Introduction to Data Science With Probability and Statistics

Lecture 24: Logistic Regression

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

What will we learn today?

- ❑ Logistic Regression
- ❑ *Introduction to Statistical Learning, Chapter 4, Think Stats 11.6*

Regression as prediction

So far, we've learned about various forms of regression.

We've viewed regression in terms of learning a relationship between one or more features and a response:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \dots + \beta_p x_{ip} + \epsilon_i$$

We've talked about using regression as a way to make predictions.

What about using regression as a classifier?

Regression as prediction

Example: Back to the Titanic data!

	age	outcome
0	25	survived
1	30	survived
2	35	survived
3	40	survived
4	45	died
5	50	died
6	55	died
7	60	died

Recode outcomes as y = {0, 1}

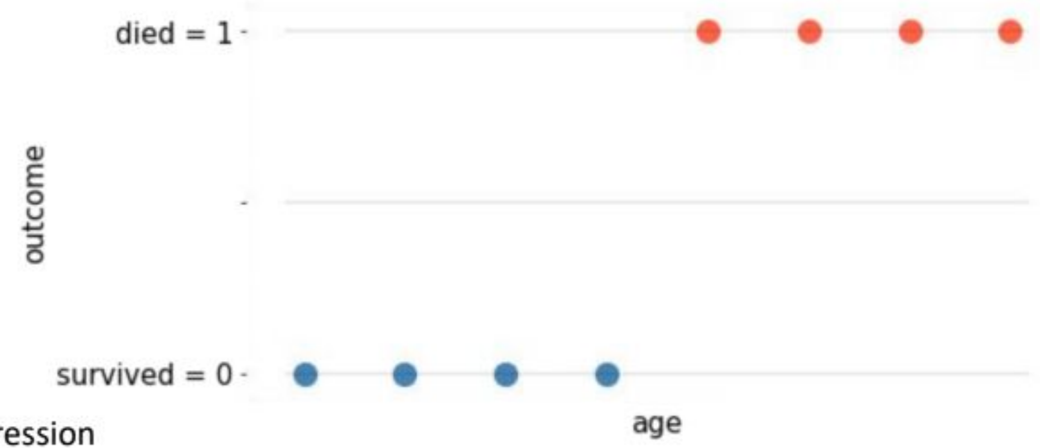
	age	outcome
0	25	0
1	30	0
2	35	0
3	40	0
4	45	1
5	50	1
6	55	1
7	60	1

- Let's try using linear regression to take the feature x = Age and predict the response y = Outcome

Regression as prediction

Example: Suppose you want to predict whether a passenger on the Titanic survived or not, based on passenger Age as the sole feature.

