Maior Dúvida da Aula

Linear Regression

MC886/MO444, September 6, 2022

Equação Normal

1. Entendi como funciona a descida do gradiente e a equação normal, porém como saber qual a melhor ferramenta para o nosso problema? Quando utilizar cada uma?
2. Existe um limite de dados que as pessoas consideram o gradiente descendente melhor do que a equação normal?

Regressão Polinomial

3. Não entendi a vantagem de executar o modelo com muitos graus. Seria para validar a importância das variáveis combinadas?

4. A função de custo da regressão polinomial é convexa?

Perguntas Gerais

5. Não entendi muito bem o que significa o Batch Size. Ele é uma parte dos exemplos de treino? Mas, se fosse assim, ele não seria um Mini-Batch?

6. Sobre o tamanho dos batches, há algum consenso de quanto ele deveria ser? Digo em termos de ordem de magnitude (10, 100, 1000, 10000, ...).

7. Quando estamos trabalhando em problemas de regressão linear/regressão polinomial existe um protocolo indicado para transformar os dados categóricos em numéricos?

Categorical/Nominal Variables

Size (feet^2) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Color x_5	Price ($) in 1000's y
2104	5	1	45	blue	460
1416	3	2	40	white	232
1534	3	2	30	pink	315
852	2	1	36	green	178

https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding

https://www.kaggle.com/code/arashnic/an-overview-of-categorical-encoding-methods/notebook

Perguntas Gerais

8. Durante a normalização, a ideia é que a entrada fique entre 0,5 e -0,5 mapeando ao valor real: Exemplo da casa é tirar 1000 e dividir por 2000. Tem problema na instanciação do modelo se minha feature nesse caso saia desse intervalo, que nesse caso seria uma casa com mais de 2000 m²?

9. What metric is the best to measure error between RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error)?

Logistic Regression

Machine Learning

(Largely based on slides from Andrew Ng)

Prof. Sandra Avila

Institute of Computing (IC/Unicamp)

MC886/MO444, September 6, 2022

Today's Agenda

- Logistic Regression
  - Classification
  - Hypothesis Representation
  - Decision Boundary
  - Cost Function
  - Simplified Cost Function and Gradient
  - Multiclass Classification

Classification

Spam Filtering

Bad Cures fast and effective! - Canadian *** Pharmacy #1 Internet Inline Drugstore Viagra Cheap Our price $1.99 …

Good Interested in your research on graphical models - Dear Prof., I have read some of your papers on probabilistic graphical models. Because I …

Classification

Email: **Spam** / **Not Spam**?

Content Video: **Sensitive** / **Non-sensitive**?

Skin Lesion: **Malignant** / **Benign**?

Classification

Email: **Spam** / **Not Spam**?

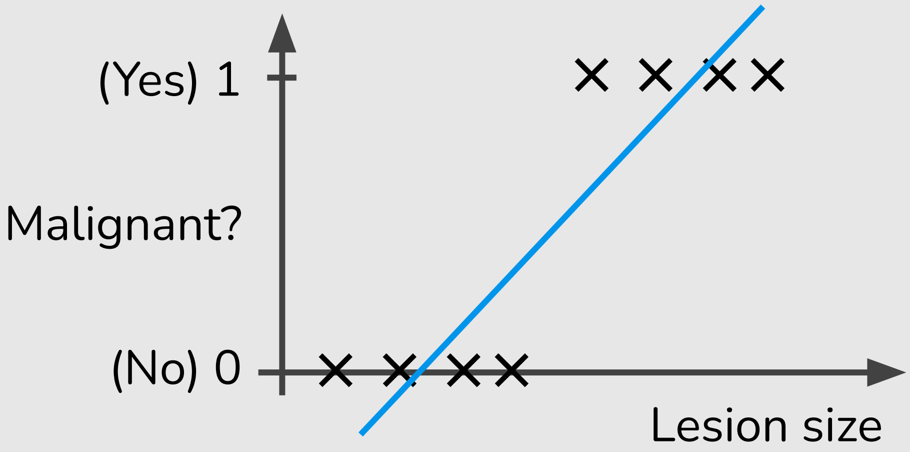
Content Video: **Sensitive** / **Non-sensitive**?

Skin Lesion: **Malignant** / **Benign**?

