CSCI 3022 intro to data science with probability & statistics

Lecture 27
April 25, 2018

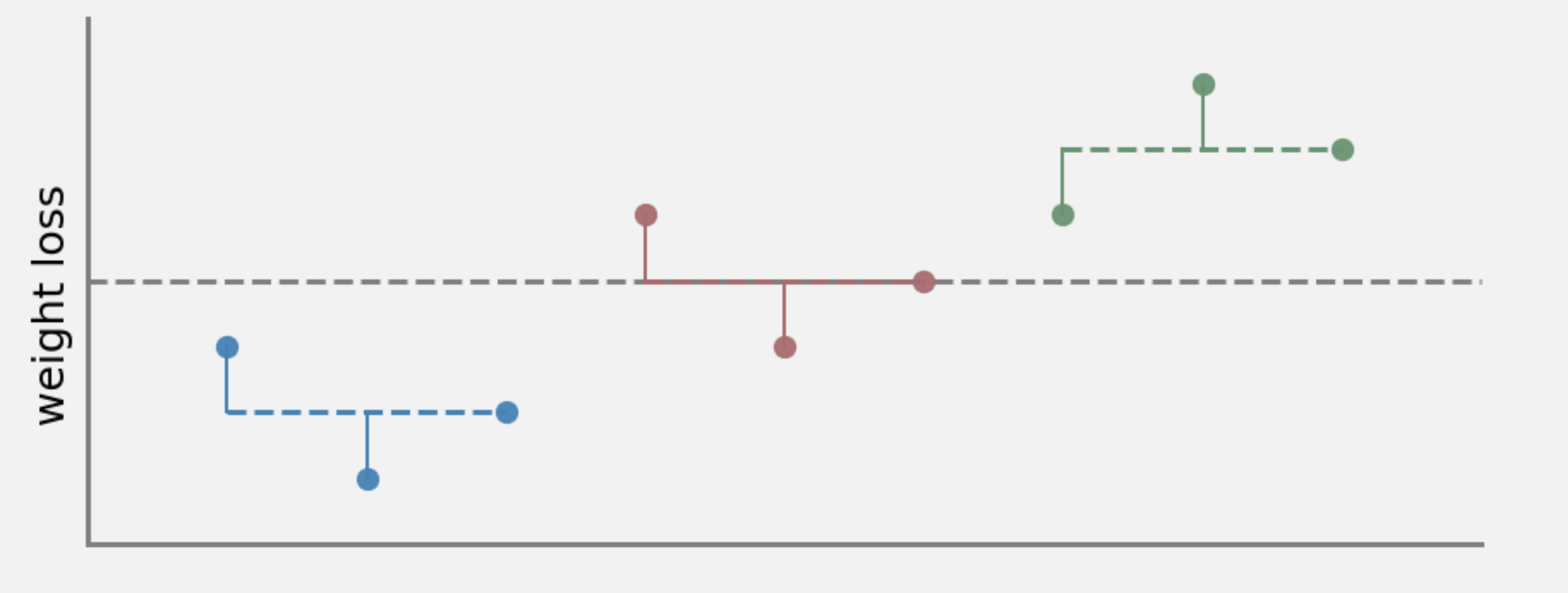
Logistic Regression

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Last time on CSCI 3022:

- If I have responses (y) from multiple different sources, categories, or experiments, how can I tell whether the mean responses are the same?
- ANOVA (ANalysis Of VAriance) answers this by computing the Sum-of-Squares Within (SSW) and Between (SSB) groups.
- **Assumptions**:
  1. Responses are IID samples from normally distr. groups.
  2. The variance of each group is the same.



