DATA SOCIETY:

Intro to classification - Logistic regression - 1

One should look for what is and not what he thinks should be. (Albert Einstein)

Logistic regression: Topic introduction

In this part of the course, we will cover the following concepts:

- Logistic regression use cases and theory behind it
- Data transformation necessary for logistic regression
- Implementation of logistic regression on a dataset
- Model performance evaluation and tuning

DATASOCIETY: © 2023

Quick Activity

- Suppose we want to predict whether a patient will return to our facility
  - What numerical data might be relevant for making this prediction?
  - What additional qualitative or categorical data might be relevant?
  - How might you handle variables like marital status, education level, or gender?

Module completion checklist

| Objectives	Complete
Determine when to use logistic regression for classification and transformation of target variable	
Summarize the process and the math behind logistic regression	

DATASOCIETY: © 2023

Logistic regression

- **Logistic regression** is a **supervised** machine learning method used for classification
- The **target** or **dependent** variable is **binary**
  - Yes or no
  - This or that
  - 1 or 0

- The outputs are numerical **probabilities** that different observations will be in the desired class ($y = 1$), rather than category labels

DATASOCIETY: © 2023

What logistic regression looks like

- The "logistic" in logistic regression comes from the `logit` function (*a.k.a. sigmoid function*)
- The model solves for coefficients to create a curve maximizing the likelihood of correct classification

DATASOCIETY: © 2023

What logistic regression looks like (cont'd)

- The model's performance can be changed by adjusting the **cut-off probability** where the curve bends, with no need to re-run the model with new parameters
- Note that we convert the target variable to binary values or either 0 or 1 depending on this cut-off or **threshold**
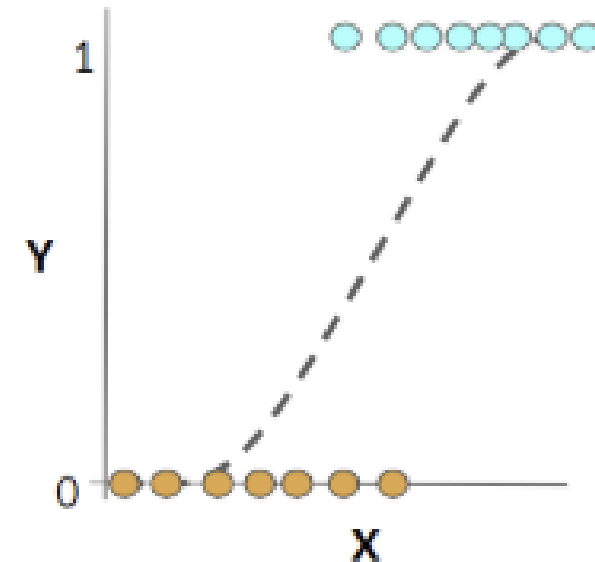

