng: mt cũng c cc k quan trng khi x lì xòc sut nh vị d liu tri dji nhieu bc ln.

(c) Võ sao đúng $\ln(x)$ thay võ $\log_{10}(x)$?

Trong hc mò, ta đúng $\ln(x)$ vò:

- ln liён quan trc tip n hím m e^x , vn xut hin liён tc trong xòc sut vị ti u húa.
- o hím ca ln rt n gin:

$$\frac{d}{dx} \ln(x) = \frac{1}{x}.$$

- Khi ti u húa Likelihood, tojn b biu thc s tr nññ gn, tròn phi mang thóm h s $\ln(10)$.

Tóm li: ln giúp cung thc p, d o hím, vị phứ hp vi mi phón phi đúng trong thng kô hin i.

2.3 Probability vi Likelihood (gii thoch lin mch bng vor d ng xu)

hiu Logistic Regression da vio xoc sut nh th no, iu quan trng nht li phón bit **Probability vi Likelihood**.

Probability li xoc sut ta *gi nh trc khi quan st d liu*. Hoy tng tng bn cu mt ng xu cung bng. Trc khi tung, bn cho rng xoc sut ra mt nga li $P(\text{Head}) = 0.5$ vi xoc sut ra mt sp cng li 0.5. Khi ton hong Probability, ta ang hi: Nu mu honh ung nh ta tin (ng xu cung bng), tho kh nng xy ra mt s kin li bao nhiu? óy chon li *probability without testing* hojn tojn da tron gi nh mu honh, khung da vio d liu the t.

Trong khi ú, **Likelihood** xut hin ton hong ngc li: **khi o quan st d liu**, vi mun bit tham s mu honh cu hp ly khung. Vor d: bn tung ng xu 50 ln, nhng ch thy 14 ln ra mt nga. iu niy khin bn nghi ng rng gi nh $P(\text{Head}) = 0.5$ cu th khung ung. Luc niy, ta hi: Vi d liu niy, gio tr no ca p li hp ly nht? óy li *probability after testing*. D liu o c nh, cùn tham s p li th thay i. Likelihood o hp ly ca tng gio tr p trong vic sinh ra d liu quan st.

Tum li:

Probability: mu honh c nh d liu thay i, Likelihood: d liu c nh mu honh thay i.

Vi Logistic Regression dung Likelihood tom ra tham s tt nht.

2.4 Bernoulli Distribution: Mu honh chon xoc cho phón loi nh phón

Trong phón loi nh phón, y ch cu hai gio tr 0 hoc 1. óy chon xoc li nh ngha ca bin Bernoulli. Phón phi Bernoulli c vit:

$$P(Y = 1) = p, \quad P(Y = 0) = 1 - p.$$

gn hn, ta vit di dng hp nht:

$$P(Y = y) = p^y(1 - p)^{1-y}.$$

iu niy phn ònh chon xoc ung ca y :

- nu $y = 1$, biu thc cùn li li p ,
- nu $y = 0$, biu thc cùn li li $1 - p$.

2.5 Mu honh Logistic Regression: Mu honh húa xoc sut $P(y = 1 | x)$

Ý tng ca Logistic Regression n gin:

$$p = P(y = 1 | x)$$

Nhng vo $\theta^T x$ cu th óm hoc dng, ta cn mt him up gio tr v on $[0, 1]$. Sigmoid lum iu niy rt t nhiun:

$$h_\theta(x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = \theta^T x.$$

2.6 Xóy dng Likelihood cho tojn b tp d liu

Trong bi cnh mu honh húa bng Bernoulli, ta gi nh rng mi im d liu $(x^{(i)}, y^{(i)})$ c sinh ra bi mt bin ngu nhiun cù xoc sut:

$$P(y^{(i)} = 1 | x^{(i)}, \theta) = h_\theta(x^{(i)}), \quad P(y^{(i)} = 0 | x^{(i)}, \theta) = 1 - h_\theta(x^{(i)}).$$

Dứng dng hp nht ca phón phi Bernoulli, ta cú th vit xòc sut cho tng mu:

$$P(y^{(i)} | x^{(i)}, \theta) = h_\theta(x^{(i)})^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}.$$

Trc giòc ca biu thc niy:

- Nu mu cú nhõn $y^{(i)} = 1$, biu thc tr thịnh $h_\theta(x^{(i)})$. Mū hñnh c thng nu $h_\theta(x^{(i)})$ ln (xòc sut cao cho lp ũng).
- Nu $y^{(i)} = 0$, biu thc tr thịnh $1 - h_\theta(x^{(i)})$. Mū hñnh c thng nu nú gim xòc sut d oòn dng.
- S m $y^{(i)}$ vñ $1 - y^{(i)}$ úng vai trù nh cñng tc (on/off). Chñng bt ũng biu thc cn dñng cho tng nhõn. Còch vit niy cc k hiu qu vñ cñng mt cñng thc x lÿ c c hai trng hp.

2.7 Vy Likelihood cho toïn b tp d liu lị gợ?

Vñ còc im d liu c gi nh lị c lp (IID independent and identically distributed), xòc sut quan sòt c tp d liu chònh lị:

$$\mathcal{L}(\theta) = \prod_{i=1}^m P(y^{(i)} | x^{(i)}, \theta) = \prod_{i=1}^m h_\theta(x^{(i)})^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}.$$

Ti sao dñng phôp nhón (Product) thay vñ phôp cng ? Cùn trong xòc sut, óy lị im rt nhieu ngi hay nhm khi mi hc xòc sut.

- Vi còc s kin c lp, xòc sut xy ra ng thi bng **tòch** còc xòc sut thịnh phn.
- Mi im d liu u lị 1 tham s θ c lp.
- Vñ th, xem toïn b tp d liu “ng h” tham s θ bao nhieu, ta phi nhón toïn b còc xòc sut li.

óy chònh lị ni xut hin khòi nim **likelihood ca mū hñnh** -- nú lị mc mū hñnh hp lỳ khi nhơn vñ toïn b d liu.