- If the groups all have the same mean, then SSB will be the same as SSW.
- If one of the groups has a different mean, SSB will be larger than SSW.

$$F = \frac{SSB/(I-1)}{SSW/(N-I)} \qquad F \geq F_{\alpha, I-1, N-1} \qquad SSB = \sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2 \quad SSW = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

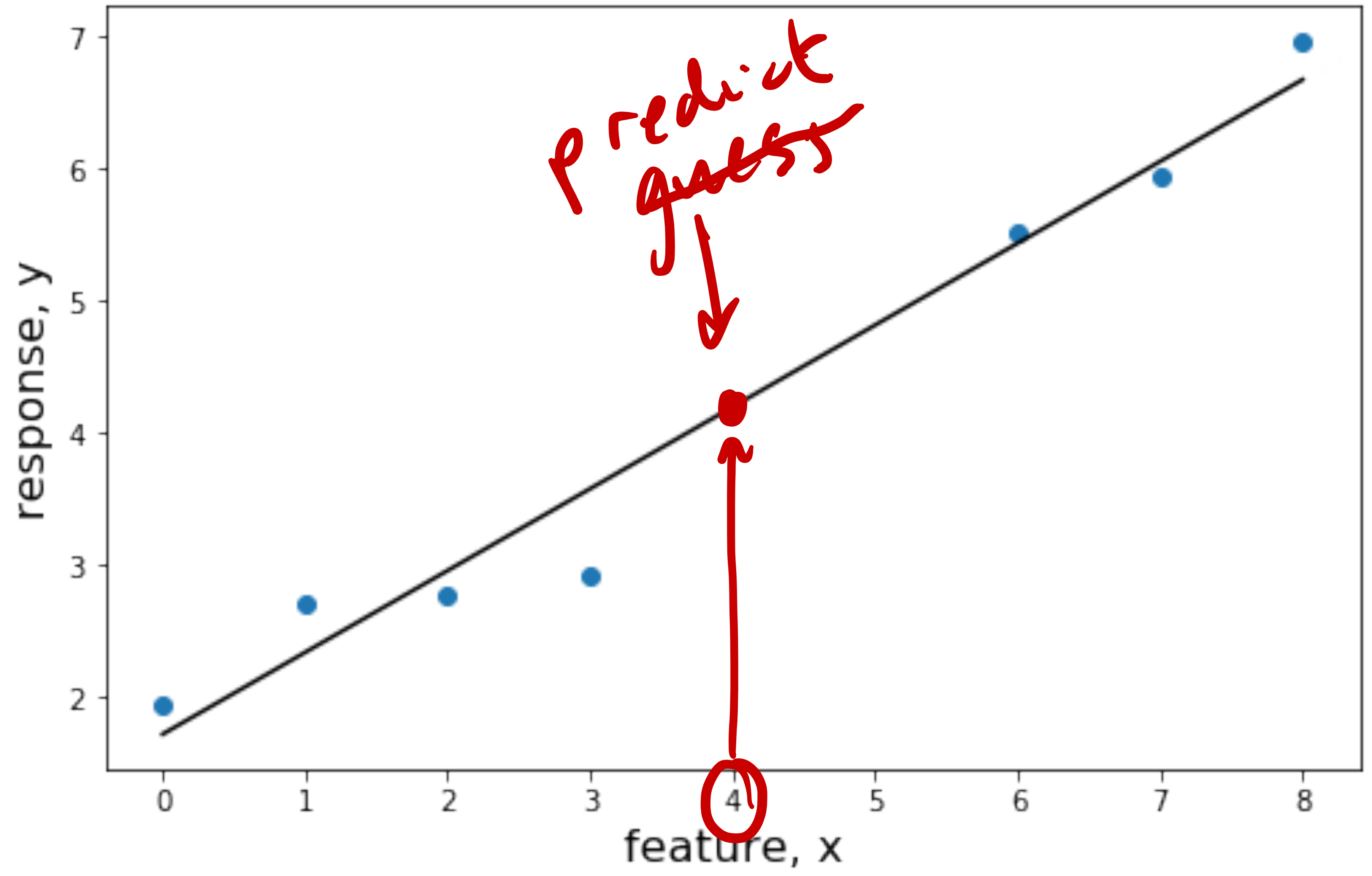
p-value : $1-\text{stats.f.cdf}(F, I-1, N-I)$