Source
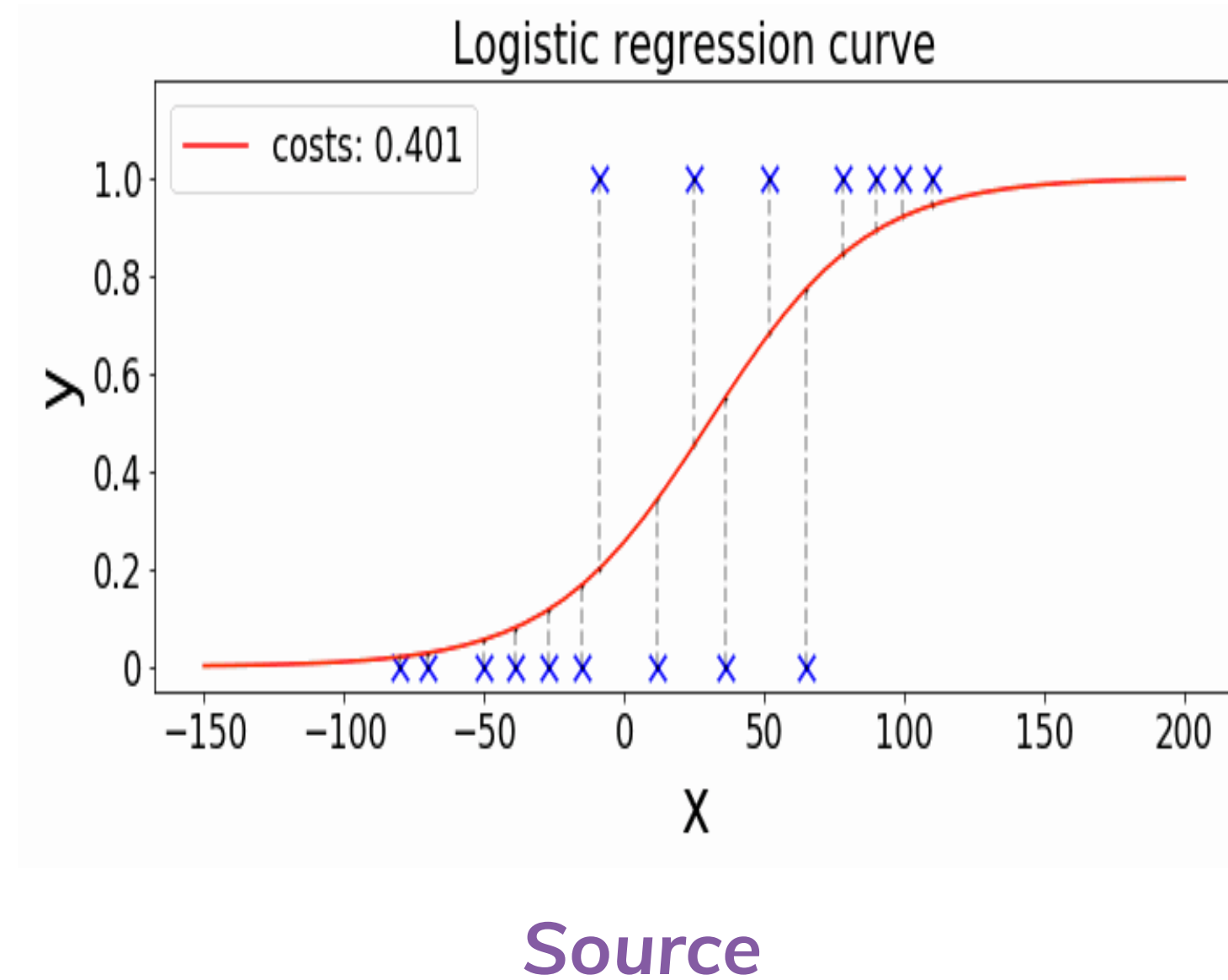
Logistic regression: process

Step 1:
Convert target variable to 1/0
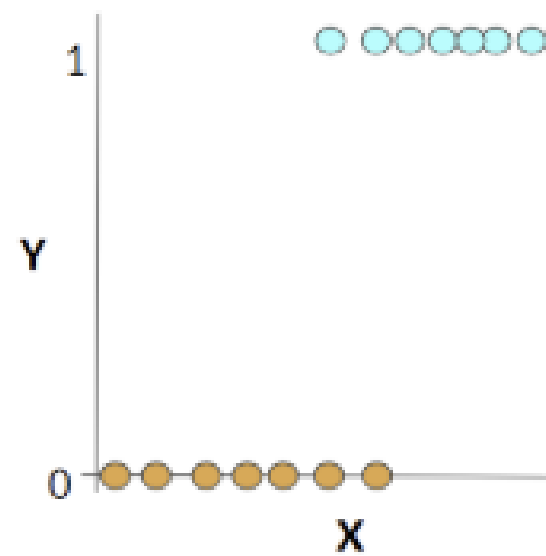
Step 2:
Logistic regression on training data
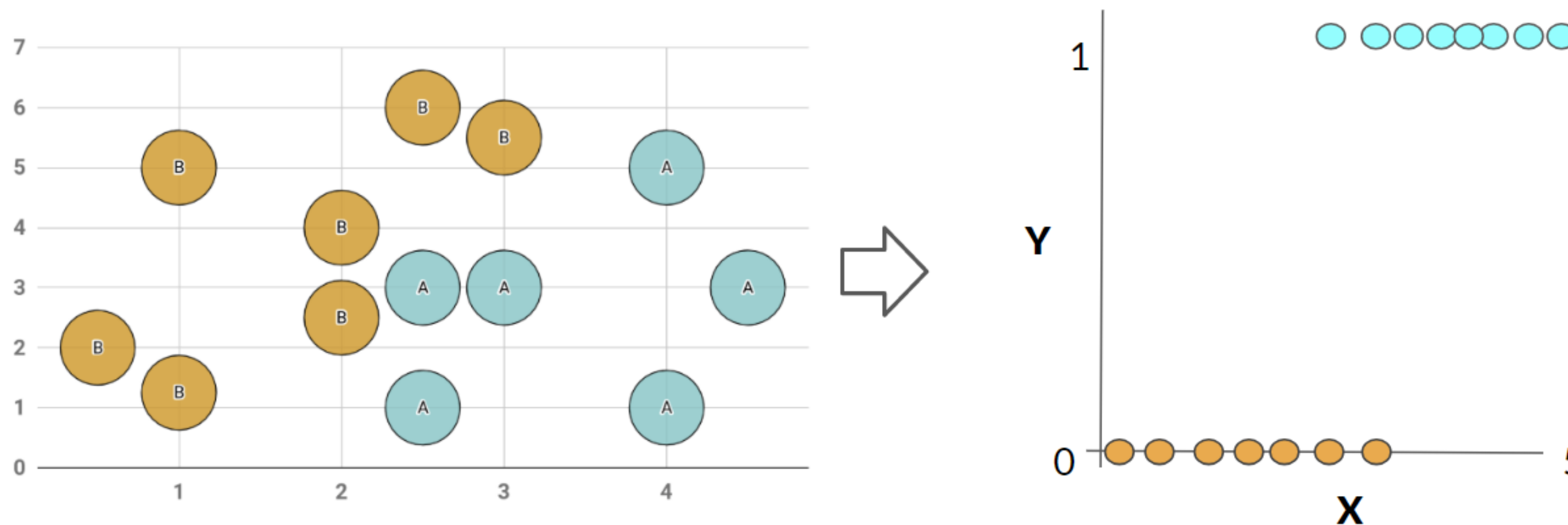
Step 3:
Use ROC curve & AUC to pick threshold