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

0: "Negative Class" (e.g., Benign skin lesion)

1: "Positive Class" (e.g., Malignant skin lesion)

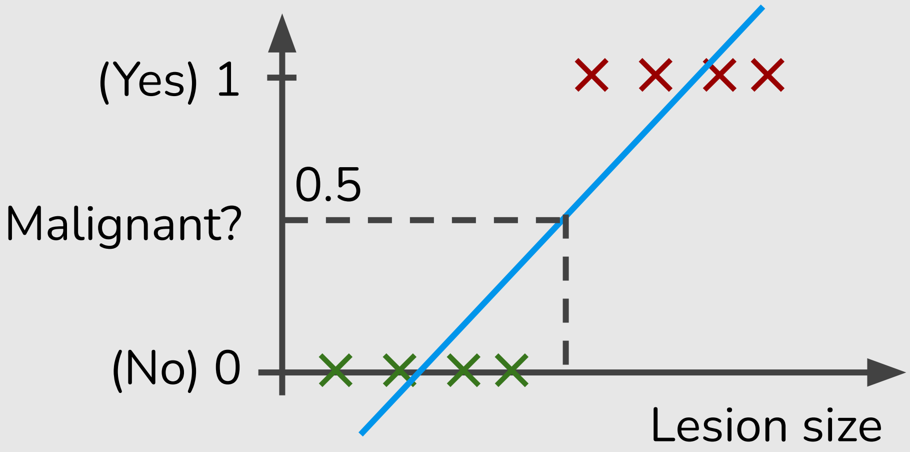


$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"



$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

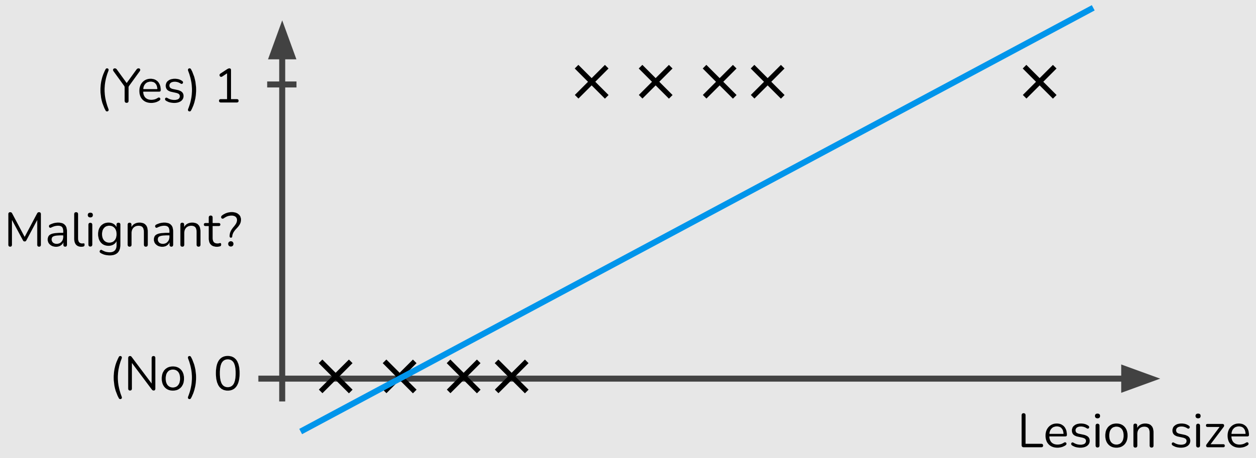
If $h_\theta(x) < 0.5$, predict "$y = 0$"

$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"

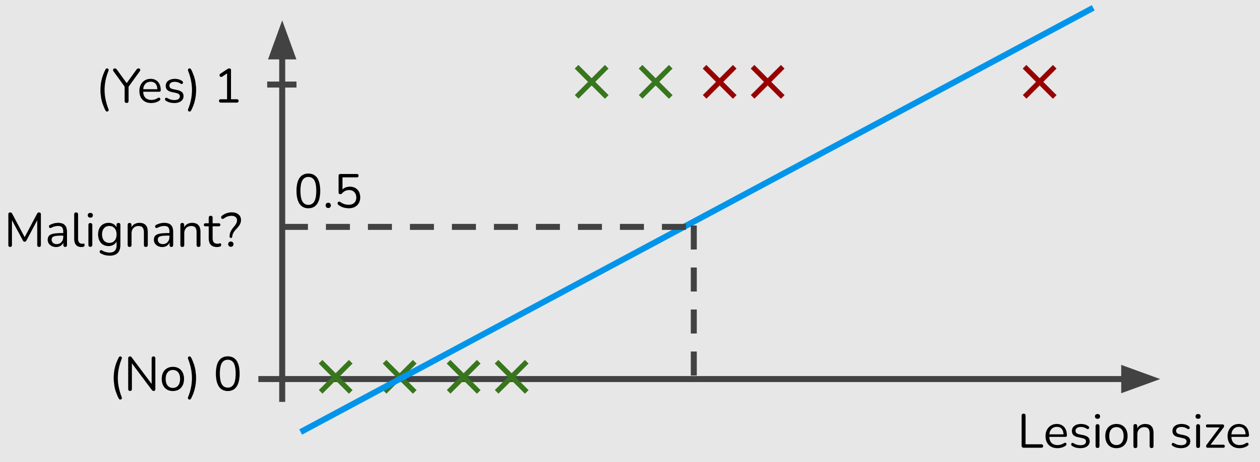


$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"



$$h_\theta(x) = \theta^T x$$

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "$y = 1$"