Model input: single feature, $x_1 = age$ Output: prediction, $y = \{0, 1\}$



First Idea: Linear regression

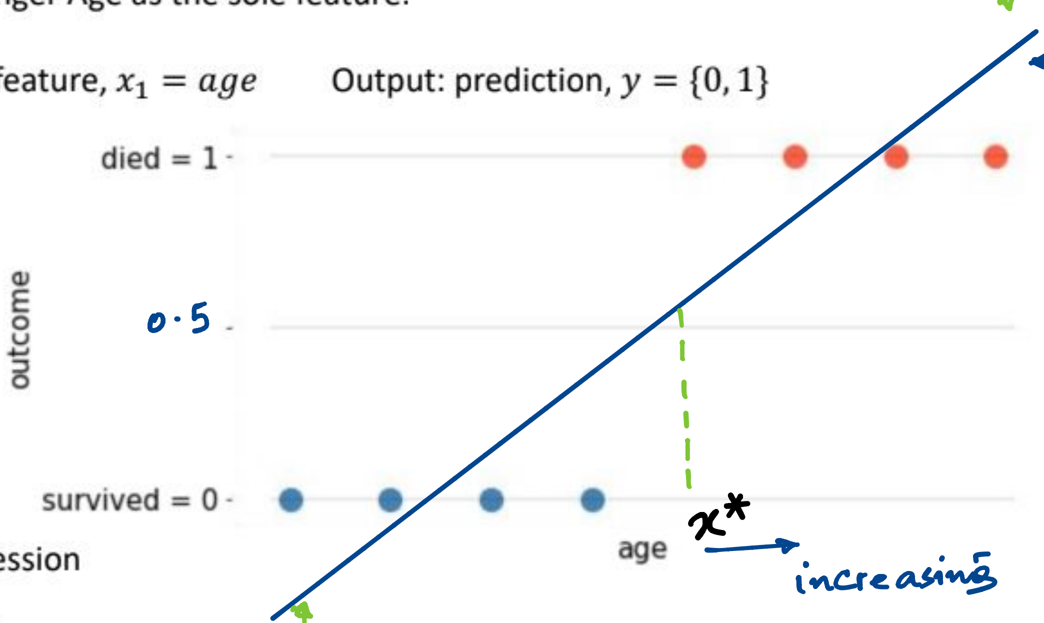
$$y = \beta_0 + \beta_1 x_1$$

Regression as prediction

Example: Suppose you want to predict whether a passenger on the Titanic survived or not, based on passenger Age as the sole feature.

as x increases; $\hat{y}$ goes upto $+\infty$

Model input: single feature, $x_1 = age$ Output: prediction, $y = \{0, 1\}$



x^* : threshold age.

Let's think of the regression results for various x_i as probability.

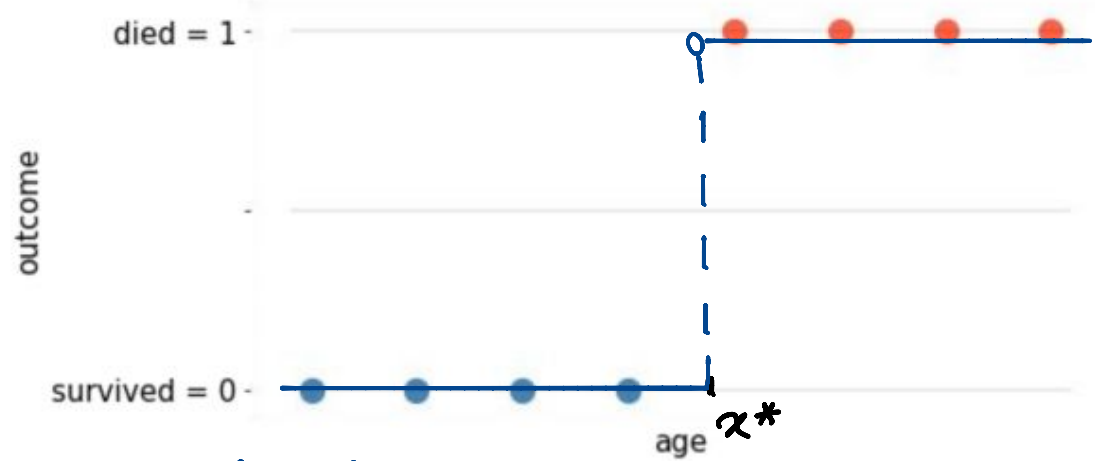
First Idea: Linear regression

$$y = \beta_0 + \beta_1 x_1$$

as x decreases, $\hat{y}$ values goes on to $-\infty$

Regression as prediction

Example: Suppose you want to predict whether a passenger on the Titanic survived or not, based on passenger Age as the sole feature.



Second Idea:

Piecewise function or Step function.