Step 4:
Check performance on test data

Converting categorical to binary variable

- There are two main ways to prepare the target variable:
  - **First method:** translate an existing binary variable (i.e., any categorical variable with 2 classes) into `1` and `0`

Converting continuous to binary variable

- **Second method:** convert a continuous numeric variable into a binary one
  - We can do this by using a **threshold** and labeling observations that are higher than that threshold as `1` and `0` otherwise
  - If the median for the example below was `100`, then any point below the median is coded as `0`, and any point above is `1`

Charge
193.89
0
39.99
201.65
117.9
200.88
79.99

Charge
1
0
0
1
1
1
0

Module completion checklist

Objectives	Complete
Determine when to use logistic regression for classification and transformation of target variable	✔
Summarize the process and the math behind logistic regression	

DATASOCIETY: © 2023

Linear vs. logistic regression

Linear regression line

- For data points $x_1, \ldots, x_n$, we have $y = 0$ or $y = 1$
- The function that "fits" the points is a simple line $\hat{y} = ax + b$

Logistic regression curve

- For the same data points $x_1, \ldots, x_n$, $y = 0$ or $y = 1$
- The function that "fits" the data points is a sigmoid $p(y = 1) = \dfrac{exp(ax+b)}{1+exp(ax+b)}$

Logistic regression: function

- For every value of x, we find p (i.e., probability of success) or probability that $y = 1$
- To solve for p, logistic regression uses an expression called a **sigmoid function:**

$$p = \frac{exp(ax + b)}{1 + exp(ax + b)}$$

- Although it may look a little scary, we can see a very familiar equation inside of the parentheses: $ax + b$
- This is virtually identical to $y = mx + b$

DATASOCIETY: © 2023

Logistic regression: the odds ratio

- Through some algebraic transformations that are beyond the scope of this course, we can change this equation

$$p = \frac{exp(ax+b)}{1+exp(ax+b)}$$

- into a logarithmic expression

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

- Since p is the **probability of success**, $1 - p$ is the **probability of failure**
- The ratio $\left(\frac{p}{1-p}\right)$ is called the **odds** ratio - it tells us the **odds** of having a successful outcome with respect to the opposite
- Knowing this provides useful insight into interpreting the resulting **coefficients**

DATASOCIETY: © 2023

Logistic regression: coefficients

- In **linear** regression, the coefficients in the equation can easily be interpreted

$$ax + b$$

- An increase in x will result in an increase in y and vice versa
- However, in **logistic** regression, the simplest way to interpret a positive coefficient is with an increase in **likelihood**
- A larger value of x increases the likelihood that $y = 1$

Knowledge check

Module completion checklist

Objectives	Complete
Determine when to use logistic regression for classification and transformation of target variable	✔
Summarize the process and the math behind logistic regression	✔

DATASOCIETY: © 2023

Congratulations on completing this module!