https://ttpoll.com/p/911058

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Lecture 08: Logistic Regression

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Class Announcements

Homework:
Homework #2 is due this Sunday.

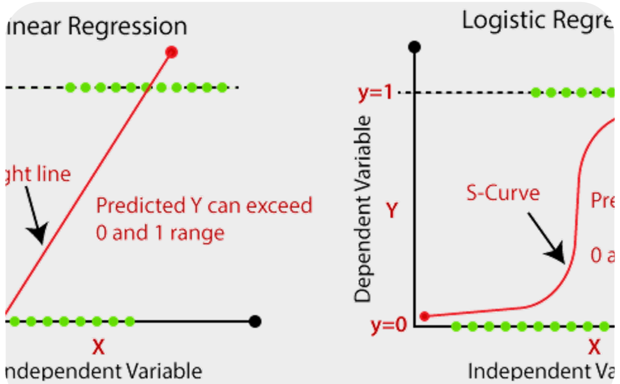
Course Project:
Check groups in Canvas.
PRFAQ is due 09/19
Check Additional Approved Datasets in the Course Project Assignment pane.

Lectures:
Absences: In your email’s subject, include the following text “[COSC325 ABSENCE]”

Exams:
Exam #1: Thursday, 10/03
Exam #2: Thursday, 11/21

Today’s Topics

Logistic Regression



Last Lecture

- Linear Regression
  - Close-form vs GD optimization
  - Parameter confidence intervals and hypothesis testing
  - Extensions
    - Interactions
    - Polynomial regression
  - Manipulation of input features while learning technique remains the same

What: UTK Machine Learning Club

Where: **MK 525**

When: **Tuesday** at **5:00** (including today)

Who: Any experience level

Everyone is welcome to the first meeting of ML club today. Whether you are a beginner looking to learn from our intro to ML lesson series, experienced practitioner who wants to learn from and discuss with other enthusiasts in our reading groups, or you just want to hear from our industry guest speakers and seminars, utkML can help you scratch your machine learning itch!

Linear Regression Wrap up

Exact Solution vs. Gradient Descent

$$\theta = (X^T X)^{-1} X^T y$$

- This gives an exact solution (modulo numerical inaccuracy for inverting the matrix)
- Gradient descent gives you progressively better solutions and eventually gets to an optimum

Exact Solution vs. Gradient Descent

- GD Solution: $O(n * m)$
- Close solution: $O(m^3 + n * m^2)$
- Guidance:
  - Typically $n > m$
  - Will you need to run more than m iterations of gradient descent?
    - Yes? Close form solution may be faster
    - No? Gradient descent may be faster
  - For m $\leq$ 100, it's probably faster to do a closed form solution
  - For m $\geq$ 10000, it's probably faster to do gradient descent
  - For in between...it's unclear

Matrix Design

- We have an input matrix X with shape (n, m)
- We want to fit a polynomial of degree d
- Polynomial feature extraction process
  - Columns for each feature polynomial power (e.g., x_1^3, x_4^5)
  - Plus, columns for each feature interaction up to $d-1$(e.g., $x_1 x_4, x_1^2 x_4$)
- Example for data with n samples, $\mathrm{m} = 2$ features, and polynomial degree $d = 3$.

$$X_{new} = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_1 x_2 \quad x_2^2 \quad x_1^3 \quad x_1^2 x_2 \quad x_1 x_2^2 \quad x_2^3]$$

- Then, apply linear regression algorithm on X_{new}

Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

Notebook Time

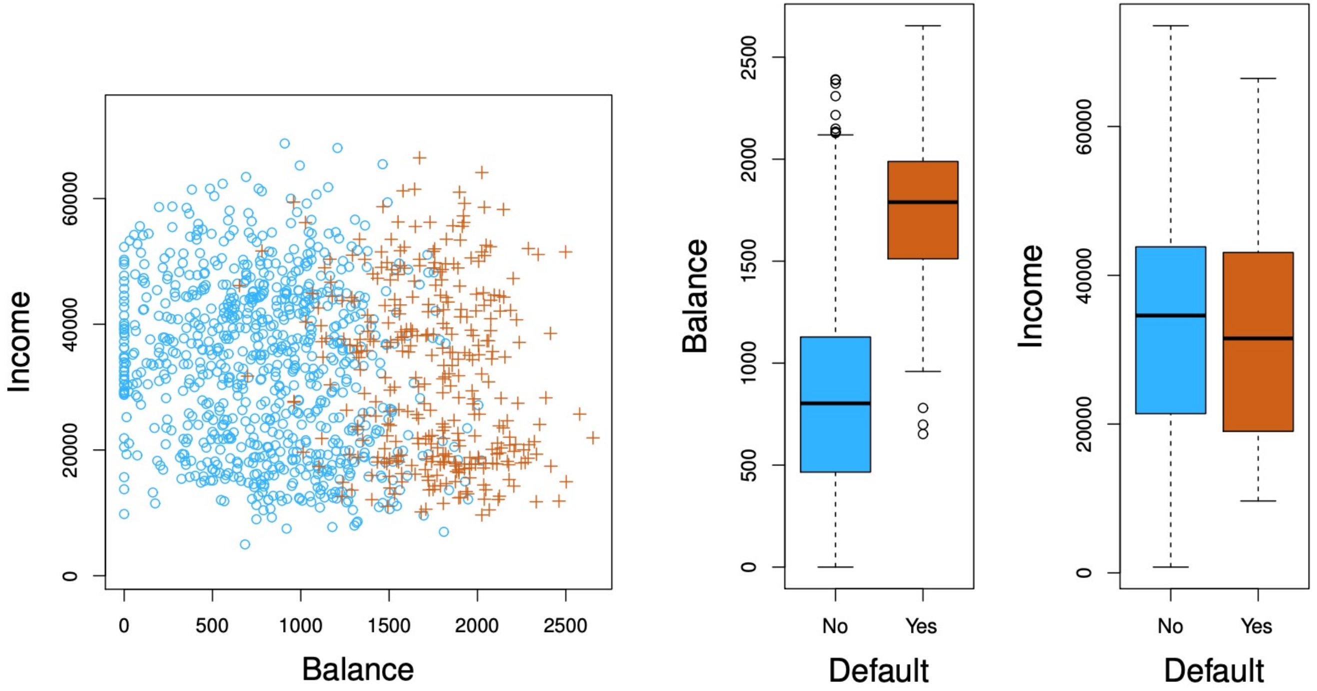
Classification Problems

- Qualitative variables take values in an unordered set C, such as:
  - $\boldsymbol{eye\_color} \in \{brown, blue, green\}$
  - $\boldsymbol{email} \in \{spam, ham\}$.
- Given a feature vector X and a qualitative response y taking values in the discrete set C, the classification task is to build a function $\mathrm{h}(X)$ that takes as input the feature vector X and predicts the value for $\hat{y}$; i.e. $h(X) \in C$.
- Often, we are more interested in estimating the ***probabilities*** that X belongs to each category in C

Examples

Fraud: It is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

Credit card default:



Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

Can we use Linear Regression?

- Suppose for the ***Default*** classification task that we code

$$y = \begin{cases} 1, & if\ Yes \\ 0, & if\ NO \end{cases}$$

- Can we simply perform a linear regression of y on X and classify as Yes if $\hat{y} \geq 0.5$?

Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

Issues with Linear Regression for classification

- For balanced binary classification problems, linear regression is a good classifier.
- Since in the population $E(y\,|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, linear regression might produce probabilities less than zero or bigger than one.

Slide Credit: James, et. al., An Intro to Statistical Learning, 2023.

Issues with Linear Regression for classification

- Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$y = \begin{cases} 1 & if \quad stroke \\ 2 & if \quad drug\ overdose \\ 3 & if \quad epileptic\ seizure \end{cases}$$

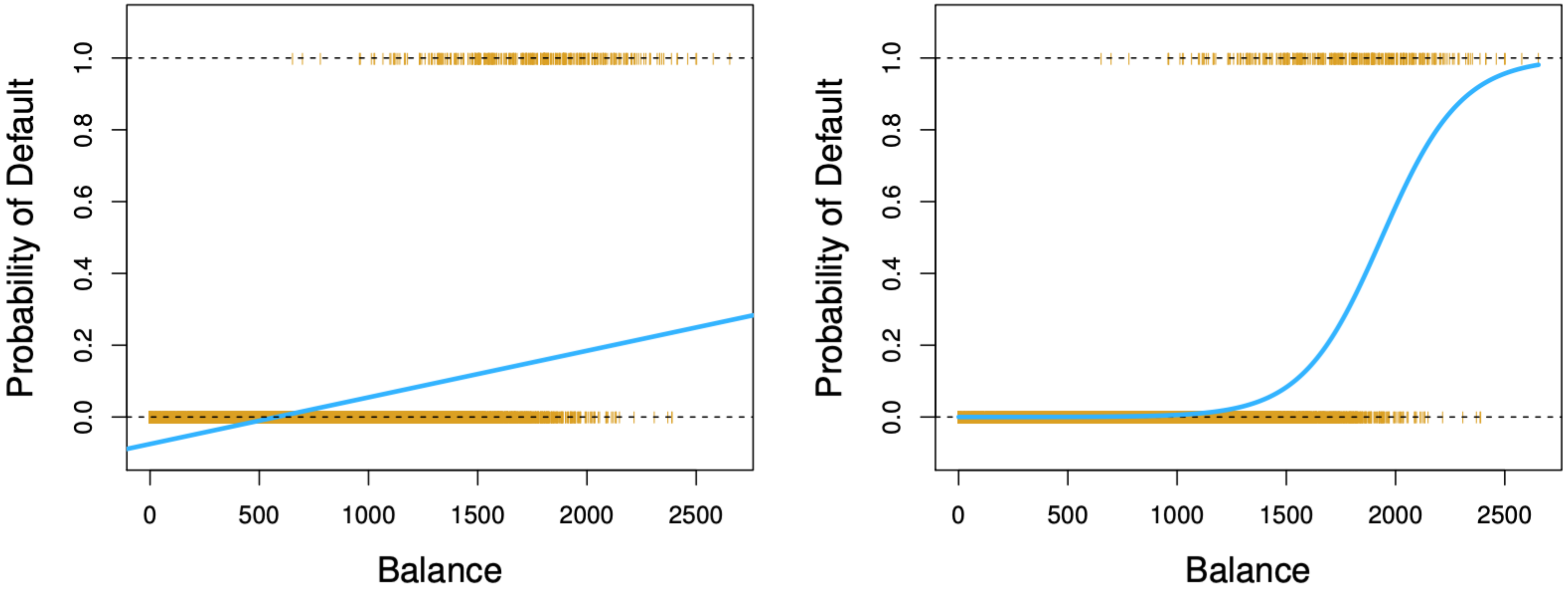
- Any issues with this coding?
  - Suggests an ordering
  - Implies that the difference between ***stroke*** and ***drug overdose*** is the same as between ***drug overdose*** and ***epileptic seizure***.