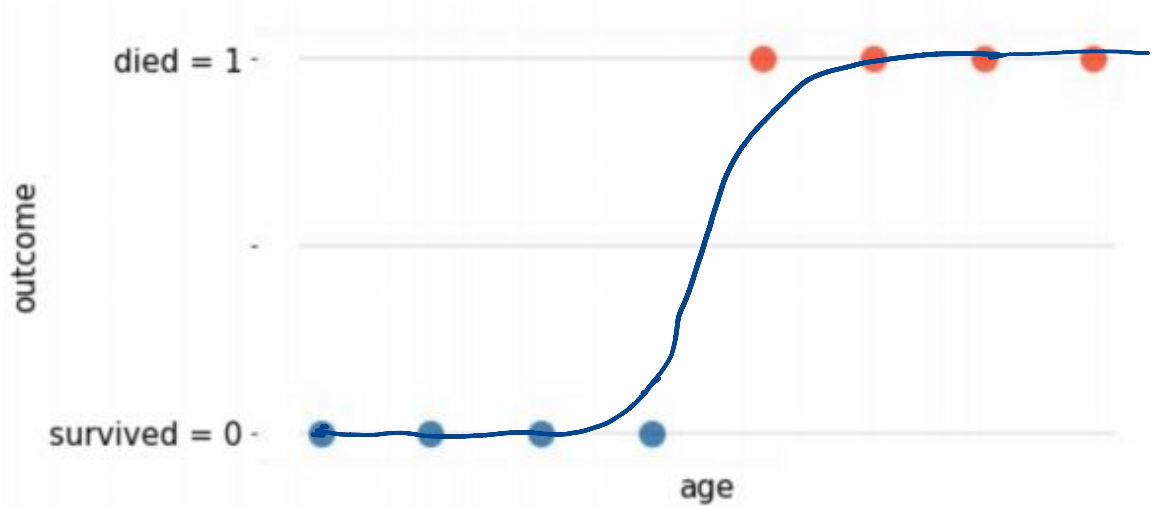
$$y = \begin{cases} 1 & \text{if } x_1 > \text{some threshold} \\ 0 & \text{otherwise} \end{cases}$$

(*) Fit ✓

(*) Continuos / Smooth / Differentiable ✗

Regression as prediction

Example: Suppose you want to predict whether a passenger on the Titanic survived or not, based on passenger Age as the sole feature.

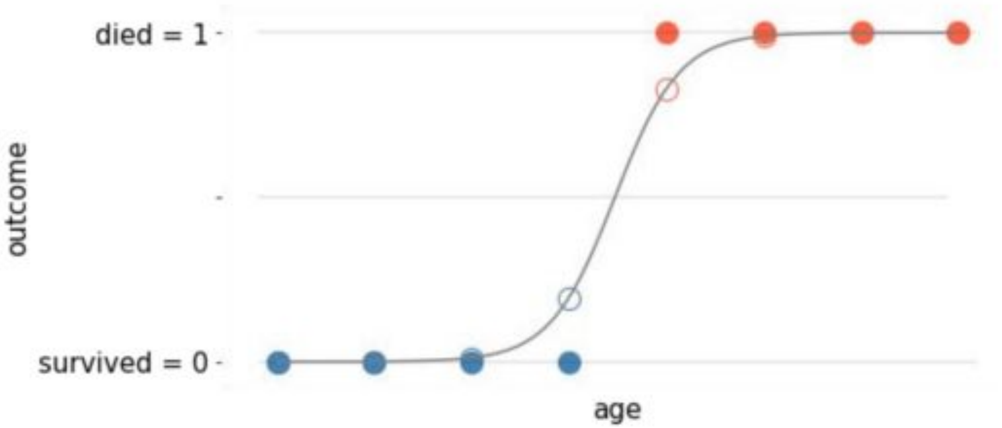


Idea:
Need something that behaves more like a probability

eg: $P(\hat{y} = 1 \mid x = age)$

Regression as prediction

Example: Suppose you want to predict whether a passenger on the Titanic survived or not, based on passenger Age as the sole feature.



This curve looks nice. What is it?