Ün tp li kin thc toòn c bn

Biu thc Likelihood s dng li hai quy tc quen thuc:

- **Product Rule** Nu còc s kin c lp,

$$P(A \cap B) = P(A)P(B).$$

- **Power Rule** Khi vit li p hoc $(1-p)$ tu theo nhõn y , ta dñng quy tc:

$$a^1 = a, \quad a^0 = 1.$$

óy lị lỳ do cñng thc Bernoulli gp c 2 trng hp vñ mt.

Nhơn vịo mt vờ d thc t hiu rủ hn

Gi s ta cn d oòn kh nng mc ung th da trồn ch s PSA (theo vờ d trong tịi liu bn cung cp). Mi bnh nhón cú:

- giò tr PSA (u vịo $x^{(i)}$), - nhõn: **Cancer** (1) hoc **Healthy** (0), - mū hñnh d oòn xòc sut $p^{(i)} = h_{\theta}(x^{(i)})$.
- Vờ d mt on d liu (rút gn):

PSA	y	$h_{\theta}(x)$
3.8	1	0.99
3.4	1	0.97
2.9	1	0.92
2.1	0	0.50
1.2	0	0.13

Likelihood ca mū hñnh lị:

$$\mathcal{L}(\theta) = 0.99^1 \cdot 0.97^1 \cdot 0.92^1 \cdot (1 - 0.50)^1 \cdot (1 - 0.13)^1.$$

Nu nhón y :

$$0.99 \cdot 0.97 \cdot 0.92 \cdot 0.50 \cdot 0.87 \approx 0.386.$$

Trc giòc:

- Mū hñnh cng chñnh xòc còc xòc sut ũng cng cao tòch cng ln.
- Mū hñnh sai t tin mt s hng rt nh kñ tojn b tòch xung gn 0. óy chñnh lị logic ct lñi ca Likelihood.

Ti óy, ta ó xóy dng xong Likelihood. phn tip theo, ta s i mt vi vn ln nht ca Likelihood: *tòch nñiu xòc sut nh khin giò tr nhanh chñng tin v 0*. Vị ú lị lñc **Log-Likelihood** xut hin .

2.8 Log-Likelihood: Khi phôp nhón khñng cùn chu ni d liu thc t

Mt trong nhng vn ln nht ca Likelihood lị giò tr ca nñ **gim rt nhanh** khi nhón nñiu xòc sut nh li vi nhau. Trong thc t, còc mū hñnh phón loi thng d oòn xòc sut nh 0.93, 0.71, 0.85, ... hoc cù nhng im khú oòn vi xòc sut 0.02 hoc 0.15.

Khi nhón 10010,000 giò tr nh vy, ta s thu c nhng con s nh ti mc:

$$10^{-50}, 10^{-200}, 10^{-500} \text{ hoc thm chò nh hn.}$$

Kt qu lị mòy tònh s trìn s (underflow), bin tojn b kt qu v 0. Hoc nge li overflow khi s quò ln. iu nñy khin vic ti u tr nññ bt kh thi.

Vy lñm th njo bin mt tòch siôu nh thịnh th cù th tònh toòn, o hñm vñ ti u c? Cóu tr li chñnh lị: **dñng Logarithm**.

Ti sao log giüp gii quytn vn ? Hiu bng trc giòc trc

Ta cù quy tc cc kñ quan trng ca log

$$\ln(a \cdot b \cdot c) = \ln(a) + \ln(b) + \ln(c).$$

Mt phôp nhón hñng trm s nh

$$0.99 \cdot 0.92 \cdot 0.85 \cdots$$

tr thịnh mt phôp cng hojn tojn vñ hi:

$$\ln(0.99) + \ln(0.92) + \ln(0.85) + \cdots$$

Phổ cung luân n nh vị d o hịm hn phosp nhón.

Khung ch vy, log cùn koo còc xòc sut v mt di giò tr d qun lỳ:

- $\ln(0.99) \approx -0.01$
- $\ln(0.92) \approx -0.083$
- $\ln(0.85) \approx -0.16$

Còc con s nh gn, khung b lt vio vung quò nh nh 10^{-50} .

Dứng log vit li Likelihood

Ta õ cú Likelihood gc:

$$\mathcal{L}(\theta) = \prod_{i=1}^m h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}.$$

Ly log hai v (dứng Product Rule v Power Rule):

$$\ln(a \cdot b) = \ln(a) + \ln(b), \quad \ln(a^k) = k \ln(a).$$

Ta c

$$\ell(\theta) = \ln \mathcal{L}(\theta) = \sum_{i=1}^m \left[y^{(i)} \ln h_\theta(x^{(i)}) + (1 - y^{(i)}) \ln (1 - h_\theta(x^{(i)})) \right].$$

Hịm c gi lị Log-Likelihood? v

- $\mathcal{L}(\theta)$ lị Likelihood (mt phosp nhón).
- $\ell(\theta) = \ln(\mathcal{L}(\theta))$ lị Log-Likelihood (mt phosp cng). iu quan trng nht lị Hịm log lị hịm n iu, tc lị vic ti a húa Likelihood vti ti a húa Log-Likelihood lị nh nhau:

$$\theta^* = \arg \max_{\theta} \ell(\theta) \iff \theta^* = \arg \max_{\theta} \mathcal{L}(\theta).$$

Tutm li mnh mt cung i qua log lị v: Log-Likelihood d tònh loss hn, d o hịm hn, d phón tòch hñh dng hn, mt hn cho vic ti u.

2.9 Mt vờ d rǔ rìng: Log-Likelihood trong bii toòn d oòn ung th

Gi s bn cù 5 bnh nhón, nhón nh sau (ging vờ d phn Likelihood):

PSA	y	$h_\theta(x)$
3.8	1	0.99
3.4	1	0.97
2.9	1	0.92
2.1	0	0.50
1.2	0	0.13

Ta õ tònh c Likelihood trc ú:

$$\mathcal{L}(\theta) \approx 0.386.$$

Gi tònh Log-Likelihood:

$$\ell(\theta) = \ln(0.99) + \ln(0.97) + \ln(0.92) + \ln(0.50) + \ln(0.87).$$

Tònh tng giò tr:

$$\begin{aligned}\ln(0.99) &\approx -0.01005, \\ \ln(0.97) &\approx -0.03046, \\ \ln(0.92) &\approx -0.08338, \\ \ln(0.50) &\approx -0.69315, \\ \ln(0.87) &\approx -0.13926.\end{aligned}$$

Cng li:

$$\ell(\theta) \approx -0.956.$$

So sòn trc giòc:

- Likelihood: 0.386 rt nh nhng khú so sòngh gia mū hñnh.