If $h_\theta(x) < 0.5$, predict "$y = 0$"

Classification: $y = 0$ or $y = 1$

$h_\theta(x)$ can be > 1 or < 0

Logistic Regression: $0 \leq h_\theta(x) \leq 1$

Hypothesis Representation

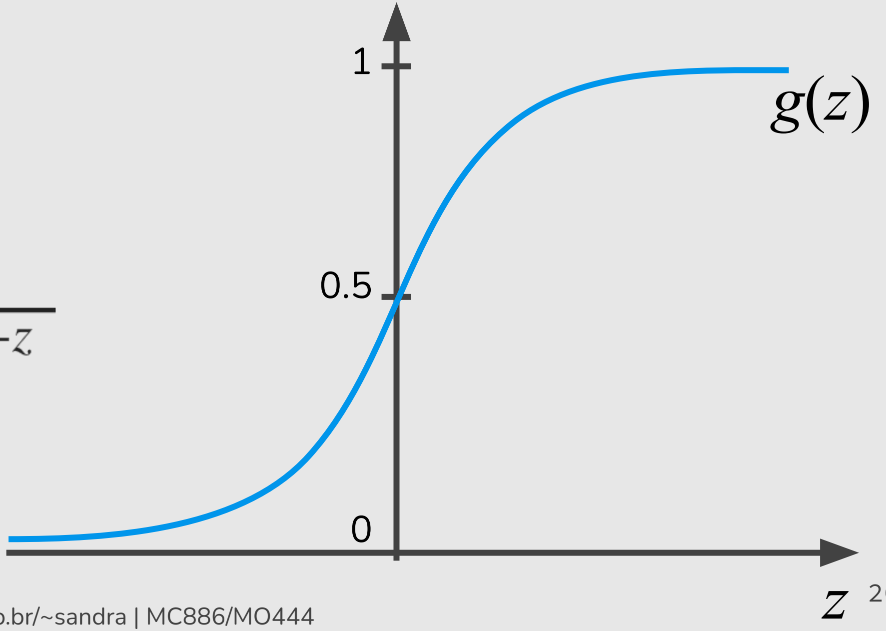
Logistic Regression Model

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

Want $0 \leq h_\theta(x) \leq 1$

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

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



Sigmoid Function

Logistic Function

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$h_\theta(x) = 0.7$

Tell patient that 70% chance of tumor being malignant

"probability that $y = 1$, given x, parameterized by θ"

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

$$P(y = 1 \mid x;\theta) = 1 - P(y = 0 \mid x;\theta)$$

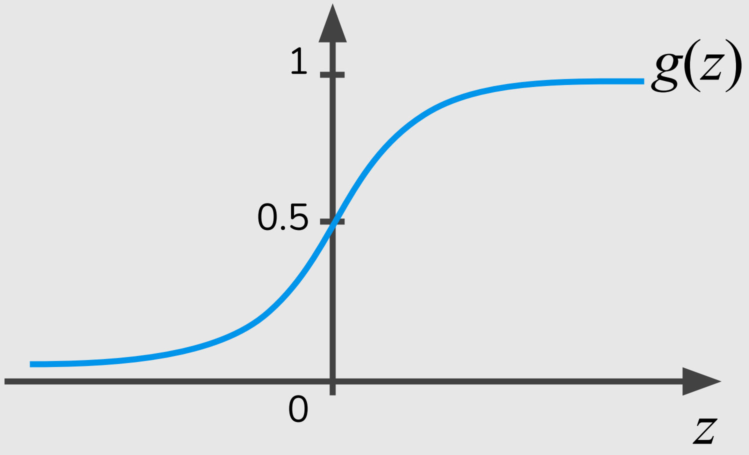
$$h_\theta(x) = P(y = 1 \mid x;\theta)$$