What do we need?

- A probability
  - Bounded between zero and one
- Categorize beyond binary problems
- Categorize unordered labels

- ***Logistic regression*** is our candidate.

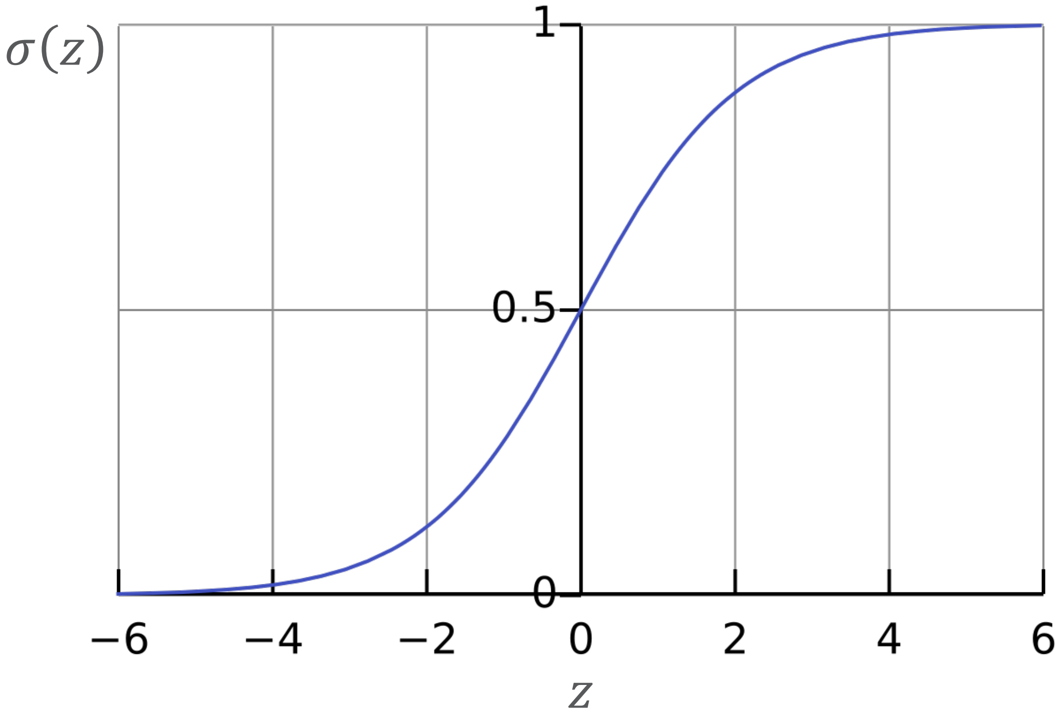
Linear vs Logistic Regression



- The orange marks indicate the response $\hat{y}$, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

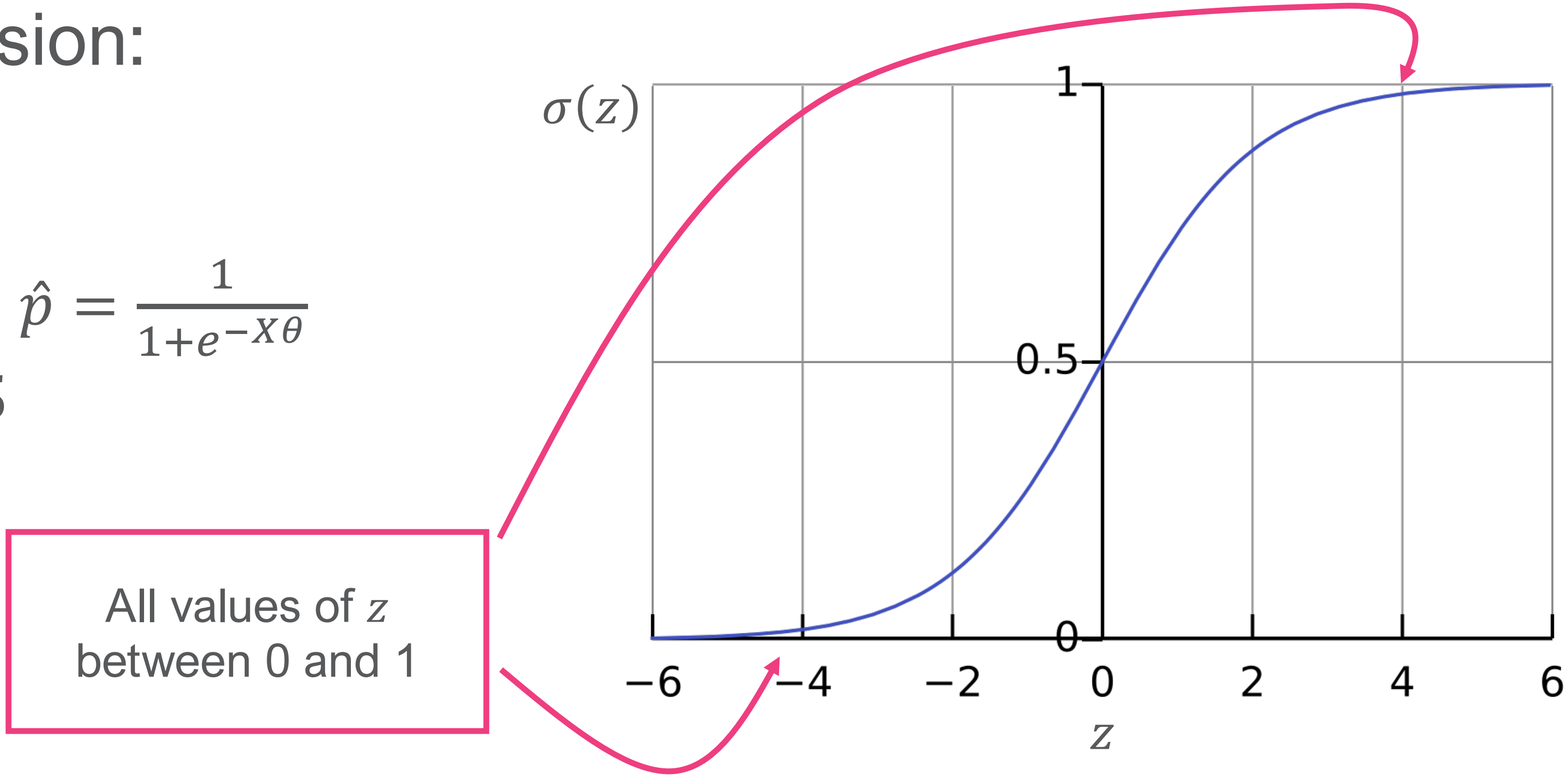
Logistic Regression

- Linear regression: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m = X\theta$
- Logistic regression:
  - $z = X\theta$
  - $\hat{p} = \sigma(z)$
  - $\sigma(z) = \frac{1}{1+e^{-z}} \Rightarrow \hat{p} = \frac{1}{1+e^{-X\theta}}$
  - $\hat{y} = \begin{cases} 1, \hat{p} \geq 0.5 \\ 0, \hat{p} < 0.5 \end{cases}$



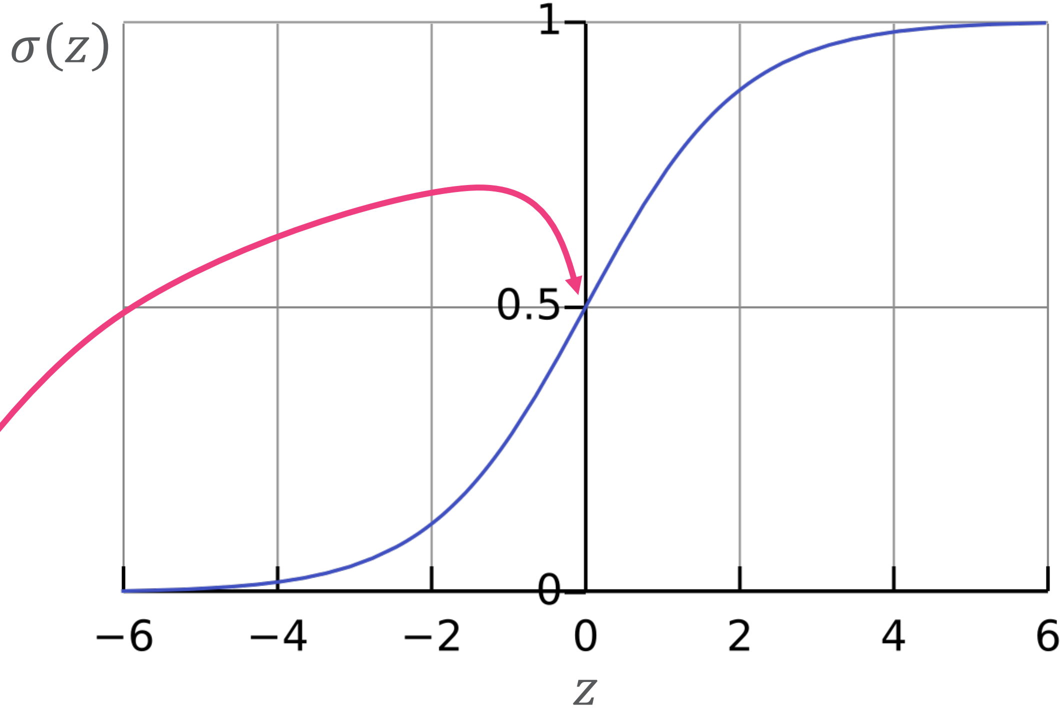
How do we compute this probability?

- Linear regression: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m = X\theta$
- Logistic regression:
  - $\boldsymbol{z} = \boldsymbol{X\theta} \in \mathbb{R}$
  - $\hat{p} = \sigma(z)$
  - $\sigma(z) = \frac{1}{1+e^{-z}} \Rightarrow \hat{p} = \frac{1}{1+e^{-X\theta}}$
  - $\hat{y} = \begin{cases} 1, \hat{p} \geq 0.5 \\ 0, \hat{p} < 0.5 \end{cases}$



How do we compute this probability?