Regression as prediction

- So far, we have learned about various forms of *regression*.
- We've talked regression in terms of *learning a relationship* between the features and response.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- We can also think of *a predictor*. If you have the coefficients $\hat{\beta}_0, \hat{\beta}_1 \ldots$ then any time I tell you the features $x_0, x_1 \ldots$ for new data, you can use the equation above to *predict* the response y



Another kind of prediction

- What is our goal is to create a mathematical *classifier*?
- **Definition**: a classifier is a predictor that takes input features $x_0, x_1 \ldots$ and $\boxed{\text{classifies}}$ the response into one of a discrete number of outcomes.
- **Examples of classification problems**:
  - Use the features of bacteria to predict whether they will survive antibiotics.

Outcomes: {survive, perish}

Possible features: motility, reproduction rate, calcium conc.

  - Use the features of freshmen to predict whether they will graduate in ≤5 years.

Outcomes: {grad ≤ 5 yrs, grad > 5 yrs or not graduate}

Possible features:

support level from friends, SAT, coffee cups per week

Regression as a classifier?

x outcomes

- Based on previous classes, it might be tempting to use *linear regression* as a classifier!
- **Example**:

	feature	outcome
0	0.0	Survive
1	1.0	Survive
2	2.4	Survive
3	1.5	Survive
4	2.7	Perish
5	4.0	Perish
6	5.0	Perish
7	4.3	Perish

Re-code the outcomes as y={0,1}. →

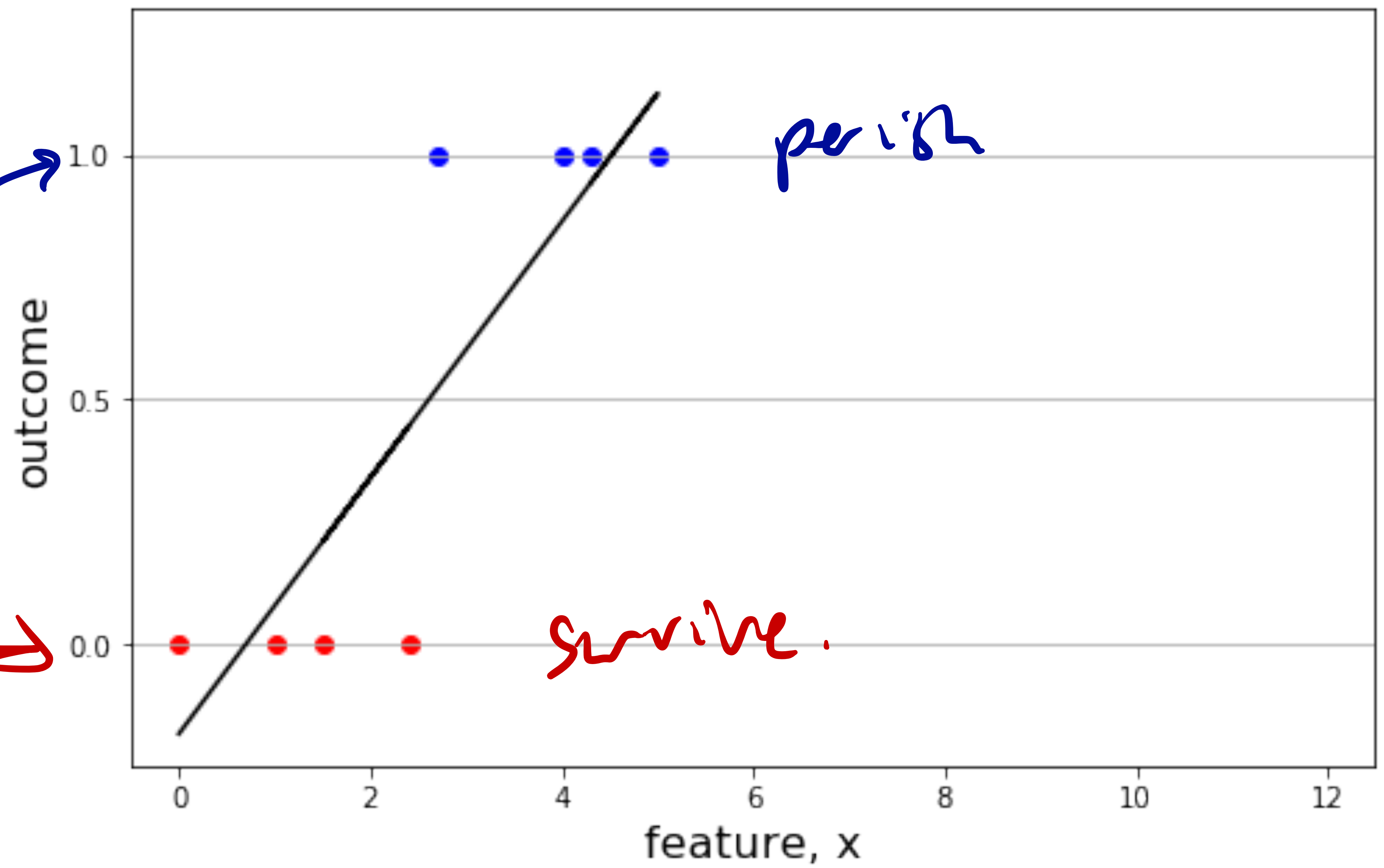
	feature	outcome
0	0.0	0.0
1	1.0	0.0
2	2.4	0.0
3	1.5	0.0
4	2.7	1.0
5	4.0	1.0
6	5.0	1.0
7	4.3	1.0