Decision Boundary

Logistic Regression

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

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



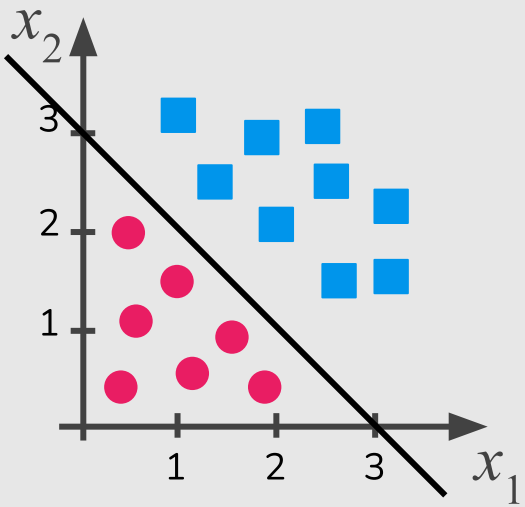
Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$

predict "$y = 0$" if $h_\theta(x) < 0.5$

$g(z) \geq 0.5$ when $z \geq 0$

$g(z) < 0.5$ when $z < 0$

Decision Boundary

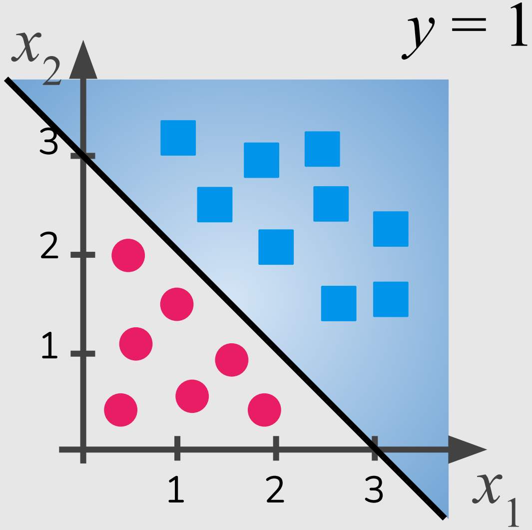


$$\begin{array}{cccc} & -3 & 1 & 1 \\ & \Uparrow & \Uparrow & \Uparrow \\ h_\theta(x) = g(& \theta_0 + & \theta_1 x_1 + & \theta_2 x_2) \end{array}$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$$x_1 + x_2 \geq 3$$

Decision Boundary



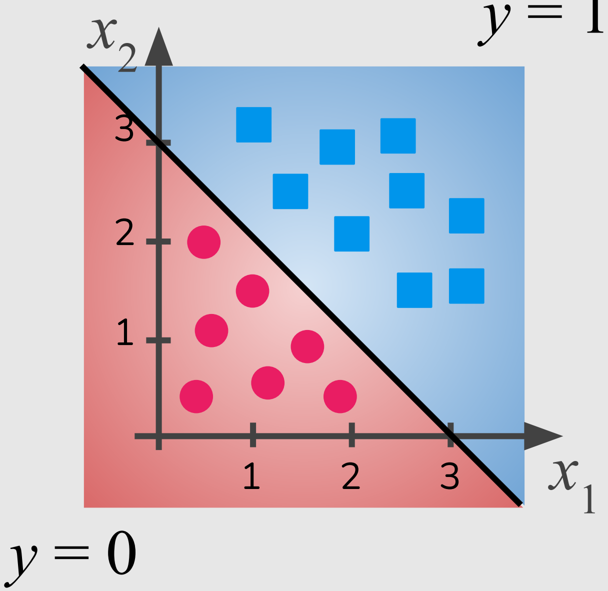
$$\begin{array}{ccc} -3 & 1 & 1 \\ \Uparrow & \Uparrow & \Uparrow \end{array}$$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$$x_1 + x_2 \geq 3$$

Decision Boundary



$$\begin{array}{ccc} -3 & 1 & 1 \\ \Uparrow & \Uparrow & \Uparrow \end{array}$$

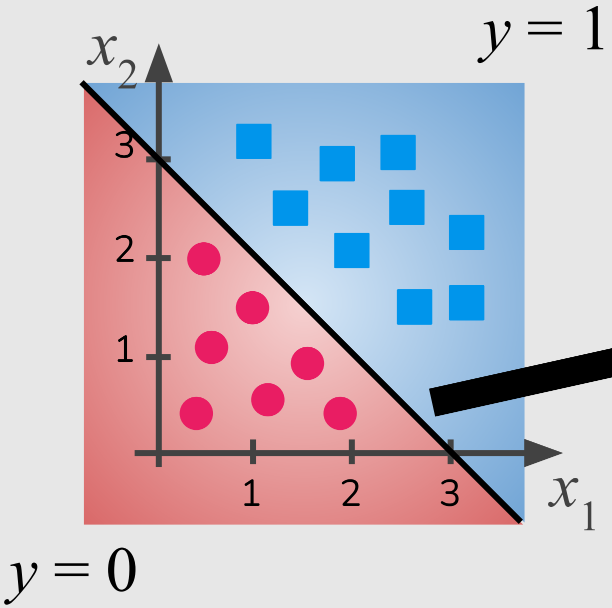
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$x_1 + x_2 \geq 3$