- Linear regression: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m = X\theta$
- Logistic regression:
  - $z = X\theta$
  - $\hat{p} = \sigma(z)$
  - $\sigma(z) = \frac{1}{1+e^{-z}} \Rightarrow \hat{p} = \frac{1}{1+e^{-X\theta}}$
  - $\hat{y} = \begin{cases} 1, \hat{p} \geq 0.5 \\ 0, \hat{p} < 0.5 \end{cases}$

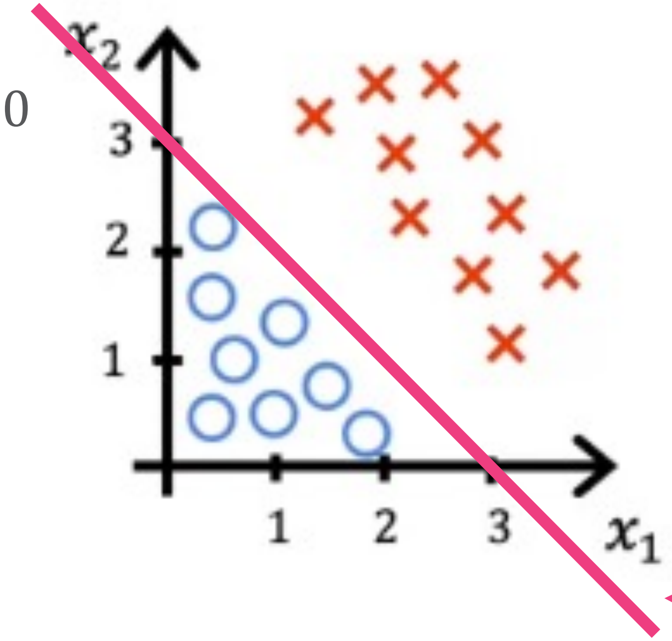


Notice model predicts 1 when $X\theta$ is positive.

Geometry of Logistic Regression

$\hat{y} = 1$

$\hat{y} = 0$

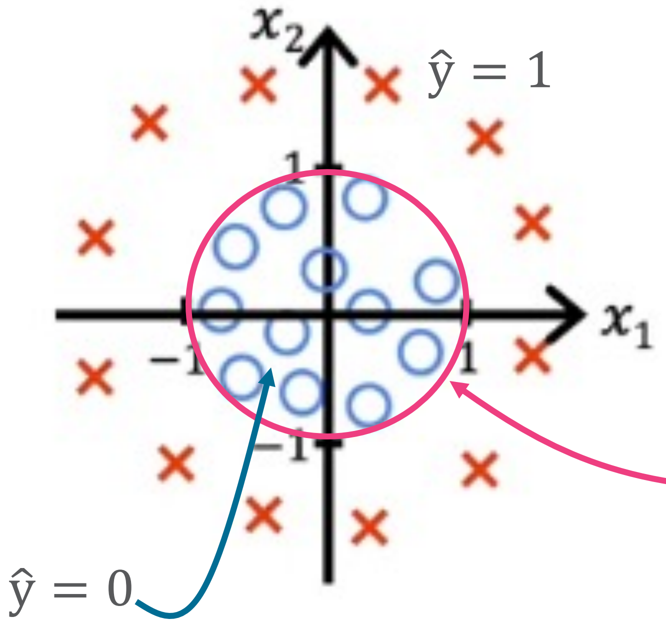


$$\sigma(z) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$
$$= \sigma(-3 + (1)x_1 + (1)x_2)$$

$\hat{y} = 1$:

$$z = -3 + x_1 + x_2 = 0 \Rightarrow$$
$$x_1 + x_2 = 3$$

Geometry of Logistic Regression



$$\sigma(z) = \sigma(\theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2)$$
$$= \sigma(-1 + (1)x_1^2 + (1)x_2^2)$$

$\hat{y} = 1$:

$$z = -1 + x_1^2 + x_2^2 = 0 \Rightarrow$$
$$x_1^2 + x_2^2 = 1$$

How do we learn the parameters?

Gradient Descent

Loss and Cost Functions

- Loss $\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right)$ is the error between the ground truth (i.e., expected response) $y^{(i)}$ and the model prediction $\hat{y}^{(i)}$.
- Cost $J(w, b)$ is a measure of expected model error for parameters w and b.

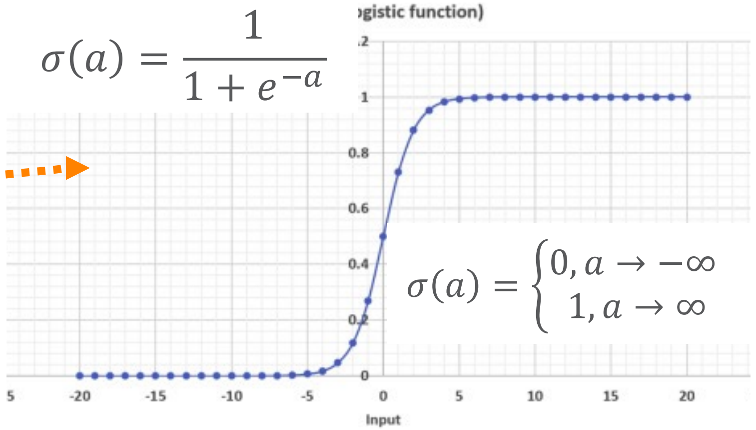
Per sample $x^{(i)}$

Expected Performance

We want both to be SMALL

Loss Function

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

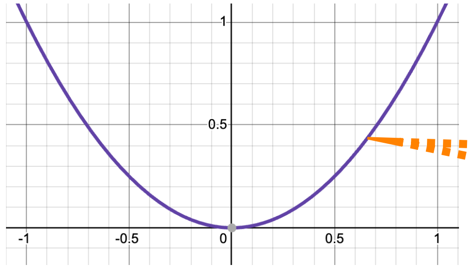


- Logistic regression model

$$\hat{y}^{(i)} = \sigma\left(x^{(i)T} w + b\right), \quad \text{where } \sigma\left(z^{(i)}\right) = \frac{1}{1+e^{-z^{(i)}}}$$

$$z^{(i)} = x^{(i)T} w + b$$

- Loss function



Ideal convex loss function

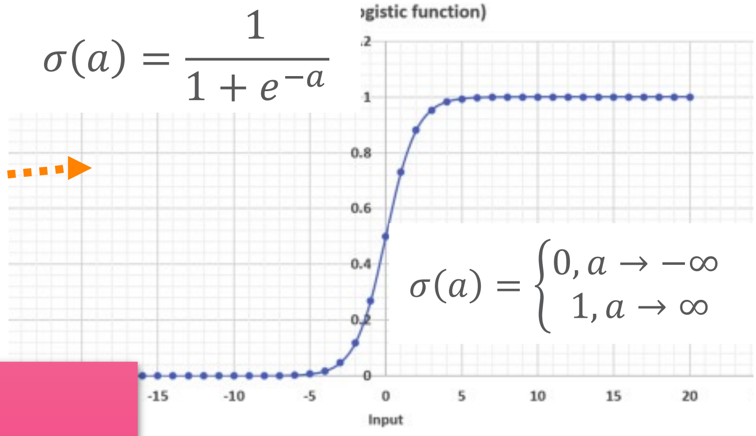
$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2 \qquad ? \qquad \mathcal{L}(\hat{y}, y) = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

Non-convex for logistic regression

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Loss Function

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



$$\sigma(a) = \begin{cases} 0, a \to -\infty \\ 1, a \to \infty \end{cases}$$

- Logistic regression model

$$\hat{y}^{(i)} = \sigma(x^{(i)T}w + b),$$

$$z^{(i)} = x^{(i)T}w + b$$

Binary Cross-Entropy Loss

- Loss function

Log-Loss

Idea vex loss
ction

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$$

?

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

Non-convex for logistic regression

THE UNIVERSITY OF TENNESSEE KNOXVILLE

https://ttpoll.com/r…58

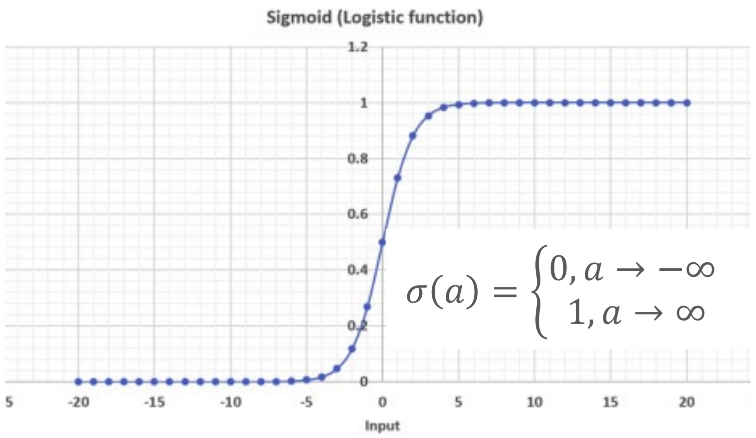
Graduate Student Alert

Loss function derivation

$$\hat{y}^{(i)} = \sigma(x^{(i)T}w + b), \qquad \text{where } \sigma(z^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}}$$

Pick one hypothesis!

We want our model to output: $\hat{y} = p(y = 1|x)$



If y = 1: $p(y|x) = \hat{y}$

If y = 0: $p(y|x) = 1 - \hat{y}$

Loss function derivation

If y = 1: $p(y|x) = \hat{y}$

If y = 0: $p(y|x) = 1 - \hat{y}$

Equivalent

$$p(y|x) = \hat{y}^{y}(1-\hat{y})^{(1-y)}$$

Correct!

If y = 1, then $\hat{y}^{1}(1-\hat{y})^{(1-1)} = \hat{y}$

If y = 0, then $\hat{y}^{0}(1-\hat{y})^{(1-0)} = 1-\hat{y}$

Log properties:

$\log(a \times b) = \log a + \log b$

$\log(a^{b}) = b \times \log(a)$

$$\log(p(y|x)) = \log\left(\hat{y}^{y}(1-\hat{y})^{(1-y)}\right) = y\log\hat{y} + (1-y)\log(1-\hat{y}) = -\mathcal{L}(\hat{y}, y)$$

Loss function derivation

Correct!