- Log-Likelihood: -0.956 con s t nhiñn, d c, d so sòngh.

Nu mt mū hñnh khòc cú LL = -10 chc chn t hn.

Nu LL = -0.3 tt hn.

Ý ngha sóu hn: Log-Likelihood thng/pht mū hñnh th njo?

Quan sòt cu trñc:

$$y \ln(h) \quad \text{vì} \quad (1 - y) \ln(1 - h).$$

Nu $y = 1$:

$$\ell_i = \ln(h_\theta(x^{(i)})).$$

Nu mū hñnh d oòn ñng vñ t tin (h gn 1), log gn 0 **im thng ln**. Nu d oòn sai t tin (h gn 0), log tin v $-\infty$ **pht rt mnh**.

Nu $y = 0$:

$$\ell_i = \ln(1 - h_\theta(x^{(i)})).$$

- Nu mū hñnh gn 0 tt

- Nu mū hñnh gn 1 log rt óm pht mnh. óy chònh lị lỳ do Log-Likelihood phn ònh tt cht lñg mū hñnh phón loi: mū hñnh sai t tin s b trng pht rt nng.

2.10 Binary Cross-Entropy: Hñm Loss xut phòt t nhiñn t Log-Likelihood

Sau khi õ xóy dng Log-Likelihood, ta cú mt biu thc mū t mc hp lỳ ca tham s θ khi nhñn viø tojn b d liu. Tuy nhiñn, trong hc móy, ta thng quen vi *ti thi u húa* thay vñ *ti a húa*. Vñ vy, theo l rt t nhiñn lị ti thi u hñm Loss:

$$J(\theta) = -\ell(\theta).$$

Biu thc niý chònh lị **Binary Cross-Entropy (BCE)**. Khñng phi c nh ngaha tu ý, mị lị kt qu tt yu khi ta ly du tr ca log-likelihood.

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

Nu mu thuc lp 1 ($y = 1$), ta nhñn viø $-\ln(h_\theta(x))$.

- Nu mū hñnh t tin ũng ($h \approx 1$), $-\ln(h)$ gn 0 **đt pht.**
- Nu mū hñnh t tin sai ($h \approx 0$), $-\ln(h)$ tin ti $+\infty$ **pht cc mnh.**

Nu mu thuc lp 0 ($y = 0$), ta nhñn vïo $-\ln(1 - h_\theta(x))$.

- Nu mū hñnh t tin ũng ($h \approx 0$), $-\ln(1 - h)$ gn 0.
- Nu mū hñnh t tin sai ($h \approx 1$), $-\ln(1 - h)$ tin ti $+\infty$.

2.11 Tng kt bc ca Logistic Regression

hiu sôu Binary Cross-Entropy, ta cn nhñn li tojn b mū hñnh Logistic Regression dng c bn nht. Mc tiôu ca Logistic Regression lì mū hñnh húa xòc sut mt im d liu thuc lp 1 di dng:

$$h_\theta(x) = P(y = 1 | x; \theta).$$

Höy phón tòch tng thịnh phn mt còch lin mch vï trc quan.

(1) Mū hñnh tuyn tònh: $z = \theta^T x$

$$z = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n.$$

Trong ú, x lì vector c trng u vïo, θ lì vector trng s mū hñnh cn hc vï $x_0 = 1$ biu din h s chn (bias). Mū hñnh ly tng cù trng s ca còc c trng to thịnh mt dng im s cú th óm, dng hoc rt ln, nhng cha th din gii nh xòc sut; vñ th ta cn n hñm Sigmoid.

(2) Bin i z thịnh xòc sut bng hñm Sigmoid

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

Sigmoid ốp u ra v on $[0, 1]$, cho phôp din gii nh xòc sut: khi z ln tho $h \rightarrow 1$, khi z nh tho $h \rightarrow 0$, cùn khi $z = 0$ th hin trng thòi khñng bit / trung lp. Vñ vy, Sigmoid úng vai trù nh b chuyn i gia khñng gian tuyn tònh vï khñng gian xòc sut.

(3) Xòc sut d oòn cho tng nhõn y

Theo phón phi Bernoulli:

$$P(y | x, \theta) = h_\theta(x)^y (1 - h_\theta(x))^{1-y}.$$

Cñng thc niy hot ng nh s m y vï $1 - y$ úng vai trù nh cñng tc: khi $y = 1$ nú gi li h , cùn khi $y = 0$ nú gi li $(1 - h)$. Nh vy, mt cñng the duy nht cù th biu din c hai trng hp.

(4) Likelihood cho c tp d liu

Vñ còc mu c lp theo gi nh IID, ta nhón tt c xòc sut li o hp lỳ tng th ca mū hñnh:

$$\mathcal{L}(\theta) = \prod_{i=1}^m h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}.$$

Phôp nhón xut hin vñ mi im d liu lì mt bng chng c lp ng h hoc bòc b θ , vï tòch ca chñng th hin mc mū hñnh phû hp vi tojn b d liu.

(5) Log-Likelihood

Tùch trồn rt nh khi nhón nhieu xòc sut dn n underflow, nőn ta ly log chuyn tùch thịnh tng d tònh vị n nh:

$$\ell(\theta) = \sum_{i=1}^m \left[y^{(i)} \ln h_\theta(x^{(i)}) + (1 - y^{(i)}) \ln (1 - h_\theta(x^{(i)})) \right].$$

Vic ly log giüp s nh tr nőn ln hn, loi b vn trjn s, ng thi gi nguyễn im cc i vơ him log lị him n iu.

(6) Binary Cross-Entropy: NegLogLikelihood

$$J(\theta) = -\ell(\theta).$$

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

Vic thóm du tr bin bịi toòn ti a húa thịnh ti thiu húa, giüp BCE tr thịnh him Loss chun. BCE thng mū hñnh oòn ũng vị pht rt mnh mū hñnh oòn sai t tin, ng thi phn ònh ũng bn cht thng kổ ca phón phi Bernoulli.