$y = 0,\ x_1 + x_2 < 3$

Decision Boundary



$$-3 \quad 1 \quad 1$$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

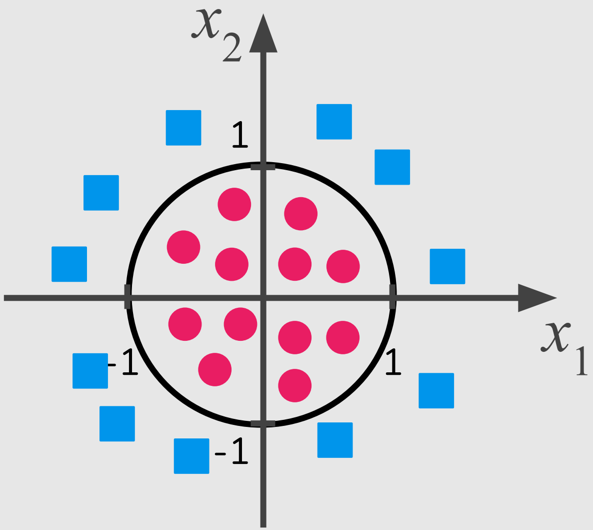
Decision Boundary $x_1 + x_2 = 3$

$$h_\theta(x) = 0.5$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

$$x_1 + x_2 \geq 3 \qquad y = 0,\ x_1 + x_2 < 3$$

Non-linear Decision Boundaries



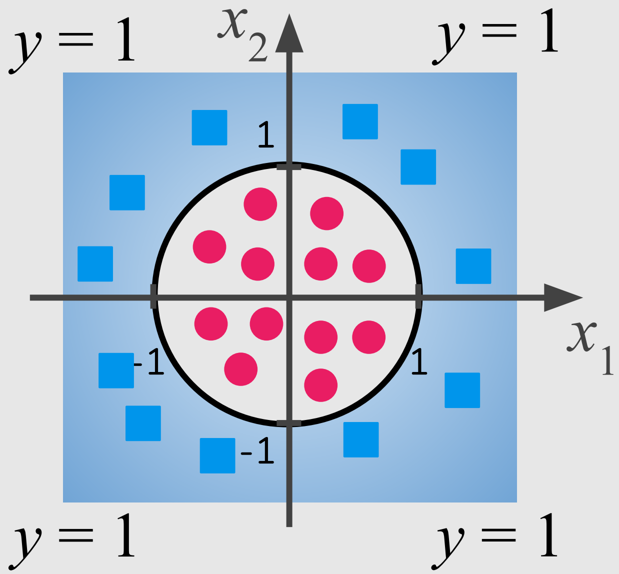
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$-1 \quad 0 \quad 0 \quad 1 \quad 1$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Non-linear Decision Boundaries



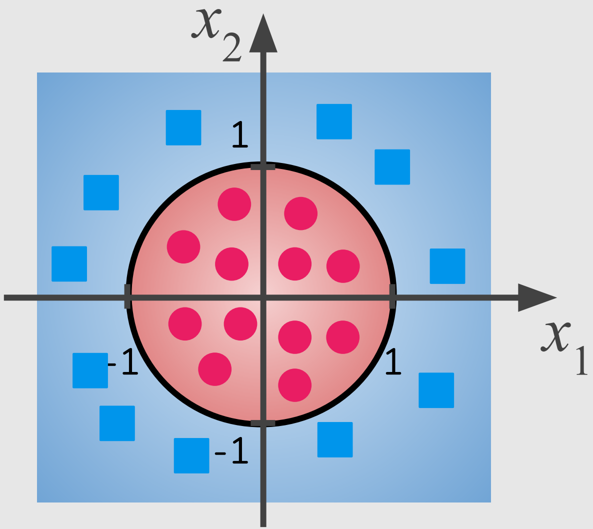
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\theta_0 = -1, \quad \theta_1 = 0, \quad \theta_2 = 0, \quad \theta_3 = 1, \quad \theta_4 = 1$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Non-linear Decision Boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\theta_0 = -1,\quad \theta_1 = 0,\quad \theta_2 = 0,\quad \theta_3 = 1,\quad \theta_4 = 1$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

Today's Agenda

- Logistic Regression
    - Classification
    - Hypothesis Representation
    - Decision Boundary
    - **Cost Function**
    - Simplified Cost Function and Gradient
    - Multiclass Classification

Cost Function

Training set: $\{(x^{(1)},y^{(1)}),\ (x^{(2)},y^{(2)}),\ \ldots,\ (x^{(m)},y^{(m)})\}$

$$h_\theta(x) = \frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}x}} \qquad x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad x_0 = 1,\ y \in \{0,1\}$$

How to choose parameters θ ?