If

If

$= \hat{y}$

$= 1 - \hat{y}$

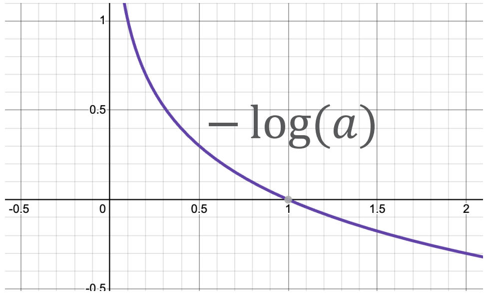
We want to make ***probabilities*** larger, but ***losses*** smaller.

$$\log(p(y|x)) = \log\left(\hat{y}^{y}(1-\hat{y})^{(1-y)}\right) = y\log\hat{y} + (1-y)\log(1-\hat{y}) = -\mathcal{L}(\hat{y}, y)$$

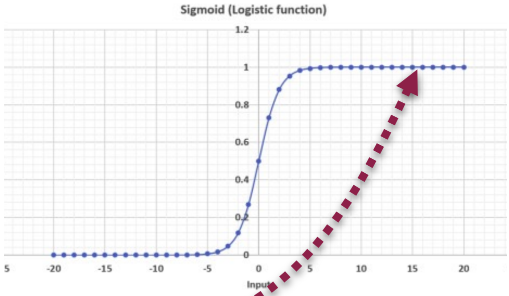
Loss Function Intuition

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

- If $y = 1$: $\mathcal{L}(\hat{y}, 1) = -\big((1)\log(\hat{y}) + (1-1)\log(1-\hat{y})\big) = -\log(\hat{y})$



As $\hat{y} \to 1, \mathcal{L}(\hat{y}, 1) \to 0$

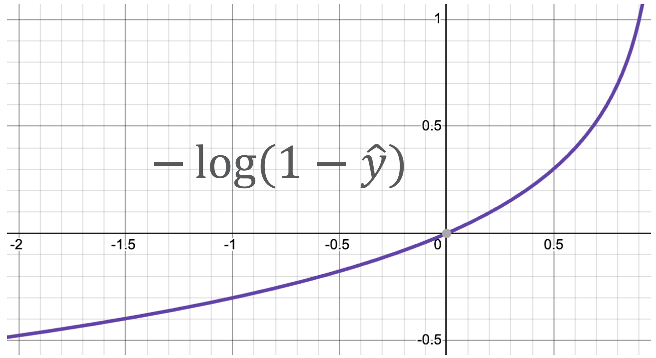


As $\hat{y} \to 0, \mathcal{L}(\hat{y}, 1) \to \infty$

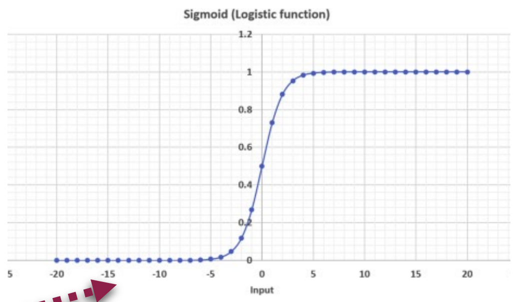
Loss Function Intuition

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

- If $y = 0$: $\mathcal{L}(\hat{y}, 0) = -\big((0)\log(\hat{y}) + (1-0)\log(1-\hat{y})\big) = -\log(1-\hat{y})$



$As\ \hat{y} \to 0, \mathcal{L}(\hat{y}, 0) \to 0$



$As\ \hat{y} \to 1, \mathcal{L}(\hat{y}, 0) \to \infty$

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Cost Function

Computes loss for sample i

- Loss function

$$\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) = -(y^{(i)} \log(\hat{y}^{(i)}) + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right))$$

- Cost function

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

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

Note this is the average over all losses.

Pop Quiz: Question #1

2 | MULTIPLE CHOICE

Why the LogLoss loss is preferred over Mean Squared Error (MSE) loss?

A. MSE is harder to compute

B. LogLoss is sensitive to outliers.

C. LogLoss is a convex function for binary problems.

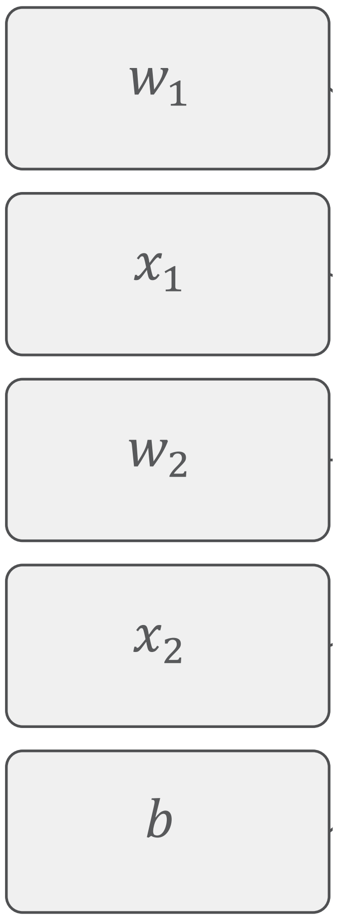
D. The functions are equivalent.

Computing dZ

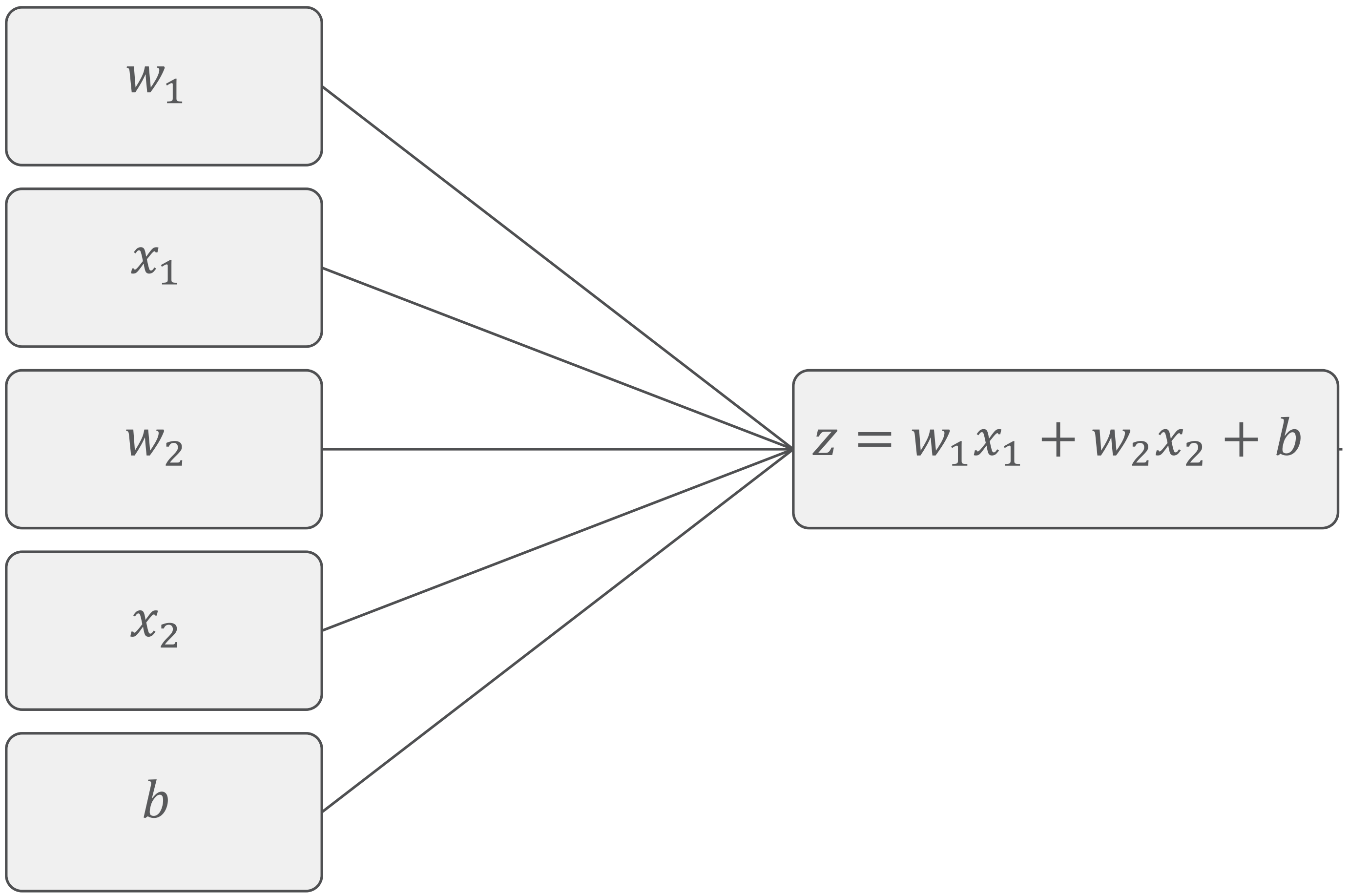
Logistic Regression Comp. Graph

- Equations ($n_x = 2$ and $a = \hat{y}$)
  - Loss function: $\mathcal{L}(\hat{y}, y) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$
  - Model output: $\hat{y} = \sigma(z) \rightarrow a = \sigma(z)$
  - Activation function: $\sigma(z) = \dfrac{1}{1 + e^{-z}}$
  - Input, weights, and bias: $z = x^T w \quad + b = w_1 x_1 + w_2 x_2 + b$