Gii thùch pipeline ca Logistic Regression theo trc giòc lin mch

- $\theta^T x$ to ra mt im s tuyn tònh da trồn c trng u viø.
- $\sigma(\theta^T x)$ chuyn im s ú thịnh xòc sut hp l nm trong khong $[0, 1]$.
- $h_\theta(x)^y(1 - h_\theta(x))^{1-y}$ mū t xòc sut mū hñnh cho im ũng sai cho tng mu d liu.
- $\prod_{i=1}^m$ kt hp tt c bng chng c lp thịnh mt mc hp lỳ tp th ca mū hñnh.
- log bin thùch thịnh tng tròn underflow vị lìm cho vic o him d đing hn.
- $- \log \text{bin}(y; \theta)$ toòn thịnh ti thiu húa, phứ hp vi còc thut toòn ti u chun nh Gradient Descent.
- Tng cui cúng tr thịnh mt giò tr Loss: mū hñnh còng chònh xòc tho Loss còng nh.

Kt lun cū ng

Logistic Regression = Bernoulli Model + Log-Likelihood \Rightarrow Binary Cross-Entropy.
--

Mi biu thc u xut phòt t nguyễn lỳ xòc sut.

Phn 3 M Rng: Nhón ma trn vi Logistic Regression

phn trc, ta õ xóy dng c him Loss ca Logistic Regression dng tng mu (khñg đúng ma trn). Tuy nñiñ, khi s lñg c trng vị s lñg mu tr nőn ln, còch vit tng minh theo tng im d liu tr nőn cng knh vị khú ti u hoò. x lỳ mū hñnh quy mū ln, ta cn chuyn tojn b bịi toòn v dng **vectormatrix**. Ngoi ra, nñ cng gng vic vit him tònh toòn trong code gn vị d đing hn.

3.1 Khi ng bng mt bij toòn c th

Gi s ta cú ba bn̄nh nhón vi ba giò tr PSA v̄i ta mun mū hñnh Logistic Regression d oòn xòc sut mc bn̄nh. Bng d liu:

Bnh nhón	$x^{(i)}$	$y^{(i)}$
1	1.0	1
2	2.0	1
3	3.0	0

H s mū hñnh:

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} -3 \\ 2 \end{bmatrix}.$$

Ta mun tònh:

$$h_\theta(x^{(i)}) = \sigma(\theta_0 + \theta_1 x^{(i)}).$$

Nu ljm th cñng tng mu:

$$\begin{aligned} z^{(1)} &= -3 + 2(1) = -1, & h^{(1)} &= \sigma(-1), \\ z^{(2)} &= -3 + 2(2) = 1, & h^{(2)} &= \sigma(1), \\ z^{(3)} &= -3 + 2(3) = 3, & h^{(3)} &= \sigma(3). \end{aligned}$$

Cóch nijy hojn tojn ũng nhng khñng th m rng cho 1000 mu, 20 c trng.

3.2 Ún li nhón ma trn t gúc nhon tng trng s

Nhón ma trn ch lị:

$$(hñng ca ma trn) \cdot (\text{vector}) = \text{tng cù trng s.}$$

Vđ d:

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 1 \cdot 3 + 2 \cdot 4 = 11.$$

Trc giòc:

- ma trn lị tp hp nhieu hñng,
- mi hñng lị mt mu d liu,
- nhón ma trn vi vector θ n gin lị tònh im s tuyn tònh z cho tt c mu cñng lñc.

3.3 Chuyn d liu sang dng ma trn

Ta gom tt c u vjо vjо mt ma trn thit k:

$$X = \begin{bmatrix} 1 & 1.0 \\ 1 & 2.0 \\ 1 & 3.0 \end{bmatrix}.$$

- Ct u tiøn lị 1 (cho bias term θ_0).
- Mi hñng lị mt bn̄nh nhón.

Khi ú:

$$z = X\theta = \begin{bmatrix} 1 & 1.0 \\ 1 & 2.0 \\ 1 & 3.0 \end{bmatrix} \begin{bmatrix} -3 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}.$$

3.4 Ôp sigmoid lǒn toịn b vector cúng lữc

$$h = \sigma(z) = \begin{bmatrix} \sigma(-1) \\ \sigma(1) \\ \sigma(3) \end{bmatrix}.$$

Bóy gi ta ō tònh c xòc sut cho toịn b dataset ch bng hai bc:

- nhón ma trn: $z = X\theta$,
- ôp sigmoid: $h = \sigma(X\theta)$.

3.5 Loss ca Logistic Regression dng vector

Theo nh ngha phn trc, hịm Loss cho Logistic Regression dng tng mu lị:

$$J(\theta) = - \sum_{i=1}^m \left[y^{(i)} \ln h^{(i)} + (1 - y^{(i)}) \ln(1 - h^{(i)}) \right].$$

x lỳ toịn b d liu ng thi, ta chuyn sang dng vector hoò. Khi ú:

$$J(\theta) = - (y^T \ln(h) + (1 - y)^T \ln(1 - h)),$$

trong ú:

- $y = [y^{(1)}, y^{(2)}, \dots, y^{(m)}]^T$ lị vector nhõn.
- $h = [h^{(1)}, h^{(2)}, \dots, h^{(m)}]^T$ lị vector xòc sut mū hñnh d oòn.
- $\ln(h)$ lị log theo tng phn t:

$$\ln(h) = [\ln h^{(1)}, \ln h^{(2)}, \dots, \ln h^{(m)}]^T.$$

Cũng thc vector hoò niy giüp Loss c vit gn gingga hn vị cho phôp tònh gradient vị hessian d dìng.