Cost Function

Linear regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2$

Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Linear regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \boxed{\frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cost Function

$$\text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Linear regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \boxed{\frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$

$$\text{Cost}(h_\theta(x), y) = \frac{1}{2} \left(h_\theta(x) - y\right)^2$$

Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

~~Linear~~ regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \boxed{\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}$

Logistic

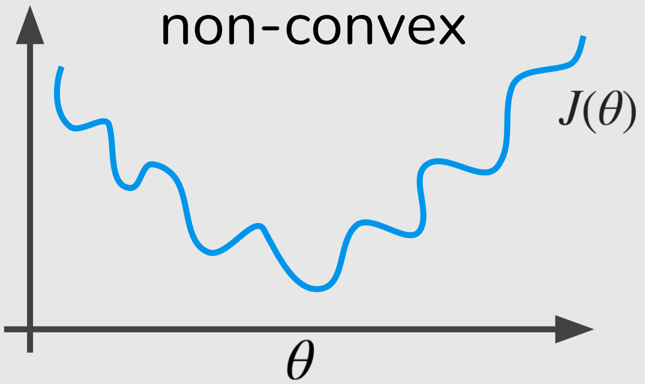
$$\mathrm{Cost}(h_\theta(x),\ y) = \frac{1}{2}(h_\theta(x) - y)^2 \qquad h_\theta(x) = \frac{1}{1+e^{-\theta^{T}x}}$$

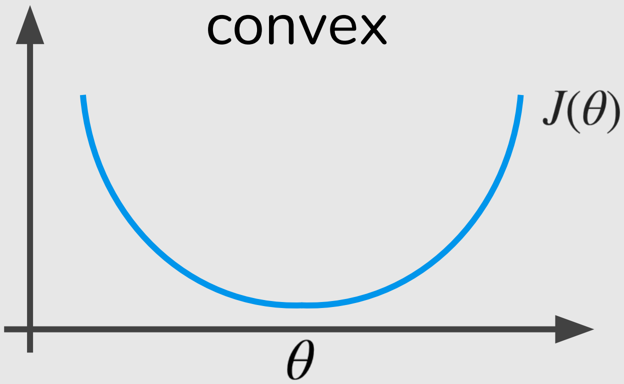
Cost Function

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Logistic regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \boxed{\frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2}$

$$\mathrm{Cost}(h_\theta(x), y) = \frac{1}{2}(h_\theta(x) - y)^2 \qquad h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$





Derivative of Logistic Function

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

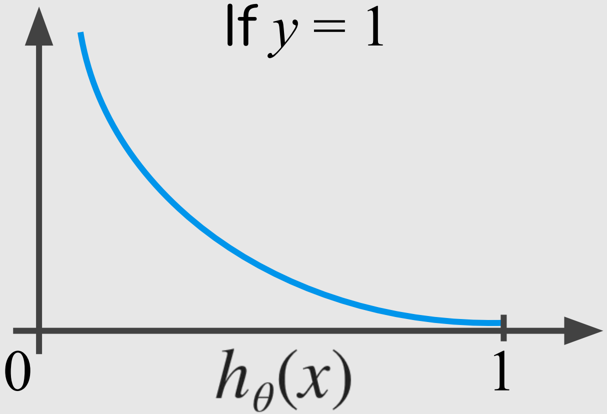
$$\begin{aligned}
g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= \frac{0\cdot(1+e^{-z}) - 1\cdot(-e^{-z})}{(1+e^{-z})^2} \qquad \text{(quotient rule)} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \left(\frac{1}{1+e^{-z}}\right)\left(1 - \frac{1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}$$

Logistic Regression Cost Function

$$\text{Cost}(h_\theta(x),\ y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Logistic Regression Cost Function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



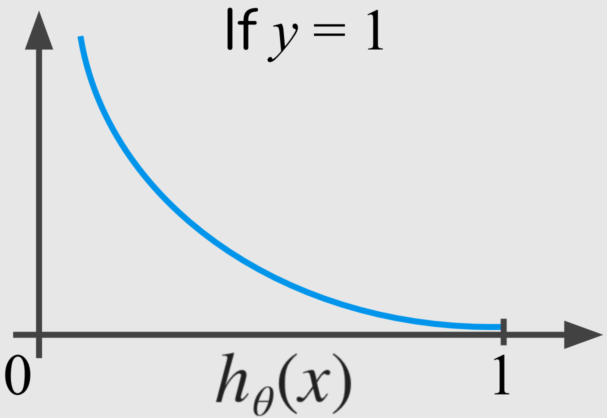
$\mathrm{Cost} = 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \to 0$