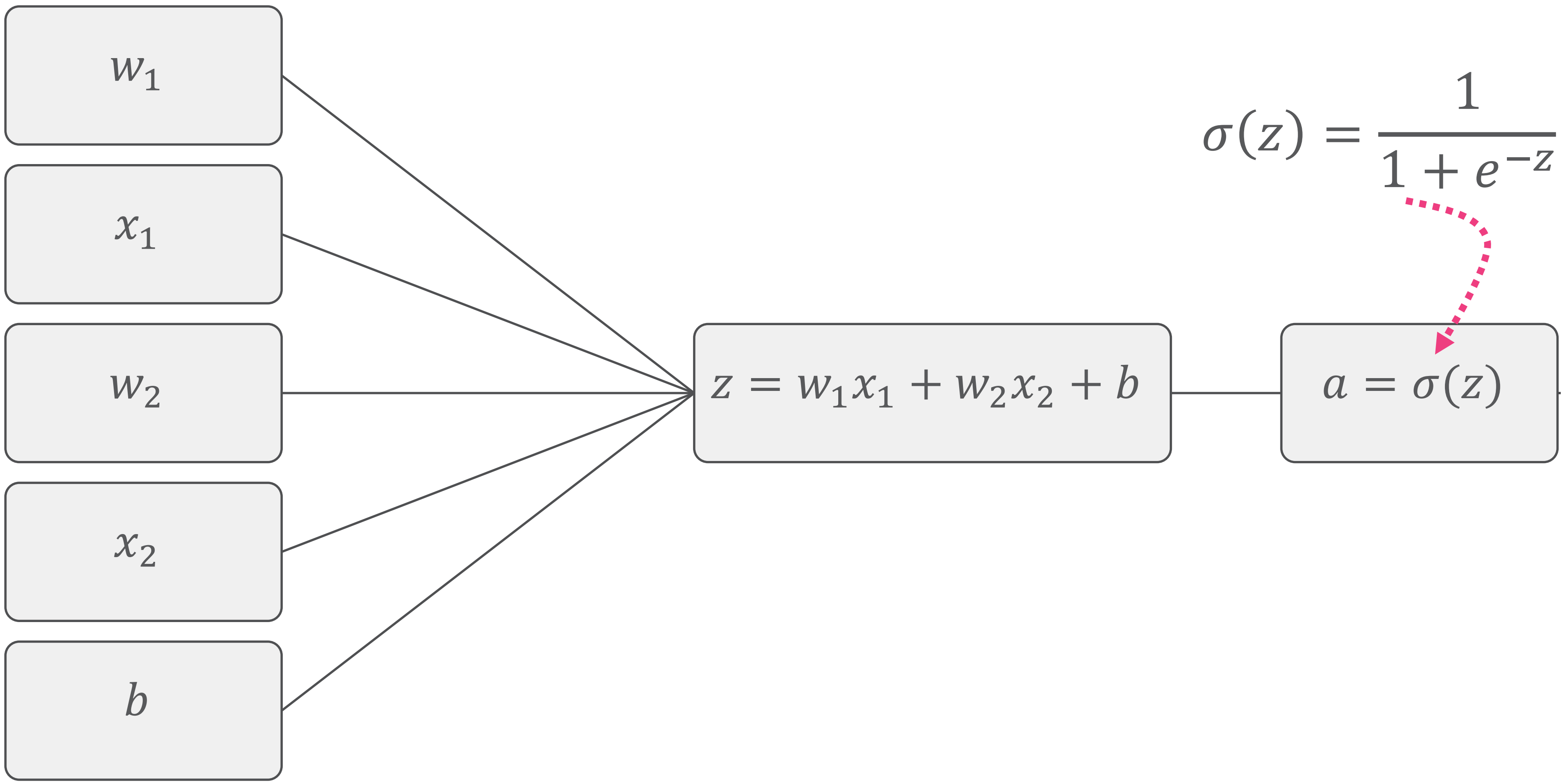
Logistic Regression Comp. Graph



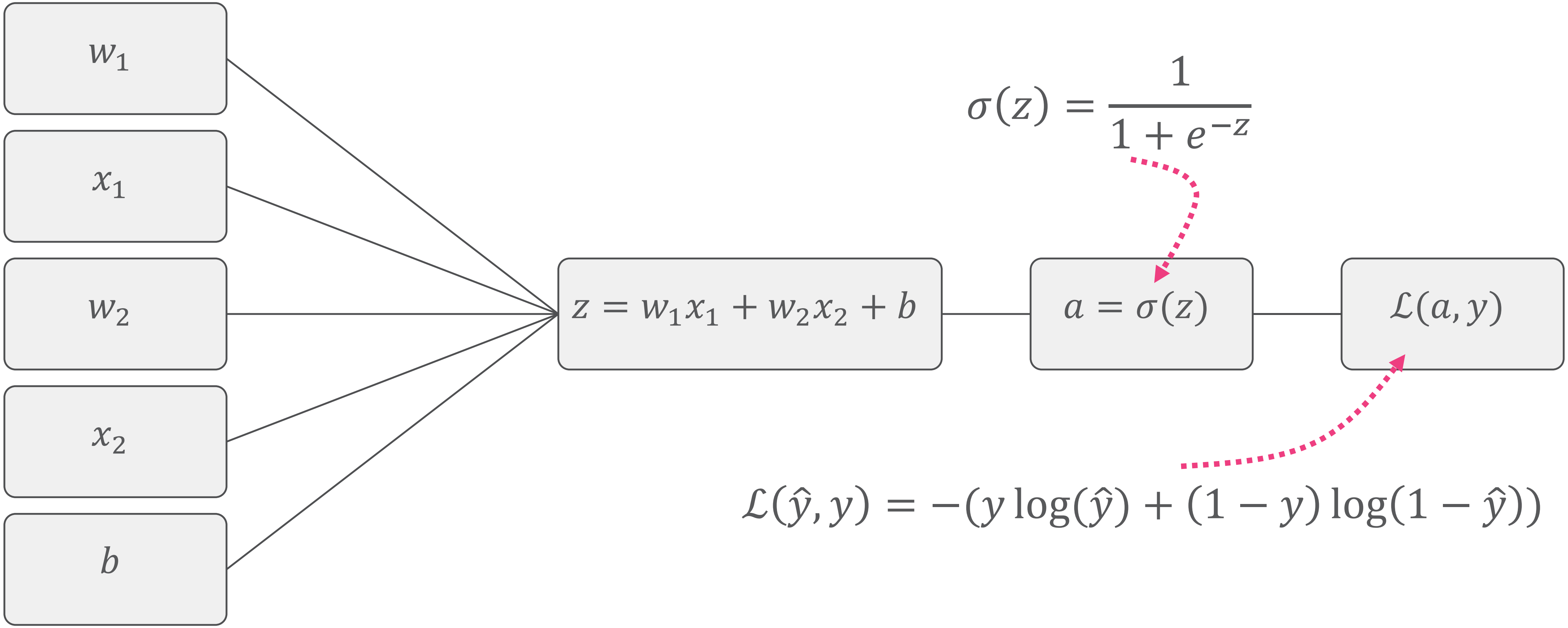
Logistic Regression Comp. Graph



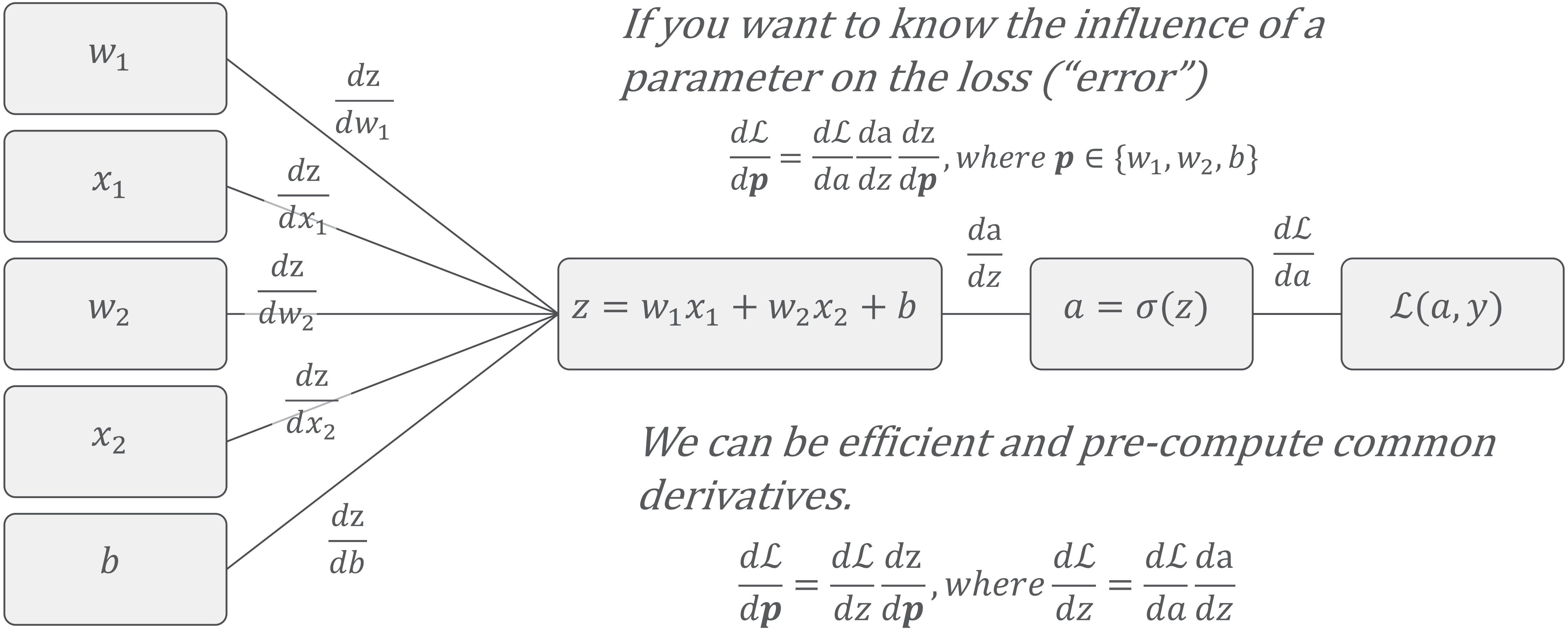
Logistic Regression Comp. Graph



Logistic Regression Comp. Graph



Logistic Regression Comp. Graph



If you want to know the influence of a parameter on the loss ("error")

$$\frac{d\mathcal{L}}{d\boldsymbol{p}} = \frac{d\mathcal{L}}{da}\frac{da}{dz}\frac{dz}{d\boldsymbol{p}}, \text{where } \boldsymbol{p} \in \{w_1, w_2, b\}$$

We can be efficient and pre-compute common derivatives.