- Perform linear regression. This would take a value for feature *x* and predict a value for *y*.

Regression as a classifier?

- Based on previous classes, it might be tempting to use *linear regression* as a classifier!
- **Example**:

	feature	outcome
0	0.0	0.0
1	1.0	0.0
2	2.4	0.0
3	1.5	0.0
4	2.7	1.0
5	4.0	1.0
6	5.0	1.0
7	4.3	1.0



- How might we interpret this predicted value y, since our goal is classification?

P(outcome is 1 (perish)) is some f(y). Rule

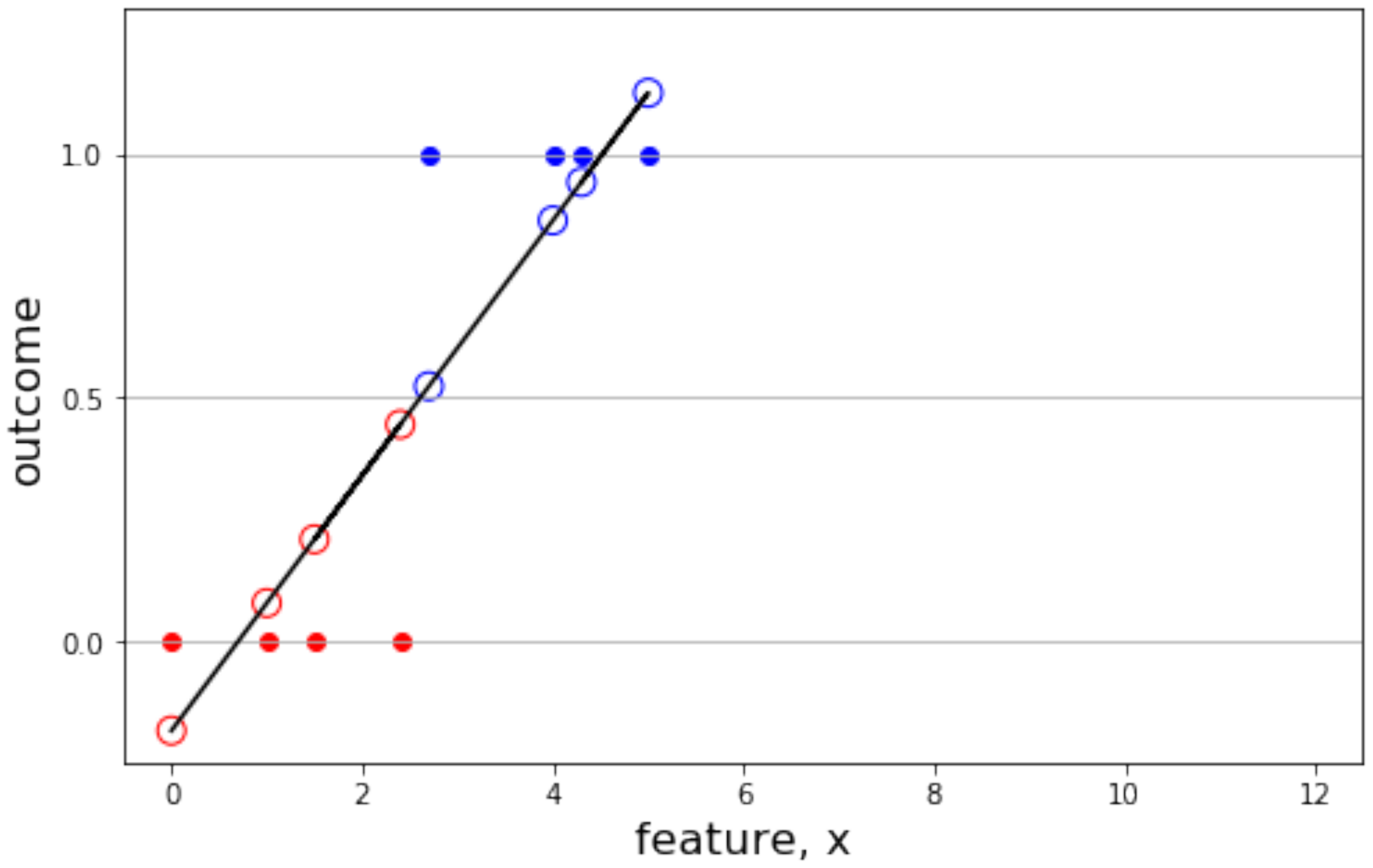
if $y(x) > 0.5 \Rightarrow$ perish

if $y(x) < 0.5 \Rightarrow$ survive

Regression as a classifier?

- Based on previous classes, it might be tempting to use *linear regression* as a classifier!
- **Example**:

	feature	outcome
0	0.0	0.0
1	1.0	0.0
2	2.4	0.0
3	1.5	0.0
4	2.7	1.0
5	4.0	1.0
6	5.0	1.0
7	4.3	1.0