$\mathrm{Cost} \to \infty$

Logistic Regression Cost Function

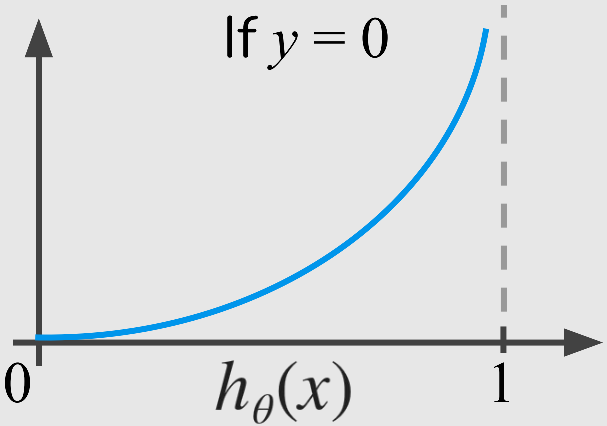
$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1 \mid x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Logistic Regression Cost Function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Simplified Cost Function and Gradient Descent

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x),\ y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\mathrm{Cost}(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\log(1 - h_\theta(x))$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\mathrm{Cost}(h_\theta(x), y) = -y\log(h_\theta(x)) - (1-y)\cancel{\log(1 - h_\theta(x))}$$

$$y = 1$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$\mathrm{Cost}(h_\theta(x), y) = -y\cancel{\log(h_\theta(x))} - (1-y)\log(1 - h_\theta(x))$$

$$y = 0$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

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

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

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

To fit parameters θ : $\min_{\theta} J(\theta)$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}),\ y^{(i)})$$

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

To fit parameters θ : $\min_{\theta} J(\theta)$

To make a new prediction given new x: Output $h_\theta(x) = \dfrac{1}{1 + e^{-\theta^{\mathrm{T}} x}}$

Gradient Descent

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

Want $\min_\theta J(\theta)$:

repeat {

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Gradient Descent

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

Want $\min_\theta J(\theta)$:

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

repeat {

$$\theta_j := \theta_j - \alpha \boxed{\frac{\partial}{\partial\theta_j} J(\theta)}$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Gradient Descent

https://math.stackexchange.com/questions/477207/derivative-of-cost-function-for-logistic-regression

Want $\min_{\theta} J(\theta)$:

repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

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

Gradient Descent

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

Want $\min_\theta J(\theta)$:

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

repeat {

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

} (simultaneously update θ_j for $j = 0, 1, \ldots, n$)

Algorithm looks identical to linear regression!