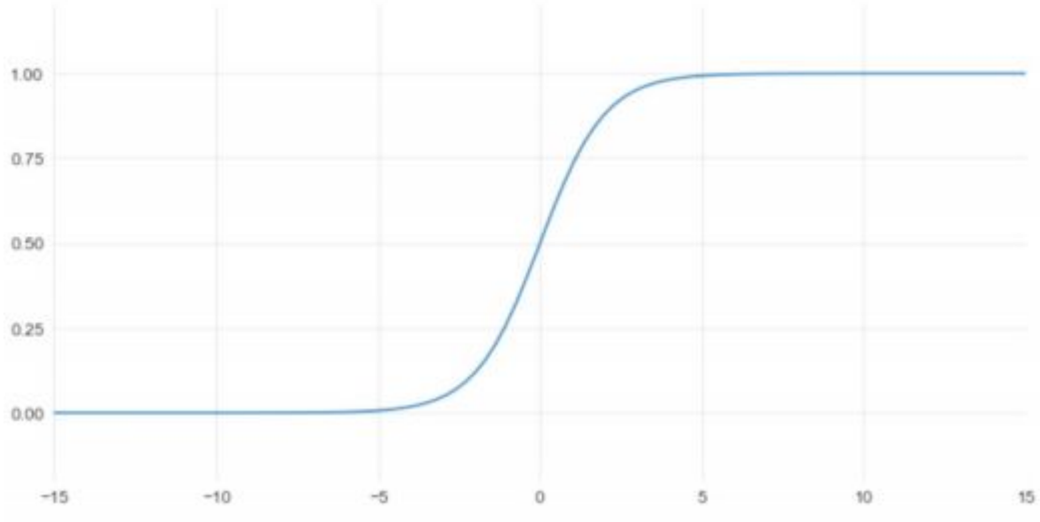
The sigmoid function

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

Has nice properties:

- Behaves like a probability [0, 1]
- Distinguishes between points
- Really smooth

differentiable



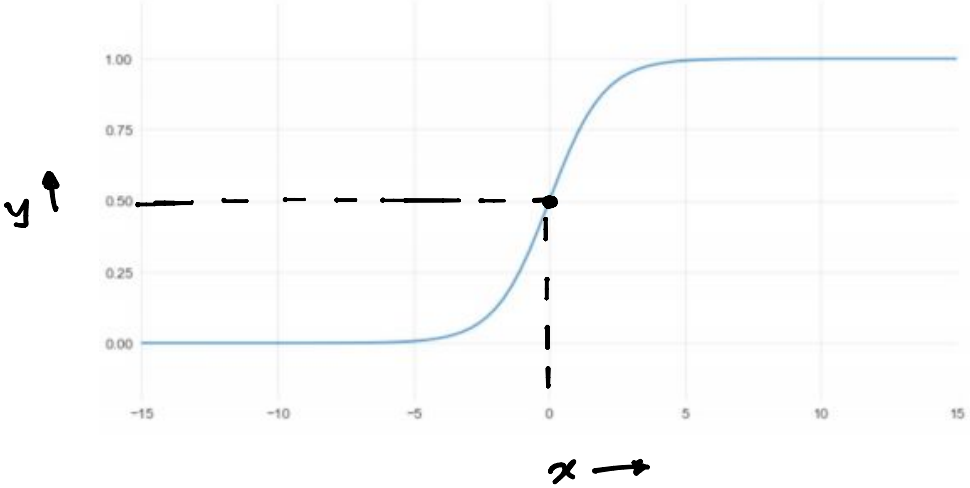
The sigmoid function

- If $x = 0$,

$$\text{sig}(0) = \frac{1}{1+e^{-0}} = \frac{1}{2} = 0.5$$

- $$\lim_{x \to \infty} \frac{1}{1+e^{-x}} = \frac{1}{1+0} = 1$$

- $$\lim_{x \to -\infty} \frac{1}{1+e^{-x}} = \frac{1}{`\infty'} = 0$$



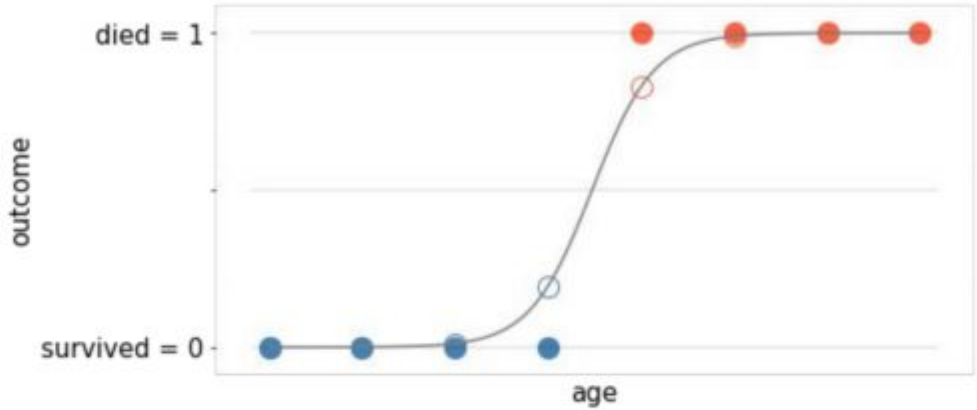
$$\boxed{\text{So, } \text{sig}(x) \in [0,1] \text{ for all } x.}$$