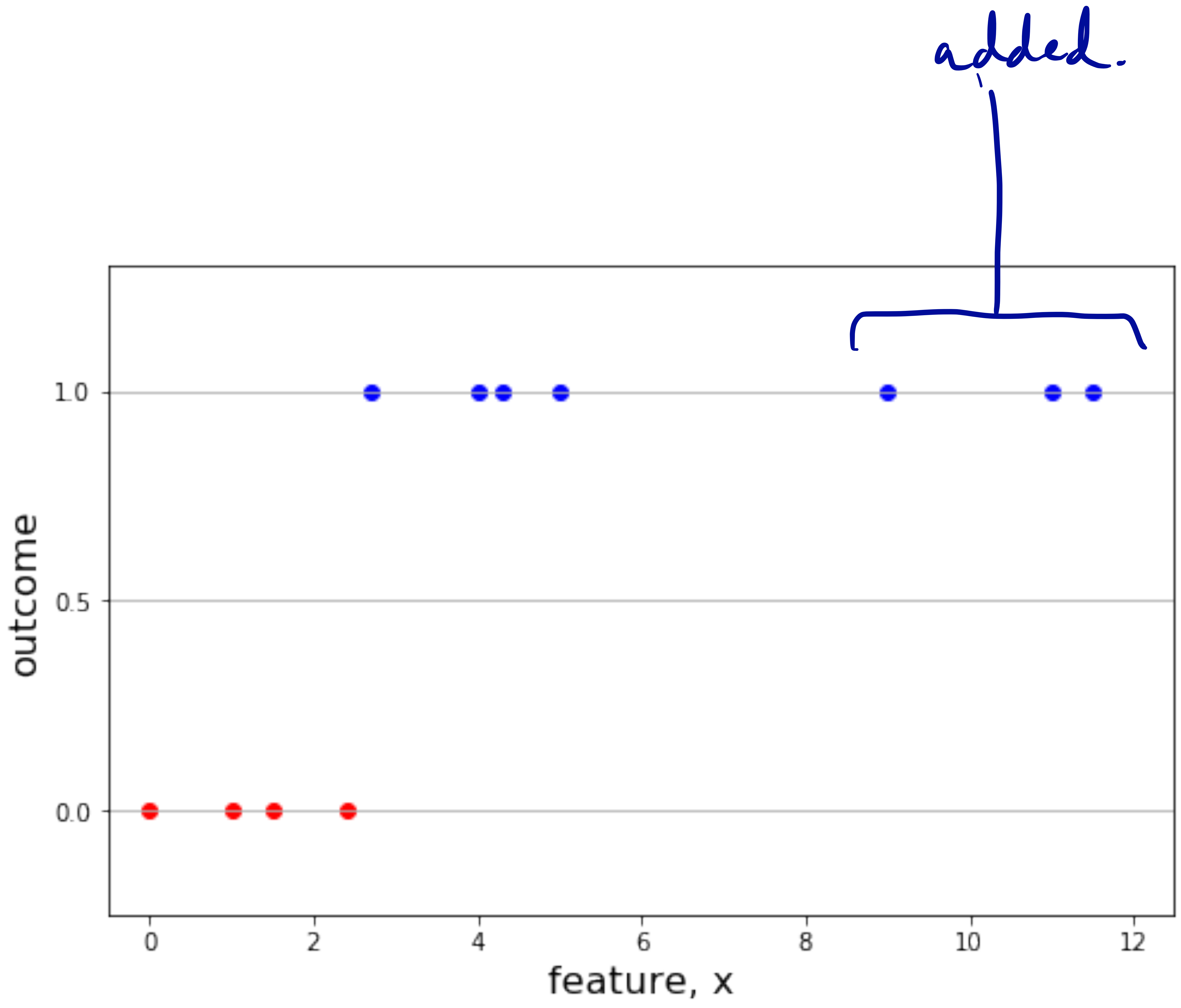


- How might we interpret this predicted value y, since our goal is classification?
- Treat y like a probability! $P(y=1|x) = \underbrace{\beta_0 + \beta_1 x}_{\text{regression line}}$

Regression as a classifier?

- Except we can quickly run into trouble if we fit a line to the data.
- Treating the classes as {0,1} and fitting a line doesn't actually do what we want.