$$\frac{d\mathcal{L}}{d\boldsymbol{p}} = \frac{d\mathcal{L}}{dz}\frac{dz}{d\boldsymbol{p}}, \text{where } \frac{d\mathcal{L}}{dz} = \frac{d\mathcal{L}}{da}\frac{da}{dz}$$

https://ttpoll.com/p/911058

$$\frac{d\mathcal{L}}{da} = \frac{a-y}{a(1-a)}$$

Property: $\frac{d(\log(a))}{da} = \frac{1}{a}da$

$$\mathcal{L}(a,y) = -(y\log(a) + (1-y)\log(1-a))$$

$$\frac{d\mathcal{L}(a,y)}{da} = -\left[y\frac{d\log(a)}{da} + (1-y)\frac{d\log(1-a)}{da}\right]$$

$$\frac{d\mathcal{L}}{da} = -\left[y\frac{1}{a} + (1-y)\frac{-1}{1-a}\right]$$

$$\frac{d\mathcal{L}}{da} = -\left[y\frac{1}{a} + \frac{y-1}{1-a}\right] = -\left[\frac{y-ay}{a(1-a)} + \frac{ay-a}{a(1-a)}\right] = -\left[\frac{y-ay+ay-a}{a(1-a)}\right]$$

$$\frac{d\mathcal{L}}{da} = -\left[\frac{y-a}{a(1-a)}\right] = \frac{a-y}{a(1-a)}$$

$\frac{da}{dz}$

$$a = \sigma(z) = \frac{1}{1+e^{-z}} \rightarrow \frac{da}{dz} = \frac{d\left(\frac{1}{1+e^{-z}}\right)}{dz} \rightarrow \begin{matrix} u = 1 + t \\ t = e^{-z} \end{matrix} \rightarrow \frac{da}{dz} = \frac{da}{du}\frac{du}{dt}\frac{dt}{dz}$$

Chain Rule !!!

$$\frac{da}{du} = \frac{d(u^{-1})}{dz} = -\frac{1}{u^2} \qquad \frac{du}{dt} = \frac{d(1+t)}{dt} = 1 \qquad \frac{dt}{dz} = \frac{d(e^{-z})}{dz} = -e^{-z}$$

Power Rule

$$\frac{da}{dz} = \frac{d\sigma}{du}\frac{du}{dt}\frac{dt}{dz} = -\frac{1}{u^2}(1)(-e^{-z}) = \frac{e^{-z}}{u^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$\frac{da}{dz} = a(a-1) = \sigma(z)(1-\sigma(z))$$

$$\frac{da}{dz} = \frac{d\sigma}{du}\frac{du}{dt}\frac{dt}{dz} = -\frac{1}{u^2}(1)(-e^{-z}) = \frac{e^{-z}}{u^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

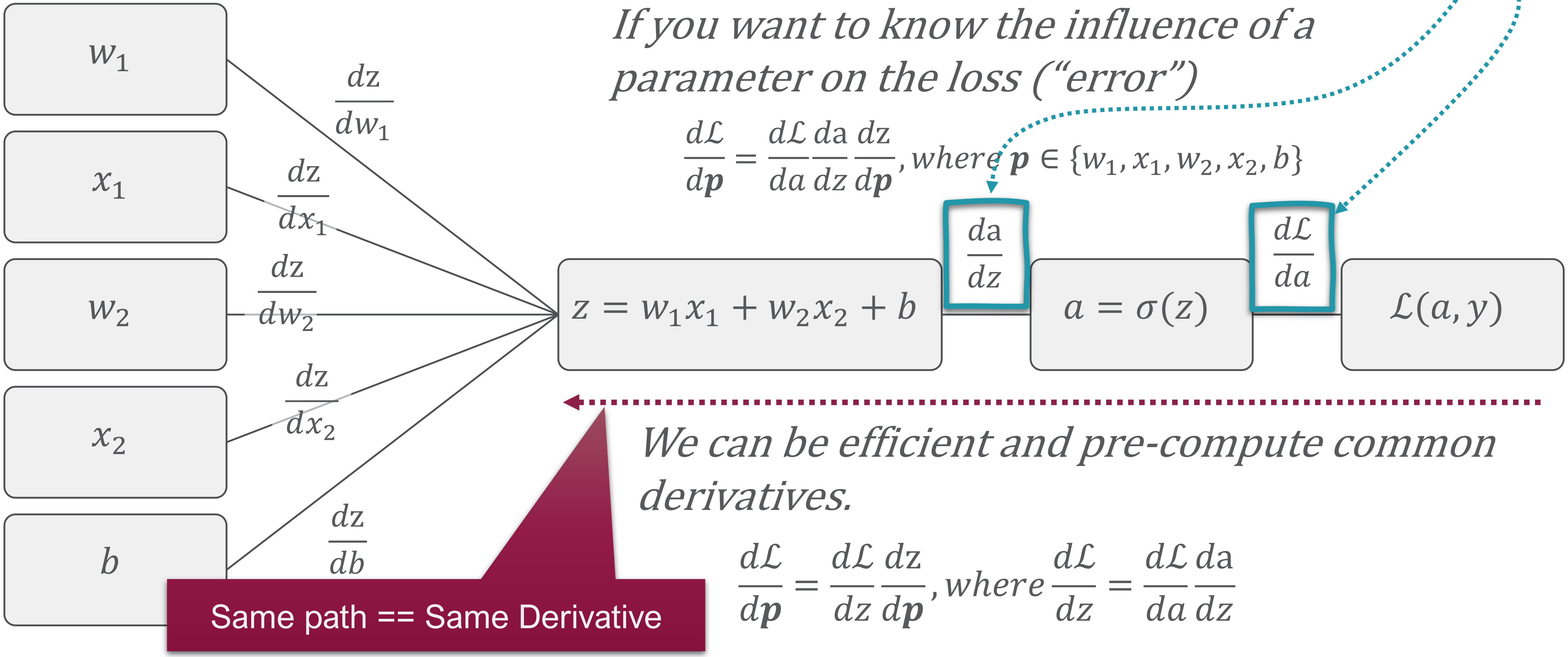
$$= \underbrace{\frac{1}{1+e^{-z}}}_{a}\frac{1+e^{-z}-1}{1+e^{-z}} = a\left(\underbrace{\frac{1+e^{-z}}{1+e^{-z}}}_{1} - \underbrace{\frac{1}{1+e^{-z}}}_{a}\right) = a(1-a) = \sigma(z)(1-\sigma(z))$$

Logistic Regression Comp. Graph

If you want to know the influence of a parameter on the loss ("error")

$$\frac{d\mathcal{L}}{d\mathbf{p}} = \frac{d\mathcal{L}}{da}\frac{da}{dz}\frac{dz}{d\mathbf{p}}, \text{where } \mathbf{p} \in \{w_1, x_1, w_2, x_2, b\}$$

Already computed!



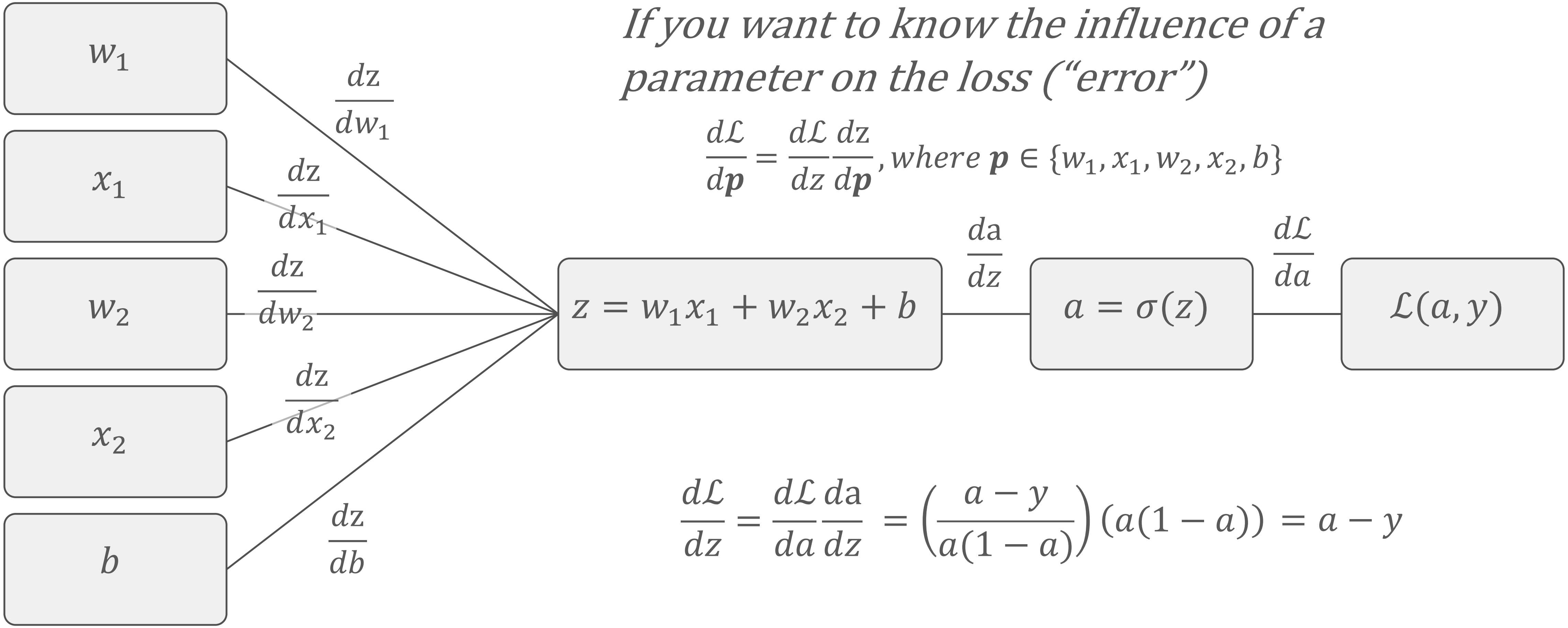
We can be efficient and pre-compute common derivatives.

$$\frac{d\mathcal{L}}{d\mathbf{p}} = \frac{d\mathcal{L}}{dz}\frac{dz}{d\mathbf{p}}, \text{where } \frac{d\mathcal{L}}{dz} = \frac{d\mathcal{L}}{da}\frac{da}{dz}$$

Logistic Regression Comp. Graph

If you want to know the influence of a parameter on the loss ("error")

$$\frac{d\mathcal{L}}{d\boldsymbol{p}} = \frac{d\mathcal{L}}{dz}\frac{dz}{d\boldsymbol{p}}, \textit{where } \boldsymbol{p} \in \{w_1, x_1, w_2, x_2, b\}$$



$$\frac{d\mathcal{L}}{dz} = \frac{d\mathcal{L}}{da}\frac{da}{dz} = \left(\frac{a-y}{a(1-a)}\right)(a(1-a)) = a - y$$