Logistic regression

The model: $p(y = 1 \mid x) = \mathrm{sigm}(\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x}_{\text{Linear regression}})$

Learn weights β_0 and β_1 from the data

Classify data point x according to: $\hat{y} = \begin{cases} 1 & \text{if } \mathrm{sigm}(\hat{\beta}_0 + \hat{\beta}_1 x) \geq 0.5 \\ 0 & \text{if } \mathrm{sigm}(\hat{\beta}_0 + \hat{\beta}_1 x) < 0.5 \end{cases}$



Idea:

Do linear regression & convert the result into the range [0,1] to make a probability measure using sigmoid function

Logistic regression Decision Boundary

for that;

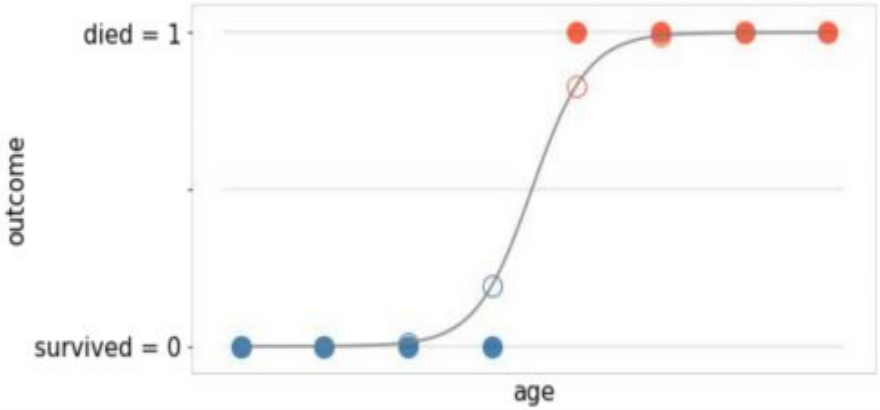
$$\frac{1}{1+e^{-\hat{y}}} = \frac{1}{2}$$

$$\Rightarrow \quad 1 + e^{-\hat{y}} = 2$$

$$\Rightarrow \quad e^{-\hat{y}} = 1 \quad \Rightarrow \quad -\hat{y} = 0$$

$$\therefore \quad -(\hat{\beta}_0 + \hat{\beta}_1 x) = 0$$

$$\Rightarrow \quad \boxed{x = \frac{-\hat{\beta}_0}{\hat{\beta}_1}} = x^*$$



Logistic regression

Our inevitable path to logistic regression and the sigmoid function began with our insistence on modeling the relationship between features and the response as a legitimate probability.

With some basic algebra, we can arrive at an interpretation of logistic regression that is very regression-like.

First we need to talk about odds.

Logistic regression

In statistics, the **odds** of an event is the ratio of the probability that the event occurs, divided by the probability that the event does not occur, and then generally flipped to get a value bigger than 1

$$\text{odds} = \frac{p}{1-p}$$

<u>Example</u>: If p = 0.75, then odds = ____________

We would say that the odds are ____________________

<u>Example</u>: If p=0.1, then odds = ____________

We would say that the odds are ____________________

Logistic regression

In statistics, the **odds** of an event is the ratio of the probability that the event occurs, divided by the probability that the event does not occur, and then generally flipped to get a value bigger than 1

$$\text{odds} = \frac{p}{1-p}$$

Example: If p = 0.75, then odds = $\frac{0.75}{1-0.75} = 3$

We would say that the odds are 3 to 1 in favour.