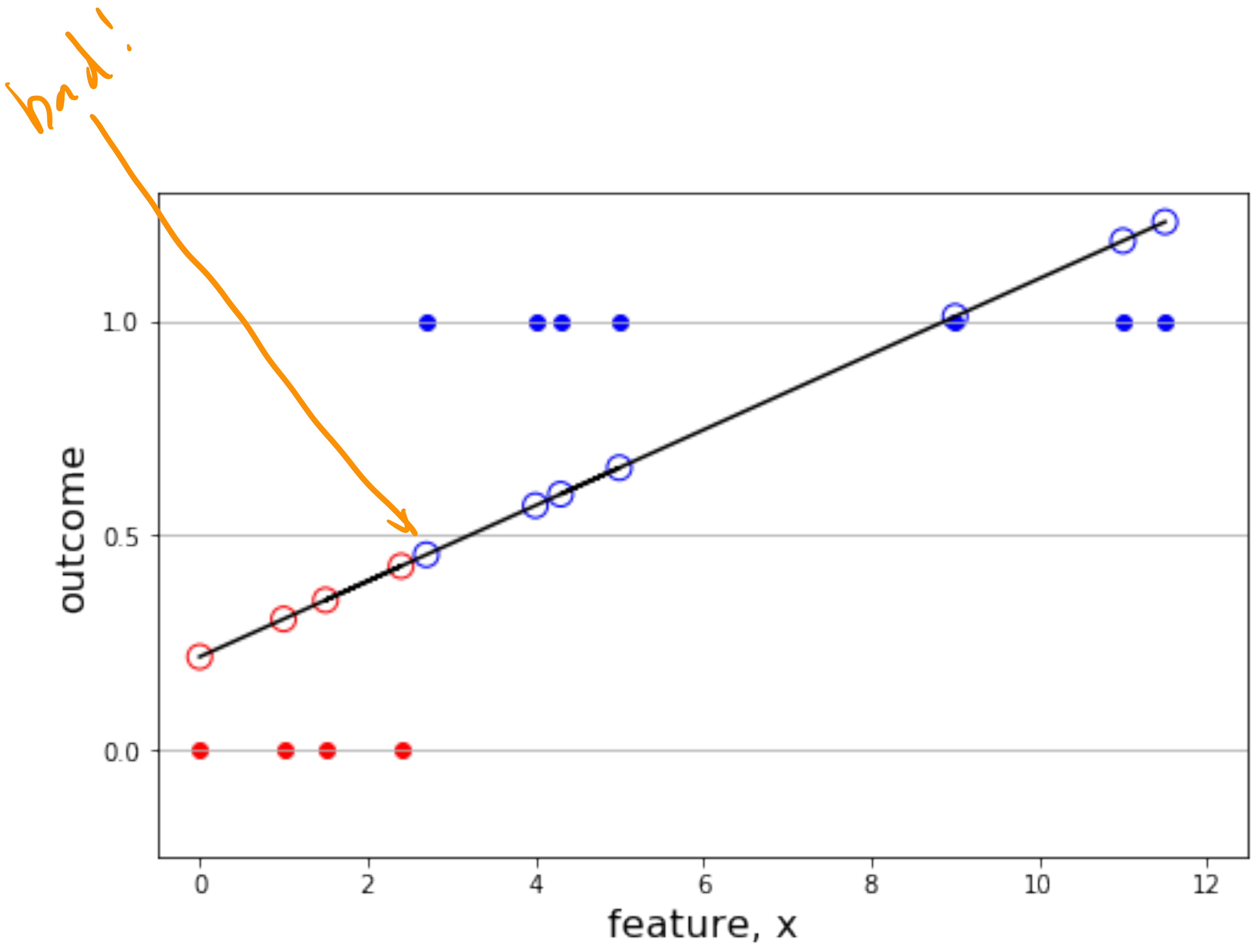
$$P(y=1|x) = \beta_0 + \beta_1 x$$



Regression as a classifier?

- Except we can quickly run into trouble if we fit a line to the data.
- Treating the classes as {0,1} and fitting a line doesn't actually do what we want.