$\frac{dz}{dp}$

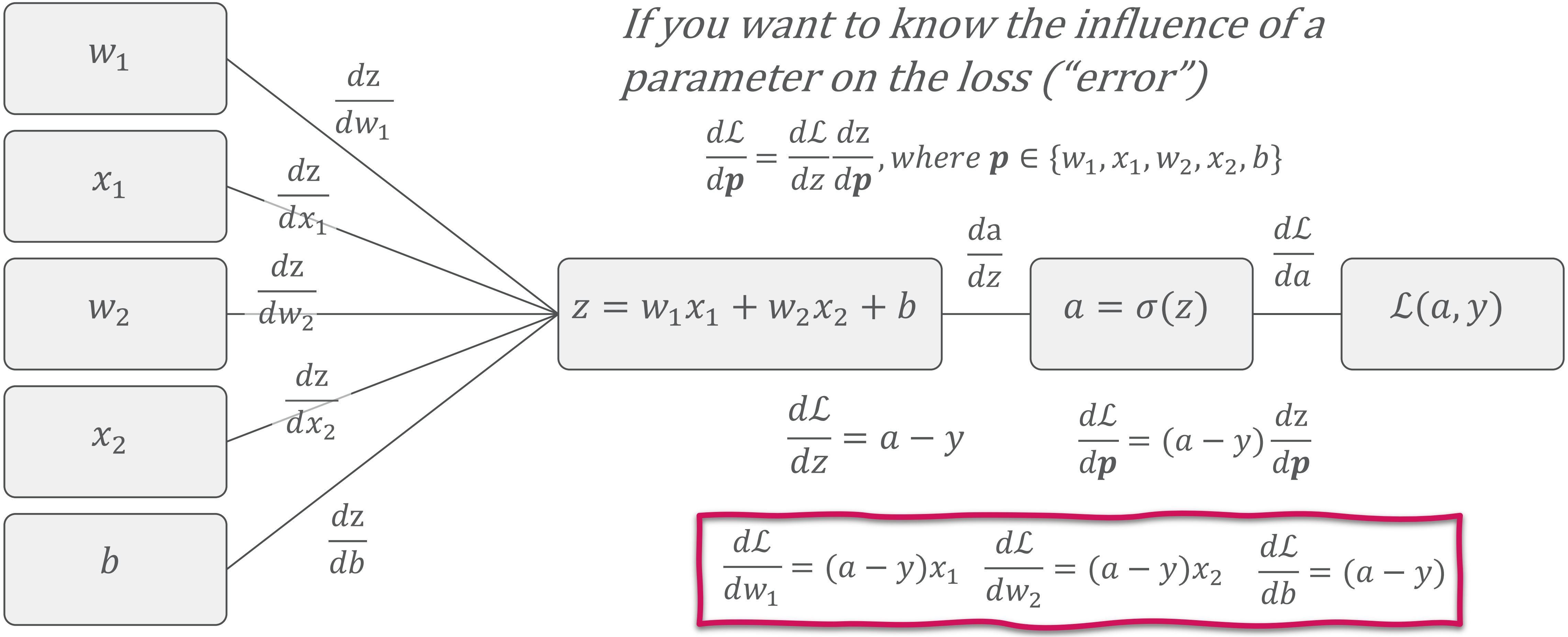
$$z = w_1x_1 + w_2x_2 + b$$

$$\frac{dz}{dw_1} = x_1 \qquad \frac{dz}{dw_1} = x_2 \qquad \frac{dz}{db} = 1$$

Logistic Regression Comp. Graph

If you want to know the influence of a parameter on the loss ("error")

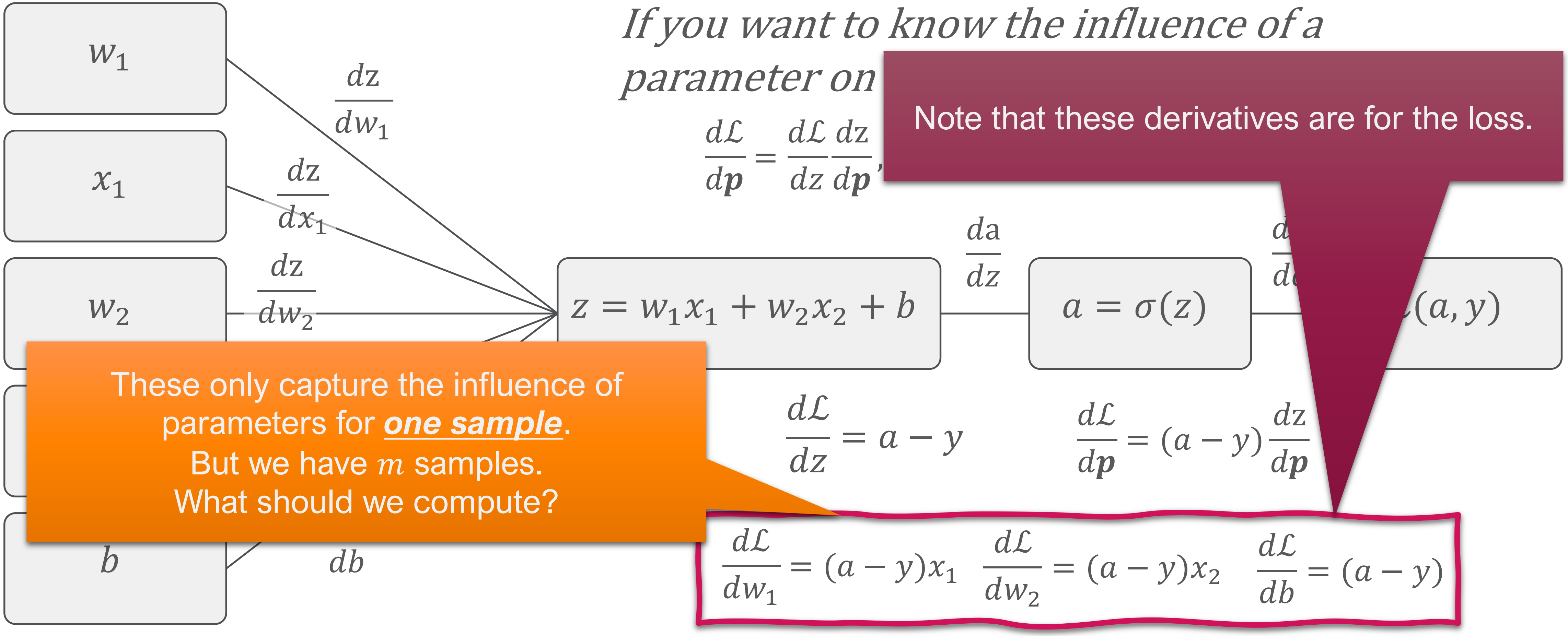
$$\frac{d\mathcal{L}}{d\boldsymbol{p}} = \frac{d\mathcal{L}}{dz}\frac{dz}{d\boldsymbol{p}}, \text{where } \boldsymbol{p} \in \{w_1, x_1, w_2, x_2, b\}$$



$$\frac{d\mathcal{L}}{dz} = a - y \qquad \frac{d\mathcal{L}}{d\boldsymbol{p}} = (a-y)\frac{dz}{d\boldsymbol{p}}$$

$$\frac{d\mathcal{L}}{dw_1} = (a-y)x_1 \qquad \frac{d\mathcal{L}}{dw_2} = (a-y)x_2 \qquad \frac{d\mathcal{L}}{db} = (a-y)$$

Logistic Regression Comp. Graph



If you want to know the influence of a parameter on

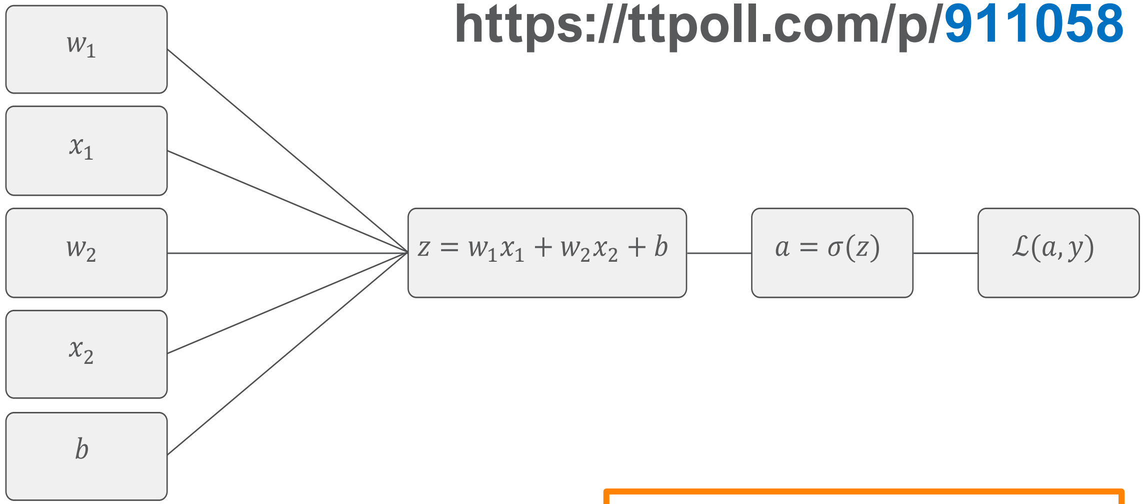
$$\frac{d\mathcal{L}}{d\mathbf{p}} = \frac{d\mathcal{L}}{dz}\frac{dz}{d\mathbf{p}},$$

Note that these derivatives are for the loss.

$$\frac{d\mathcal{L}}{dz} = a - y \qquad \frac{d\mathcal{L}}{d\mathbf{p}} = (a-y)\frac{dz}{d\mathbf{p}}$$

$$\frac{d\mathcal{L}}{dw_1} = (a-y)x_1 \qquad \frac{d\mathcal{L}}{dw_2} = (a-y)x_2 \qquad \frac{d\mathcal{L}}{db} = (a-y)$$

These only capture the influence of parameters for ***one sample***.
But we have m samples.
What should we compute?



Scaling to m samples.