Example: If p=0.1, then odds = $\frac{0.1}{1-0.1} = \frac{1}{9}$

We would say that the odds are 9 to 1 against

Logistic regression

In logistic regression, we model $p = p(y = 1 \mid x) = \mathrm{sigm}(\beta_0 + \beta_1 x)$

What is we calculate the odds that y = 1, given the data x?

$$odds = \frac{p}{1-p} = \frac{\frac{1}{1+e^{-y}}}{1 - \frac{1}{1+e^{-y}}} = \frac{\frac{1}{1+e^{-y}}}{\frac{1+e^{-y}-1}{1+e^{-y}}} = \frac{1}{\cancel{1+e^{-y}}} \cdot \frac{\cancel{1+e^{-y}}}{e^{-y}}$$

$$= \frac{1}{e^{-y}} = \underline{\underline{e^{y}}}$$

Logistic regression

In logistic regression, we model $p = p(y = 1 \mid x) = \mathrm{sigm}(\beta_0 + \beta_1 x)$

What is we calculate the odds that y = 1, given the data x?

$$odds = \frac{p}{1-p} = \quad \text{(contd.)}$$

$$odds = e^{y} = e^{(\beta_0 + \beta_1 x)}$$

$$\Rightarrow \quad \ln(odds) = \beta_0 + \beta_1 x$$

{Taking Natural log both sides}

Logistic regression

Taking the natural log of both sides, we get:

$$\log(\text{odds}) = \beta_0 + \beta_1 x$$

➢ We have been doing linear regression all along, but for the log-odds instead of probability.

Let's look at the coefficient β_1: odds = $\exp(\beta_0 + \beta_1 x)$ $\{= e^{(\beta_0+\beta_1 x)}\}$

With a unit increase in x, we get: odds = $\exp\big(\beta_0 + \beta_1(x+1)\big) = e^{\beta_0+\beta_1+\beta_1 x} = \underbrace{e^{\beta_1}} \cdot \left(e^{\beta_0+\beta_1 x}\right)$

So we have a new interpretation of the Logistic Regression weight β_1:

For a unit increase of x, the odds change by e^{β_1}.

Logistic regression

The Logistic Regression model with a single feature looks like:

$p(y = 1 \mid x) = sigm(\beta_0 + \beta_1 x)$

But in real life we typically have many features

<u>**Example**</u>:
Predict the probability of precipitation
Features: temperature, pressure, humidity, wind speed, whether it rained yesterday …

Multiple features Logistic Regression model:

$$p(y = 1 \mid x) = sigm\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots + \beta_p x_p\right)$$

Logistic regression

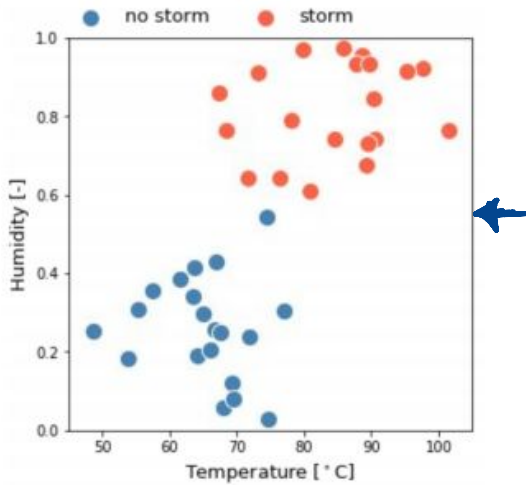
Example:
Predict the probability of precipitation
Features: temperature, pressure, humidity, wind speed, whether it rained yesterday …

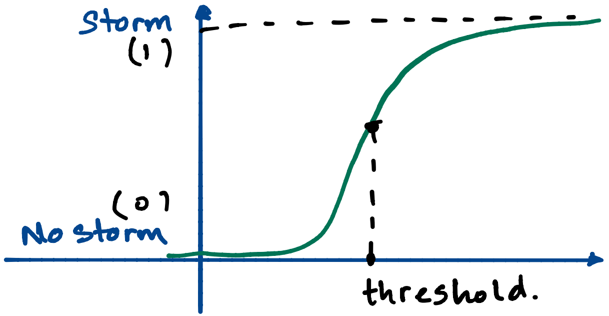
Multiple features Logistic Regression model:

$$p(y = 1 \mid x) = sigm(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

Predict: y = 1 = storm
y = 0 = no storm

Features: x_1 = temperature
x_2 = humidity





Logistic regression

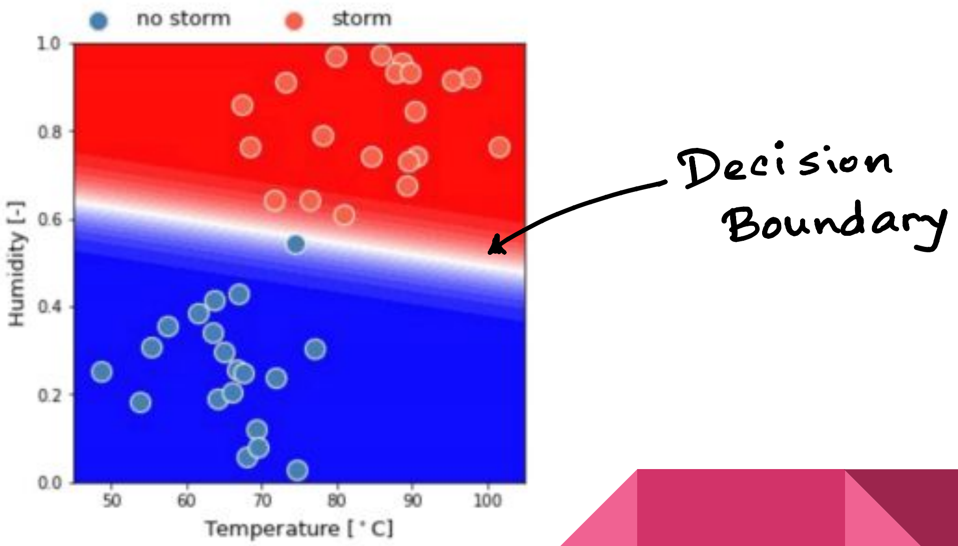
Example:
Predict the probability of precipitation
Features: temperature, pressure, humidity, wind speed, whether it rained yesterday …

Multiple features Logistic Regression model:

$$p(y = 1 \mid x) = sigm(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

Predict: y = 1 = storm
y = 0 = no storm

Features: x_1 = temperature
x_2 = humidity



Logistic regression

The decision boundary is the line/surface that separates predictions into Class 0 and Class 1

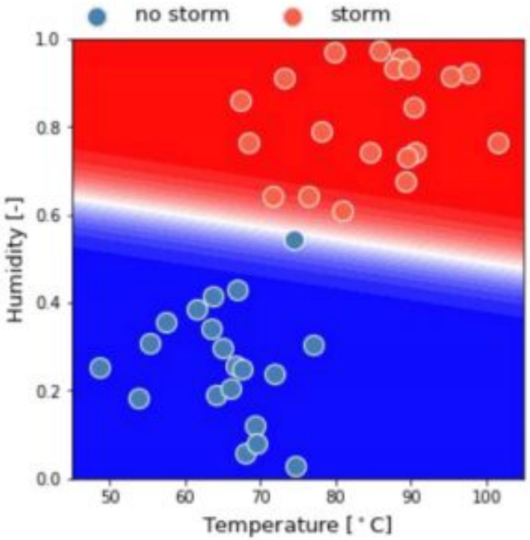
For a 2-feature model, it is described by:

$p(y = 1 \mid x) = sigm(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = 0.5$

Which is just a line in 2D space:

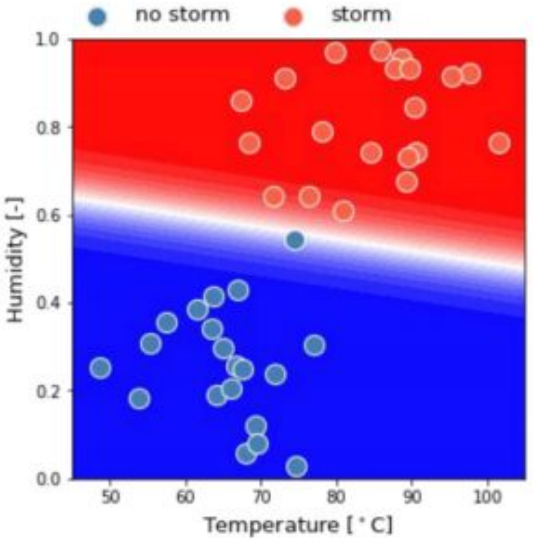
$$\frac{1}{1+e^{-y}} = \frac{1}{2} \Rightarrow e^{-y} = 1.$$

$$\Rightarrow \quad -y = 0$$



Logistic regression

The decision boundary is the line/surface that separates predictions into Class 0 and Class 1