Và d y : Tònh toịn b Loss bng vector cho bji toòn PSA

Ta s dng li dataset nh gm 3 bnh nhón:

$$X = \begin{bmatrix} 1 & 1.0 \\ 1 & 2.0 \\ 1 & 3.0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \theta = \begin{bmatrix} -3 \\ 2 \end{bmatrix}.$$

Bc 1: Tònh $z = X\theta$

$$z = \begin{bmatrix} 1 & 1.0 \\ 1 & 2.0 \\ 1 & 3.0 \end{bmatrix} \begin{bmatrix} -3 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}.$$

Bc 2: Tính $h = \sigma(z)$

$$h = \begin{bmatrix} \sigma(-1) \\ \sigma(1) \\ \sigma(3) \end{bmatrix} = \begin{bmatrix} 0.2689 \\ 0.7311 \\ 0.9526 \end{bmatrix}.$$

Bc 3: Tính log tng phn t

$$\ln(h) = \begin{bmatrix} \ln 0.2689 \\ \ln 0.7311 \\ \ln 0.9526 \end{bmatrix} = \begin{bmatrix} -1.313 \\ -0.313 \\ -0.048 \end{bmatrix},$$

$$\ln(1 - h) = \begin{bmatrix} \ln(1 - 0.2689) \\ \ln(1 - 0.7311) \\ \ln(1 - 0.9526) \end{bmatrix} = \begin{bmatrix} -0.314 \\ -1.313 \\ -3.046 \end{bmatrix}.$$

Bc 4: Tính tng phn ca Loss

$$y^T \ln(h) = [1 \ 1 \ 0] \begin{bmatrix} -1.313 \\ -0.313 \\ -0.048 \end{bmatrix} = (-1.313) + (-0.313) + (0) = -1.626.$$

$$(1 - y)^T \ln(1 - h) = [0 \ 0 \ 1] \begin{bmatrix} -0.314 \\ -1.313 \\ -3.046 \end{bmatrix} = -3.046.$$

Bc 5: Loss cui cứng

$$J(\theta) = -(-1.626 + (-3.046)) = 4.672.$$

Ý nghĩa vò d

- Hai mu u cú nhõn 1 nhng mū hñnh d oòn khñng quò cao (0.26 vñ 0.73), dn n giò tr $\ln(h)$ óm òng k.
- Mu th ba cú nhõn 0 nhng mū hñnh li d oòn gn 1 (0.95), dn n $\ln(1 - h)$ rt óm vñ to Loss ln.
- Loss tng cng 4.672 lị khò cao cho dataset nh nh th niy, phn ònh mū hñnh cha tt.

Nh vò d niy, ta thy rù:

- dng vector hoò ca Logistic Regression hojn tojn tng ng dng tng mu,
- nhng cù ng vñ phñt hp cho vic trin khai bng phn mm vñ cho phón tòch o hñm vector phn sau.

Kt ni sang phn o hñm vñ Hessian

Ngay sau óy, khi tònh o hñm ca Loss, ta s thy dng vector em li li th vt tri.

Cñng thc p:

$$\nabla_{\theta} J(\theta) = X^T(h - y)$$

3.6 Convexity of BCE for Multiple-Variable using Hessian Matrix

Phn trc, ta ō thit lp Logistic Regression di dng vector vị tōnh c Gradient ca h̄m Loss. Phn n̄y s m rng l̄n o h̄m bc hai (Second Derivative) xóy dng **Hessian Matrix** cho ta tinh chnh nhieu θ weight cung 1 l̄c, giúp tóm loss hiu qu hn, t ú chng minh rng Binary Cross-Entropy (BCE) l̄i mt h̄m convex.

1. Ứn li Second Derivative: t 1 bin n nhieu bin

Trong h̄m mt bin $f(x)$:

$$f'(x) = \text{tc thay i ca } f(x), \quad f''(x) = \text{tc thay i ca tc thay i.}$$

Second Derivative

	1 variable	2 variables
Function	$f(x)$	$f(x, y)$
First derivative	$f'(x)$ Rate of change of $f(x)$	$f_x(x, y)$ Rate of change w.r.t x $f_y(x, y)$ Rate of change w.r.t y $\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$
Second derivative	$f''(x)$ Rate of change of the rate of change of $f(x)$???

Figure 22: So s̄nh o h̄m bc nht v̄i bc hai gia mt bin v̄i hai bin. (w.r.t x - with respect to / o h̄m 1 phn theo x)

Vi h̄m hai bin $f(x, y)$, o h̄m bc nht tr th̄nh vector gradient:

$$\nabla f = \begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}.$$

Vy o h̄m bc hai ca h̄m hai bin tr̄ng th n̄o?

Ta s xem x̄t mt v̄o d c th trc giòc r̄u hn.

2. V̄o d thc t: Tōnh o h̄m bc hai cho $f(x, y)$

X̄t h̄m:

$$f(x, y) = 2x^2 + 3y^2 - xy.$$

Second Derivative

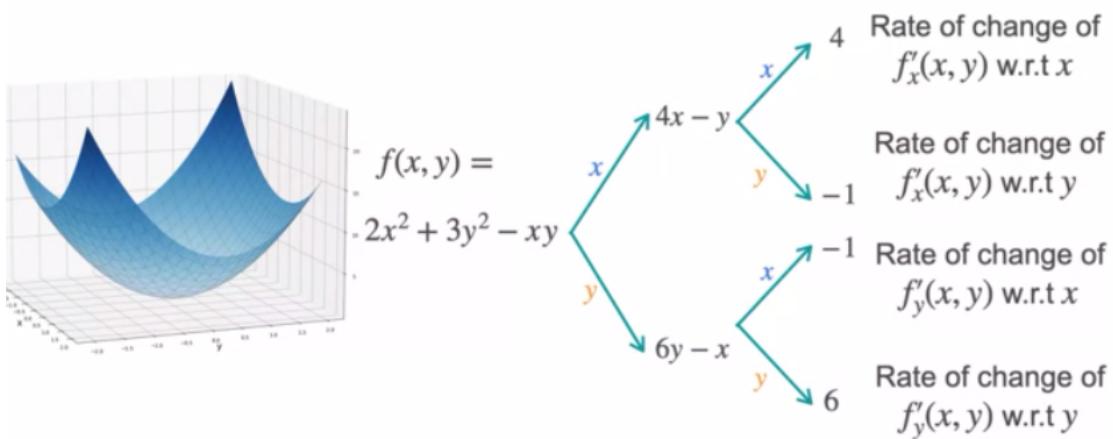


Figure 23: Cây o him bc nht vị bc hai cho him $f(x, y) = 2x^2 + 3y^2 - xy$.

o him bc nht:

$$f_x = 4x - y, \quad f_y = 6y - x.$$

o him bc hai:

$$f_{xx} = 4, \quad f_{yy} = 6, \quad f_{xy} = f_{yx} = -1.$$

What Do These Mean?