Multiclass Classification: One-vs-all

Classification

Email tagging: Work, Friends, Family

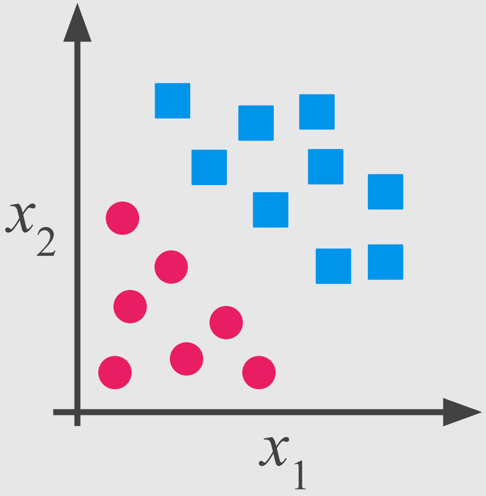
$y = 1$ $y = 2$ $y = 3$

Skin Lesion: Melanoma, Carcinoma, Nevus, Keratosis

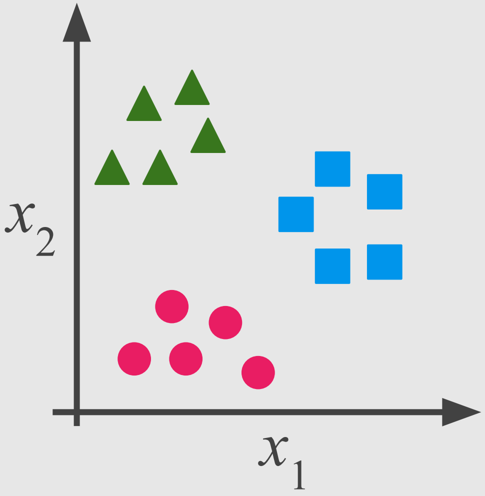
$y = 1$ $y = 2$ $y = 3$ $y = 4$

Video: Pornography, Violence, Gore scenes, Child abuse

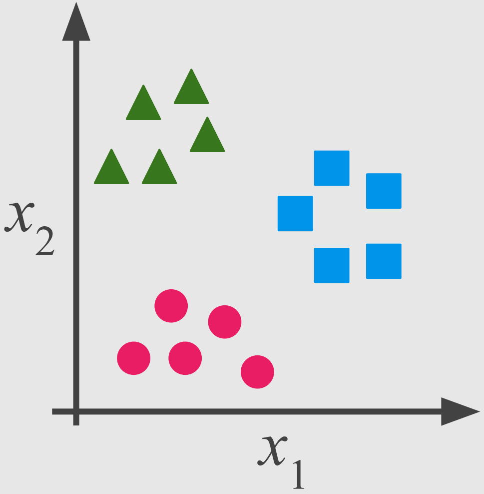
Binary Classification

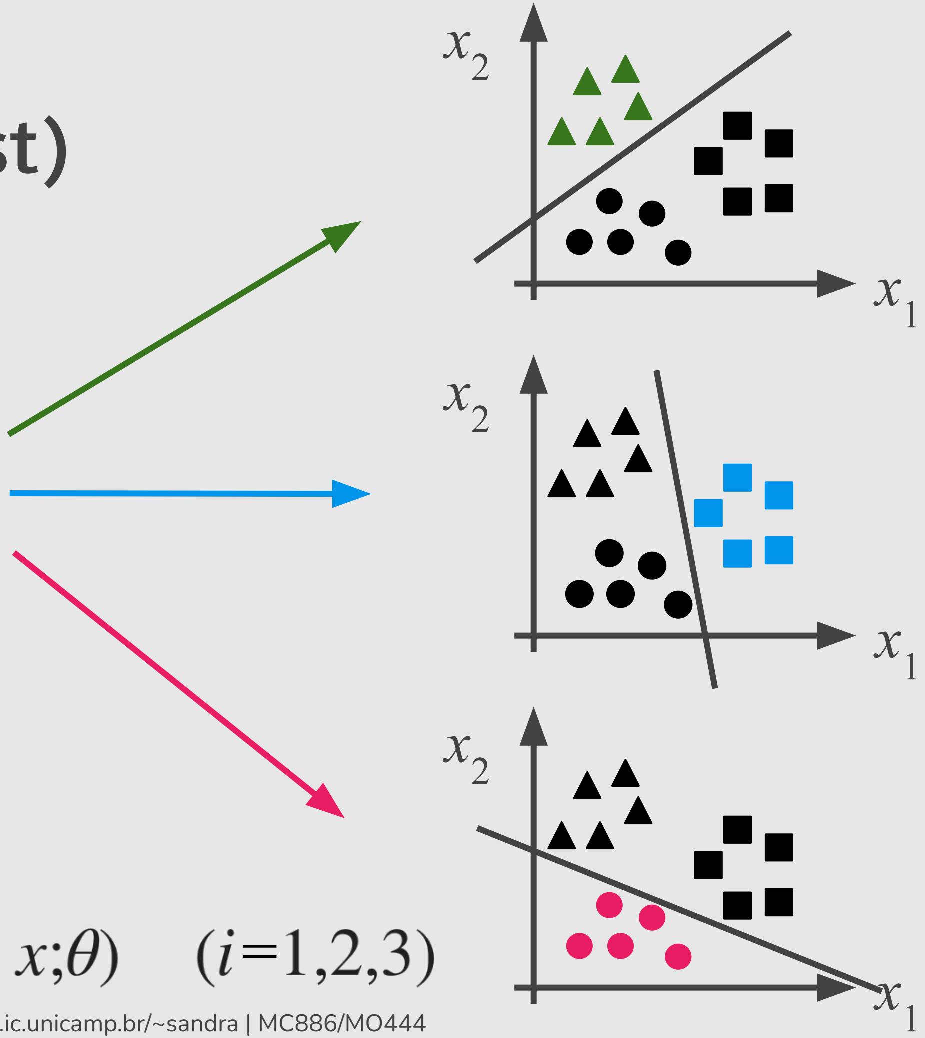


Multi-class Classification



One-vs-All (One-vs-Rest)





Class 1: ▲

Class 2: ■

Class 3: ●

$$h_\theta^{(i)}(x) = P(y = i \mid x;\theta) \quad (i=1,2,3)$$

One-vs-All (One-vs-Rest)

Train a logistic regression classifier $h_\theta^{(i)}(x)$ for each class i to predict the probability that $y = i$.

One a new input x, to make a prediction, pick the class i that maximizes

$$\max_i \; h_\theta^{(i)}(x)$$

References

Machine Learning Books

- Hands-On Machine Learning with Scikit-Learn and TensorFlow, Chap. 2 & 4
- Pattern Recognition and Machine Learning, Chap. 4.3
- Probabilistic Machine Learning: An Introduction, Chap. 10

Machine Learning Courses

- https://www.coursera.org/learn/machine-learning, Week 3
- http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes1.pdf