For a 2-feature model, it is described by:
$p(y = 1 \mid x) = sigm(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = 0.5$

Which is just a line in 2D space:

(Contd.)

$$-\beta_0 - \beta_1 x_1 - \beta_2 x_2 = 0$$

$$\Rightarrow \boxed{x_2 = \frac{-\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_1}$$

Straight line

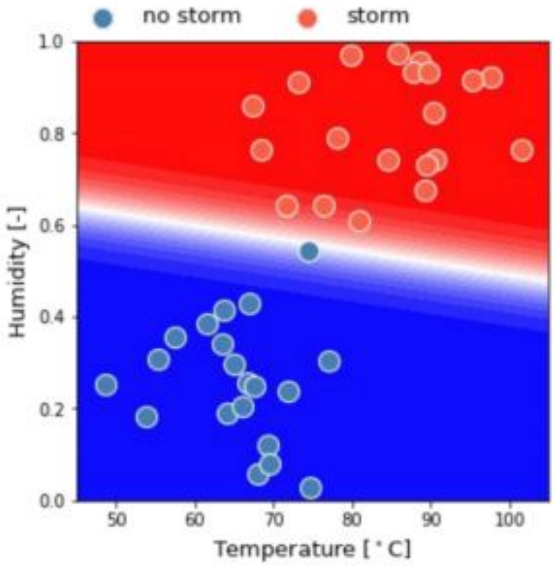
Logistic regression

The Sigmoid function has some nice differential properties that we'll explore next time.

The most important of these is that:

If $f(z) = \mathrm{sigm}(z)$,

then $f'(z) = \mathrm{sigm}(z)(1 - \mathrm{sigm}(z))$



Proof:

$$f(z) = \frac{1}{1+e^{-z}} \quad ; \quad f'(z) = \frac{-(-e^{-z})}{(1+e^{-z})^2}$$

$$= \frac{e^{-z}}{(1+e^{-z})} \cdot \overbrace{\frac{1}{(1+e^{-z})}}^{\mathrm{sig}(z)}$$

Logistic regression

The Sigmoid function has some nice differential properties that we'll explore next time.

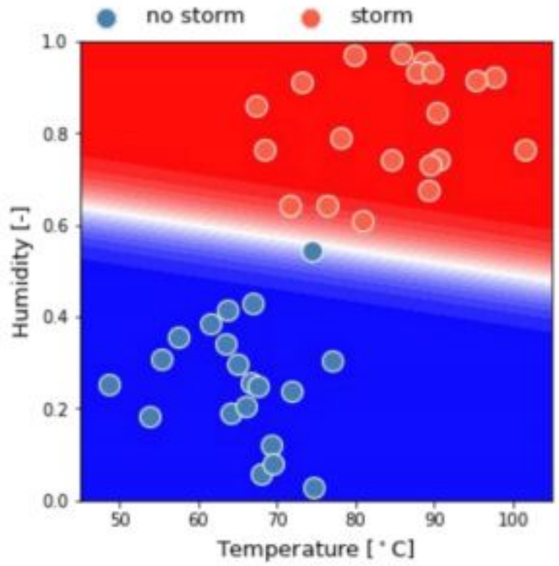
The most important of these is that:

$$\text{If } f(z) = \text{sigm}(z),$$
$$\text{then } f'(z) = \text{sigm}(z)(1 - \text{sigm}(z))$$

(Contd.)

$$\Rightarrow f'(z) = f(z) \cdot \frac{e^{-z}}{1+e^{-z}} = f(z) \cdot \frac{(1+e^{-z}-1)}{1+e^{-z}}$$

$$= f(z)\left(\frac{1+e^{-z}}{1+e^{-z}} - \underbrace{\frac{1}{1+e^{-z}}}_{\text{sig}(z)}\right) = \boxed{f(z)(1-f(z))}$$



Next:

- Maximum Likelihood Estimators
- Confidence Intervals, Hypothesis testing etc