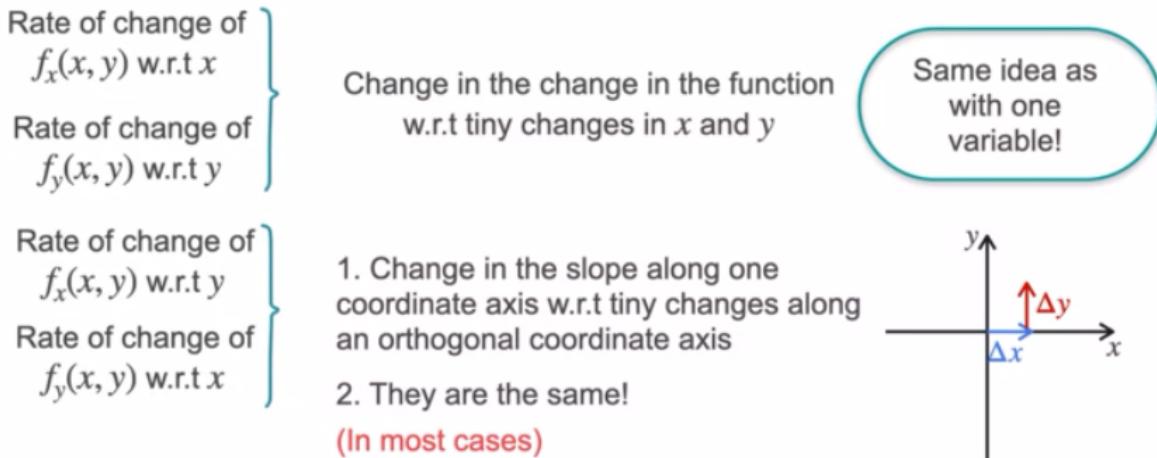


Figure 24: Ý nghia ca còc o him bc hai vị o him chõo.

Còc o him bc hai mñ t change in the change: thay i ca o him theo mt hng khi ta bin i theo mt hng khõc (cú th trc giao - orthogonal - 2 vector vuông góc).

3. Ký hiu o him bc hai: Leibniz vị Lagrange

Hình 25 minh ha hai kiu ký hiu:

	Leibniz's notation	Lagrange's notation
Rate of change of $f'_x(x, y)$ w.r.t x	$\frac{\partial^2 f}{\partial x^2}$	$f_{xx}(x, y)$
Rate of change of $f'_y(x, y)$ w.r.t y	$\frac{\partial^2 f}{\partial y^2}$	$f_{yy}(x, y)$
Rate of change of $f'_x(x, y)$ w.r.t y	$\frac{\partial^2 f}{\partial x \partial y}$	$f_{xy}(x, y)$
Rate of change of $f'_y(x, y)$ w.r.t x	$\frac{\partial^2 f}{\partial y \partial x}$	$f_{yx}(x, y)$

Figure 25: Hai h kỲ hiu cho o hÌm bc hai: Leibniz vÌ Lagrange.

- Leibniz: $\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial x \partial y}$.
- Lagrange: f_{xx}, f_{xy} .

4. Gom tojn b o hÌm bc hai vÌo mt cu trÙc: Hessian Matrix

Nh minh ho HÌnh 26:

$$H(f) = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}.$$

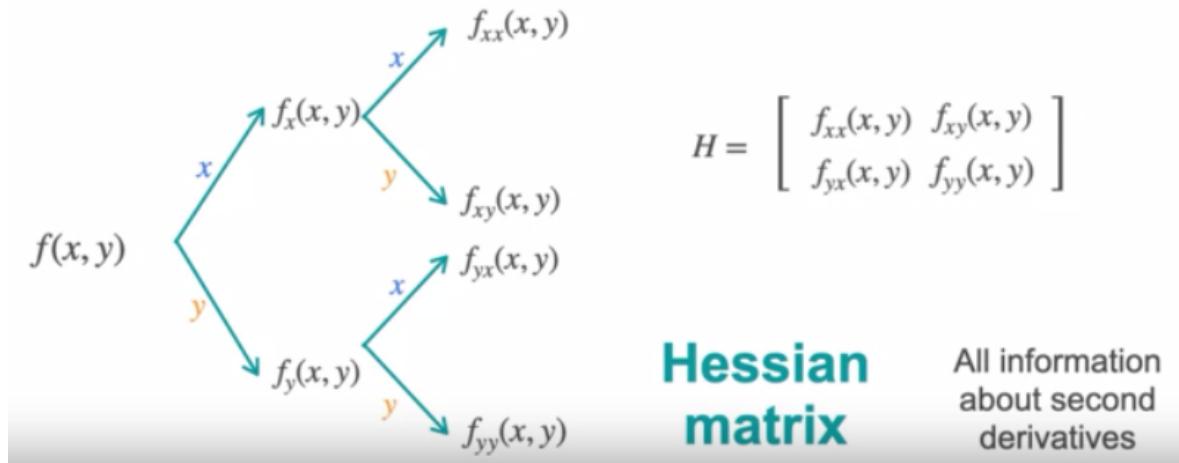


Figure 26: CÙch gom cÙc o hÌm bc hai vÌo ma trÙn Hessian.

Hessian cha tojn b thÙng tin vÌ cong (curvature) ca hÌm $f(x, y)$. Tng t nh lÙc ta cÙng gii thÙch o him bc 2 phn trc.

5. Dứng Hessian phón loi im ti u

Bng còch phón tòch ma trn Hessian, ta cú th xòc nh 1 im lị cc tiu, cc i hay lị im yển nga (khi cc tiu ca hịm bỗ hn 0, nhng o hịm ra 0) da vịo Eigen value ca nú ((giò tr riổng ca ma trn).