- Computing the cost $\boldsymbol{J(w,b)}$

$$J(w,b)=\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(a^{(i)},y^{(i)}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\sigma\left(w^T x^{(i)} + b\right),y^{(i)})$$

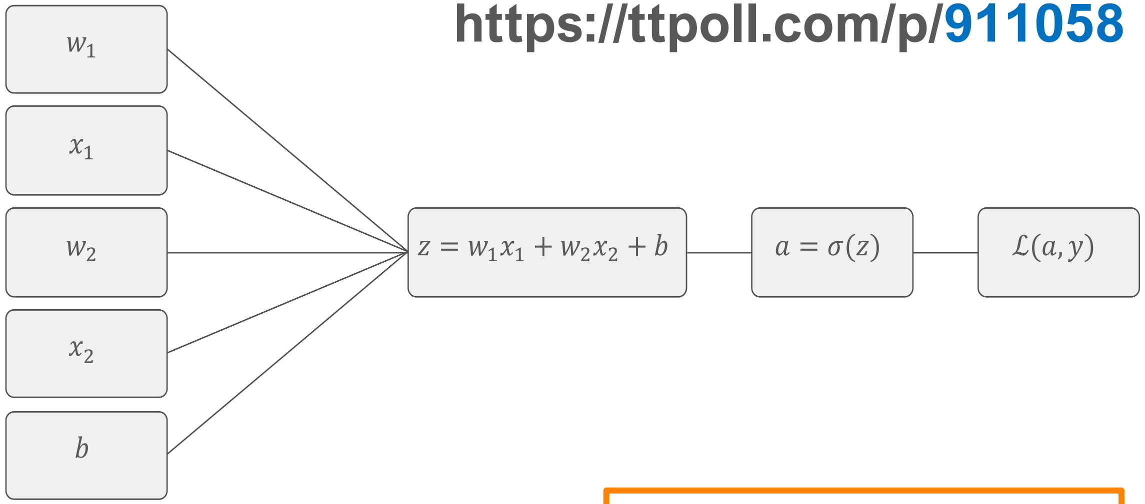
Recall:

$$\frac{d\mathcal{L}}{dw_1} = (a-y)x_1$$

- To our benefit, $J(w,b)$ is the average of the measured losses

$$\frac{\partial J(w,b)}{\partial w_1}=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial \mathcal{L}(a^{(i)},y^{(i)})}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}\left(a^{(i)} - y^{(i)}\right)x_1^{(i)} = \frac{1}{m}\sum_{i=1}^{m}\left(\sigma\left(w^T x^{(i)} + b\right) - y^{(i)}\right)x_1^{(i)}$$

Scaling to m samples.



- Computing the cost $\boldsymbol{J(w,b)}$

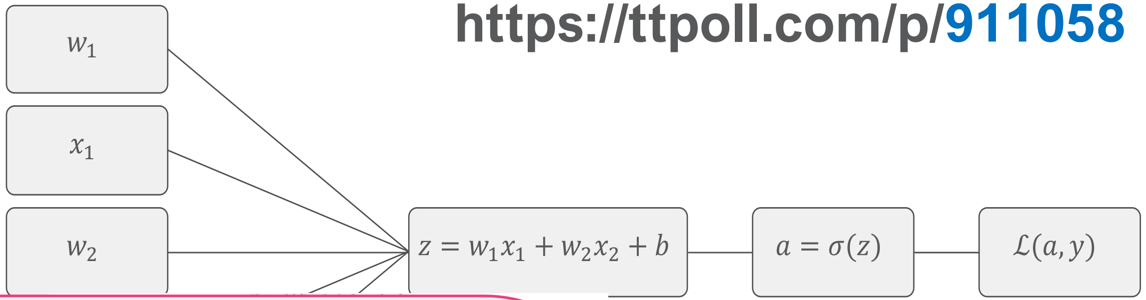
$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(a^{(i)}, y^{(i)}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\sigma(w^T x^{(i)} + b), y^{(i)})$$

Recall:

$$\frac{d\mathcal{L}}{dw_1} = (a-y)x_1$$

- To our benefit, $J(w,b)$ is the average of the measured losses

$$\frac{\partial J(w,b)}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial \mathcal{L}(a^{(i)}, y^{(i)})}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}(a^{(i)} - y^{(i)})x_1^{(i)} = \frac{1}{m}\sum_{i=1}^{m}(\sigma(w^T x^{(i)} + b) - y^{(i)})x_1^{(i)}$$



Scaling to m samples.

- Computing

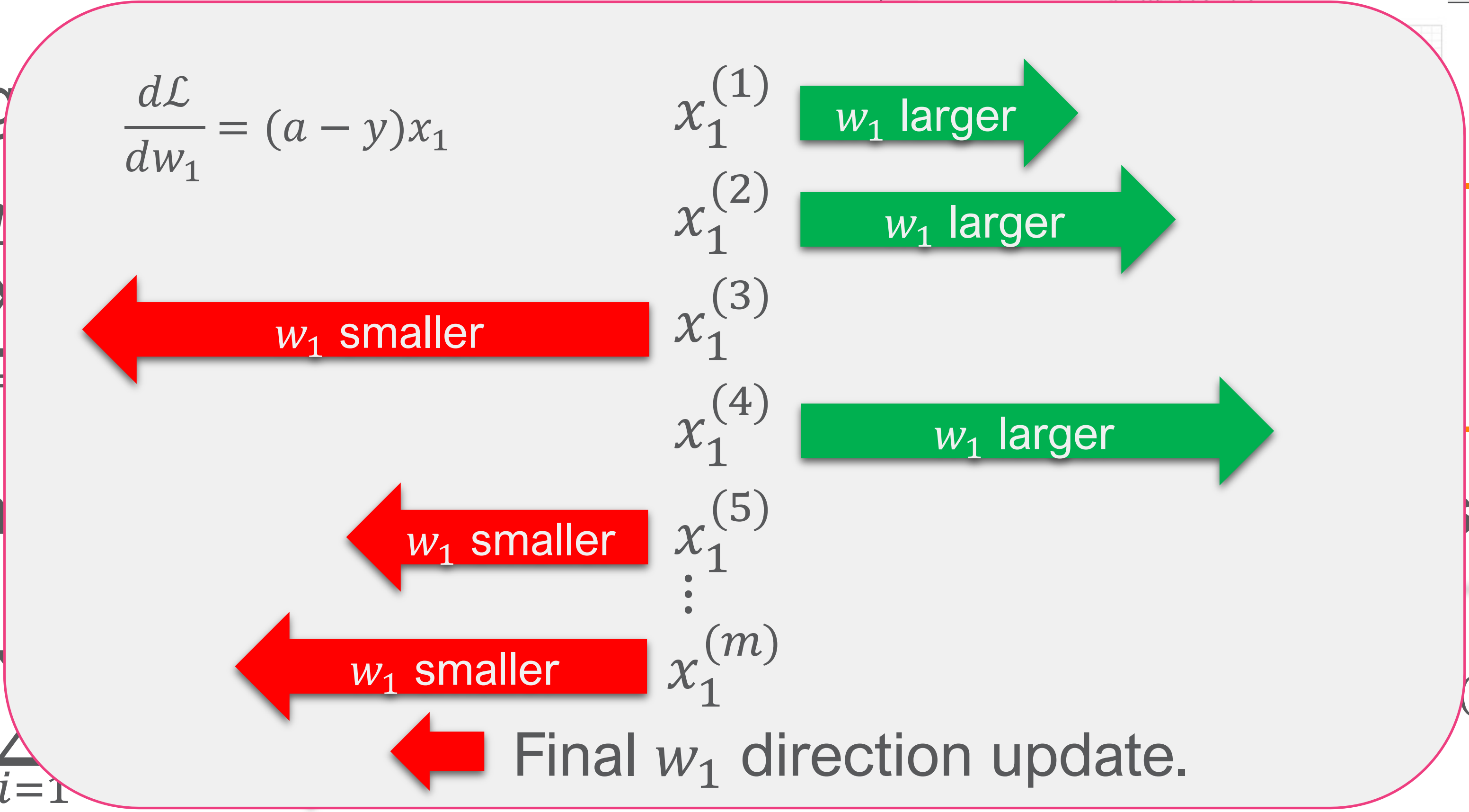
$J(w,b) = \frac{1}{m} \sum_{i=}^{m}$

- To our ben ... ses

$\frac{\partial J(w,b)}{\partial w_1} = \frac{1}{m} \sum_{i=1} \ldots {}^{(i)} + b) - y^{(i)}) x_1^{(i)}$

Recall:

$$\frac{d\mathcal{L}}{dw_1} = (a - y) x_1$$



Gradient Descent Algorithm

$J = 0, dw_1 = 0, dw_2 = 0, \text{and } db = 0$

Initialization of aggregating variables

$Repeat\ from\ i = 1{:}\,m$

$z^{(i)} = w_1 x_1^{(i)} + w_2 x_2^{(i)} + b$

$a^{(i)} = \sigma\left(z^{(i)}\right)$

$J := J - \left[y^{(i)} \log(a^{(i)}) + \left(1 - y^{(i)}\right) \log\left(1 - a^{(i)}\right)\right]$

$dz^{(i)} = \left(a^{(i)} - y^{(i)}\right)$

$dw_1 := dw_1 + \left(dz^{(i)}\right) x_1^{(i)}$

$dw_2 := dw_2 + \left(dz^{(i)}\right) x_2^{(i)}$

$db := db + \left(dz^{(i)}\right)$

Parameters derivatives

$End\ of\ Loop$

$J := J/m\,, dw_1 := dw_1/m\,, dw_2 := dw_2/m, and\ db := db/m$

Notebook Time

Review

- Why is linear regression not a good choice for classification problems?
- Logistic Regression for Classification
  - Decision boundary geometry
  - Computational graph
  - Derivatives for GD algorithm
- Binary Cross Entropy Loss
  - Convex for binary problems
  - Derivatives for GD algorithms

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Next Lecture

- Multi-class classification
- Overfitting/Underfitting
- Bias-Variance Tradeoff
- Regularization

THE UNIVERSITY OF TENNESSEE KNOXVILLE

What: UTK Machine Learning Club

Where: **MK 525**

When: **Tuesday** at **5:00**
(including today)

Who: Any experience level

Everyone is welcome to the first meeting of ML club today. Whether you are a beginner looking to learn from our intro to ML lesson series, experienced practitioner who wants to learn from and discuss with other enthusiasts in our reading groups, or you just want to hear from our industry guest speakers and seminars, utkML can help you scratch your machine learning itch!