? Bonus Fact: By analyzing the Hessian matrix, we can determine whether a point is a minimum, maximum, or a saddle point

Second Derivative

Function	1 variable		2 variables	
	$f(x)$		$f(x, y)$	
First derivative	$f'(x)$	Rate of change of $f(x)$	$f_x(x, y)$	Rate of change w.r.t x
			$f_y(x, y)$	Rate of change w.r.t y
			$\nabla f = \begin{bmatrix} f_x(x, y) \\ f_y(x, y) \end{bmatrix}$	
Second derivative	$f''(x)$	Rate of change of the rate of change of $f(x)$	$H(x, y) = \begin{bmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{bmatrix}$	

Figure 27: So sòngh bc hai trong 1 bin vị nhieu bin.

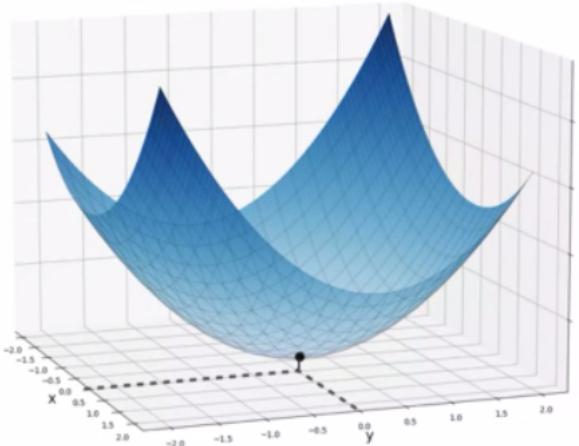
Quy tíc phón loi:

- nếu mi eigenvalue ca Hessian > 0 hịm cong lõn (local minimum),
- nếu mi eigenvalue < 0 cong xung (local maximum),
- nếu có eigenvalue > 0 và < 0 saddle point,
- nếu có eigenvalue $= 0$ khũng thũng tin.

Vở d minh ho qua còch xòc nh cc tiu da trồn giò tr ca ma trn Hessian:

Concave Up (cc tiu)

Concave Up



$$f(x,y) = 2x^2 + 3y^2 - xy$$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

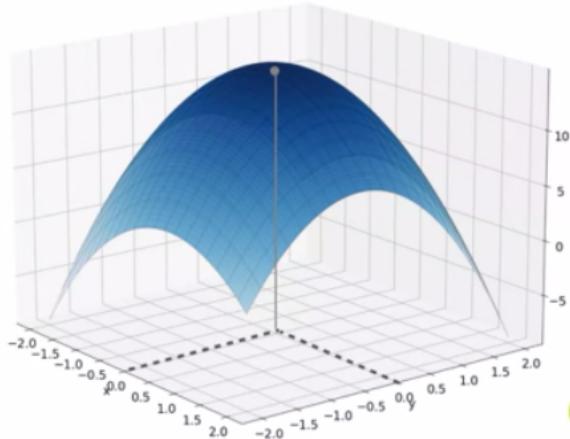
$$\begin{aligned} \det(H(0,0) - \lambda I) &= \det \left(\begin{bmatrix} 4-\lambda & -1 \\ -1 & 6-\lambda \end{bmatrix} \right) \\ &= (4-\lambda)(6-\lambda) - (-1)(-1) \\ &= \lambda^2 - 10\lambda + 23 \end{aligned}$$

$$\begin{aligned} \lambda_1 &= 6.41 \\ \lambda_2 &= 3.59 \end{aligned}$$

Figure 28: Hessian có eigenvalue dng concave up im (0,0) lị cc tiu.

Concave Down (cc i)

Concave Down



$$f(x,y) = -2x^2 - 3y^2 - xy + 15$$

$$\nabla f(x,y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\begin{aligned} \det(H(0,0) - \lambda I) &= \\ &(-4-\lambda)(-6-\lambda) - (-1)(-1) \\ &= \lambda^2 + 10\lambda + 23 \end{aligned}$$

(0,0) is a maximum!

$$< 0$$

Figure 29: Hessian có eigenvalue óm concave down cc i.

Saddle Point (im Yển Nga) Khi o hịm bng 0 thợ ngoịj còc trng hp ri vịo còc im cc tr thợ cùn mt im na chòn lị im yển nga. Mt im yển nga lị im mị ti ú o hịm bng 0 nhng khũng phi lị mt im cc tiu. Xõt hịm s $f(x) = 5^3$ hịm s nịy cù o hịm bng 0 ti $x=0$ nhng $x=0$ li khũng phi giò tr cc tiu ca hịm s nịy vơ nú cù th óm.

Saddle Point

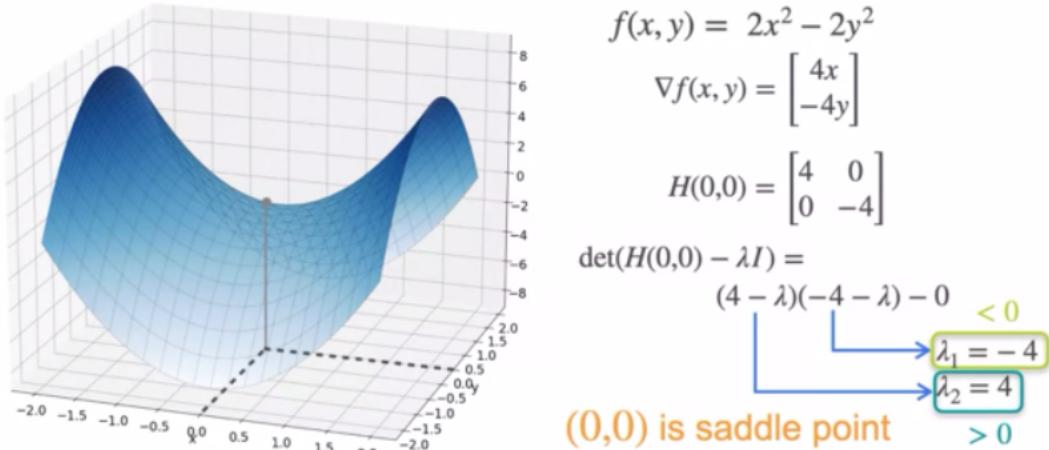


Figure 30: Eigenvalue trồi du → saddle point.

6. ng dng trc tip ṿo Logistic Regression

Sau khi ō hiu Hessian ca h̄m hai bin c̄c phn trc, ta quay tr li Logistic Regression xóy dng Hessian ca Binary Cross-Entropy (BCE). Tojn b qūo tronh di óy c vit cc k chi tit nhm gīp ngi c ūn li tōn t gc.

6.1 Nhc li dng Loss ṿi Gradient

H̄m Loss dng vector ca Logistic Regression:

$$J(\theta) = -[y^T \ln(h) + (1-y)^T \ln(1-h)], \quad h = \sigma(X\theta).$$

Gradient (o h̄m bc nht theo vector θ):

$$\nabla_{\theta} J(\theta) = X^T(h - y).$$

BCE l̄i convex nu Hessian (o h̄m bc hai) ca $J(\theta)$ l̄i Positive Semi-Definite (i.e. ch cn c̄c tr rīng ca n̄u l̄i kh̄ng óm l̄i positive semi-definite / b̄n dng).

6.2 Bt u t Gradient t̄m Hessian

Ta cn o h̄m ca:

$$X^T(h - y)$$

theo θ .

6.3 o h̄m ca $h = \sigma(X\theta)$

u tīn ta x̄t:

$$z = X\theta.$$

Gii thòch: óy lị phôp nhón ma trn c bn: mi hịng ca X nhón vi vector θ to nőn mt giò tr $z^{(i)}$. Nh rng vi ma trn $(m \times n)$ nhón vector $(n \times 1)$:

$$(X\theta)_i = \sum_{j=1}^n X_{ij}\theta_j.$$

(a) o hịm ca Sigmoid

Sigmoid:

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

o hịm:

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

Khi օp dng cho vector:

$$h = \begin{bmatrix} h^{(1)} \\ h^{(2)} \\ \vdots \\ h^{(m)} \end{bmatrix}, \quad h^{(i)} = \sigma(z^{(i)}).$$

Do ú:

$$\frac{\partial h^{(i)}}{\partial z^{(i)}} = h^{(i)}(1 - h^{(i)}).$$

(b) o hịm ca $z = X\theta$

Vđ:

$$z^{(i)} = X_{i,:}\theta$$

trong ú i lị hịng, : ngha lị mi ct nőn:

$$\frac{\partial z^{(i)}}{\partial \theta} = X_{i,:}$$

óy lị quy tc tuyn tòngh quen thuc: o hịm ca mt tng tuyn tòngh theo vector chòngh lị vector h s.

(c) Օp dng Chain Rule ly $\frac{\partial h}{\partial \theta}$

Theo chain rule a bin:

$$\frac{\partial h^{(i)}}{\partial \theta} = \frac{\partial h^{(i)}}{\partial z^{(i)}} \cdot \frac{\partial z^{(i)}}{\partial \theta}.$$

Thay vđo:

$$\frac{\partial h^{(i)}}{\partial \theta} = h^{(i)}(1 - h^{(i)}) \cdot X_{i,:}$$

Gp tt c mu thịnh mt ma trn:

$$\frac{\partial h}{\partial \theta} = DX,$$

vi:

$$D = \text{diag} \left(h^{(i)}(1 - h^{(i)}) \right).$$

trong ú $\text{diag}()$ lị ng chõo ca ma trn, i din cho còc giò tr Eigenvaleu (vector riõng) ca mi phõp tònh trong ma trn.

6.4 Bóy gi o him Gradient tóm Hessian

Gradient (o him bc 1):

$$\nabla_{\theta} J(\theta) = X^T(h - y).$$

o him bc 2 theo θ :

$$H = \nabla_{\theta}^2 J(\theta) = X^T \frac{\partial h}{\partial \theta}.$$

Thay $\frac{\partial h}{\partial \theta} = DX$:

$$H = X^T D X.$$

óy chònh lị Hessian ca BCE.

6.5 Gii thòch hñh thc ca $H = X^T D X$

hiu sóu hn, ta phón rõ theo tng lp:

- X : mū t c trng ca d liu.
- X^T : tng hp nghch o tng tham s θ_j .
- D : mū t cong ca Sigmoid ti tng mu, vø $h(1 - h)$ chònh lị nhý (sensitivity) ca Sigmoid.

Do ú $X^T D X$ lị còch mū hñh kt hp cong t Sigmoid vi cu trñc ca d liu.

Hñh 26 ō minh ha vic gom còc o him bc hai viø ma trn Hessian; ta ang lịm iu tng t nhng m rng lñn nhieu bin vñ nhieu mu.

6.6 Vø sao D luñ khñg óm (PSD component)?

Ta nhc li:

$$D_{ii} = h^{(i)}(1 - h^{(i)}).$$

Ta bit rng Sigmoid tr v giò tr t 0 n 1:

$$0 < h^{(i)} < 1.$$

Do ú:

$$h^{(i)}(1 - h^{(i)}) > 0 \quad \text{vì} \quad h^{(i)}(1 - h^{(i)}) \leq \frac{1}{4}.$$

Ngha lị mi phn t ng chõo ca D luñ khñg óm.

Vø vy, D lị mt ma trn ng chõo **Positive Semi-Definite**.

6.7 Chng minh Hessian PSD: phón rō tng bc

Ta cn chng minh:

$$v^T H v \geq 0 \quad \forall v.$$

Thay $H = X^T D X$:

$$v^T X^T D X v.$$

Nhóm li:

$$(Xv)^T D (Xv).$$

t:

$$z = Xv.$$

Khi ú:

$$z^T D z = \sum_{i=1}^m D_{ii} z_i^2.$$

Vo: $D_{ii} \geq 0$ ví $z_i^2 \geq 0$
nǎn mi s hng u khūng óm tng cng khūng óm:

$$\Rightarrow z^T D z \geq 0.$$

Suy ra:

H lị Positive Semi-Definite.

Ví do ú:

$J(\theta)$ lị mt him convex.

6.8 Trc giòc hñnh hc cui cùng

- Sigmoid luñn cú cong dng: $h(1 - h)$.
- D liu ch nh hng ca cong thñng qua X .
- $H = X^T D X$ thu gom tojn b cong ca Sigmoid theo mi hng trong khñng gian tham s.
- Khñng tn ti hng njo mị Loss cong xung khñng có local minima.

Convexity ca BCE n t cong ni ti ca Sigmoid + cu trñc tuyn tñnh ca mñ hñnh.