$$P(y=1|x) = \beta_0 + \beta_1 x$$

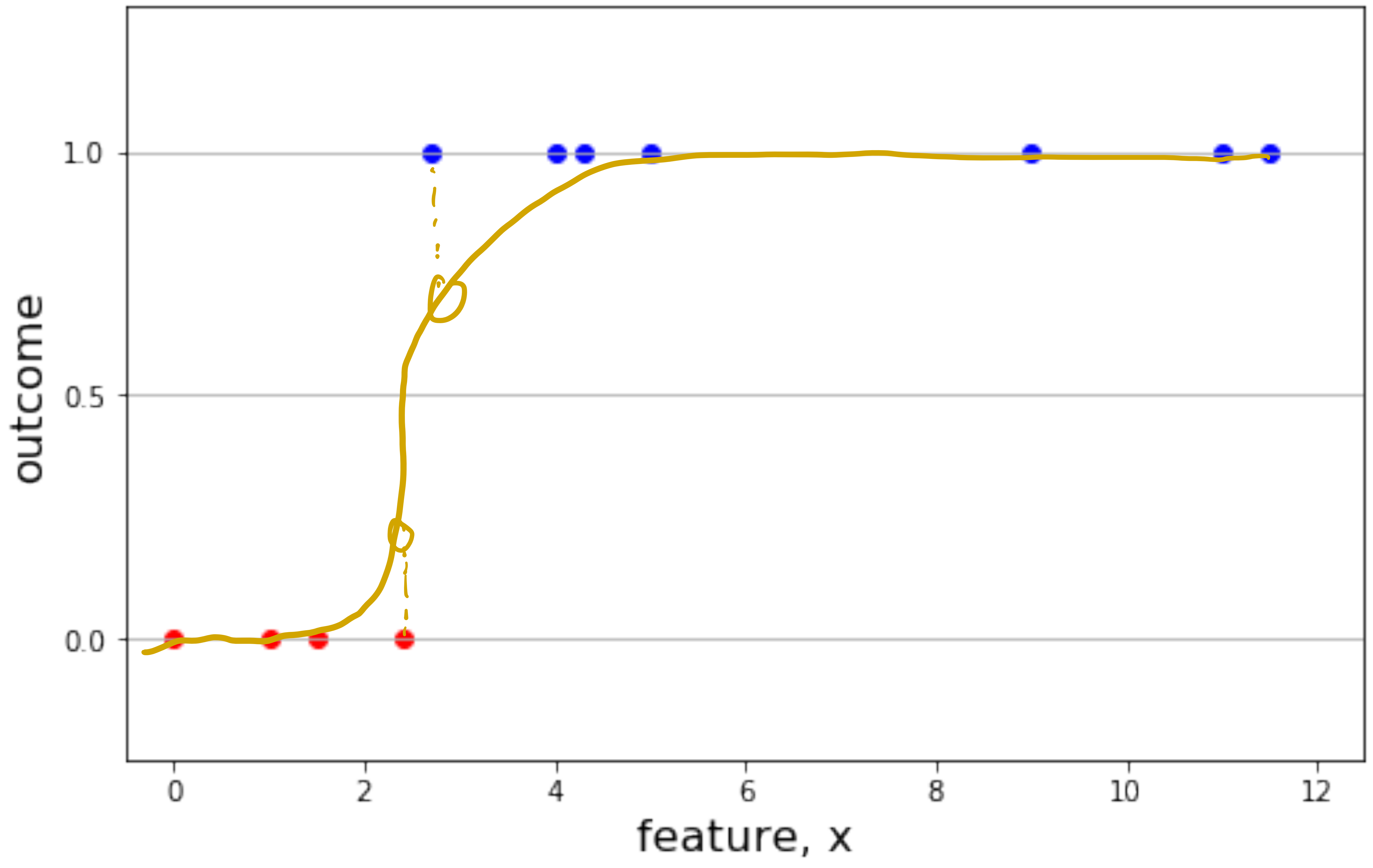


Logistic regression

✓
- Treating the classes as {0,1} seems like the right way to start.
- But modeling with a line clearly is not the right way to go.

$$P(y = 1|x) = \text{something else(x)}$$

- What are the properties we might want in this function?
  - $0 \leq \text{something else}(x) \leq 1$
  - variable slope
  - continuous, differentiable



"sigmoid" "s-curve"

Logistic regression

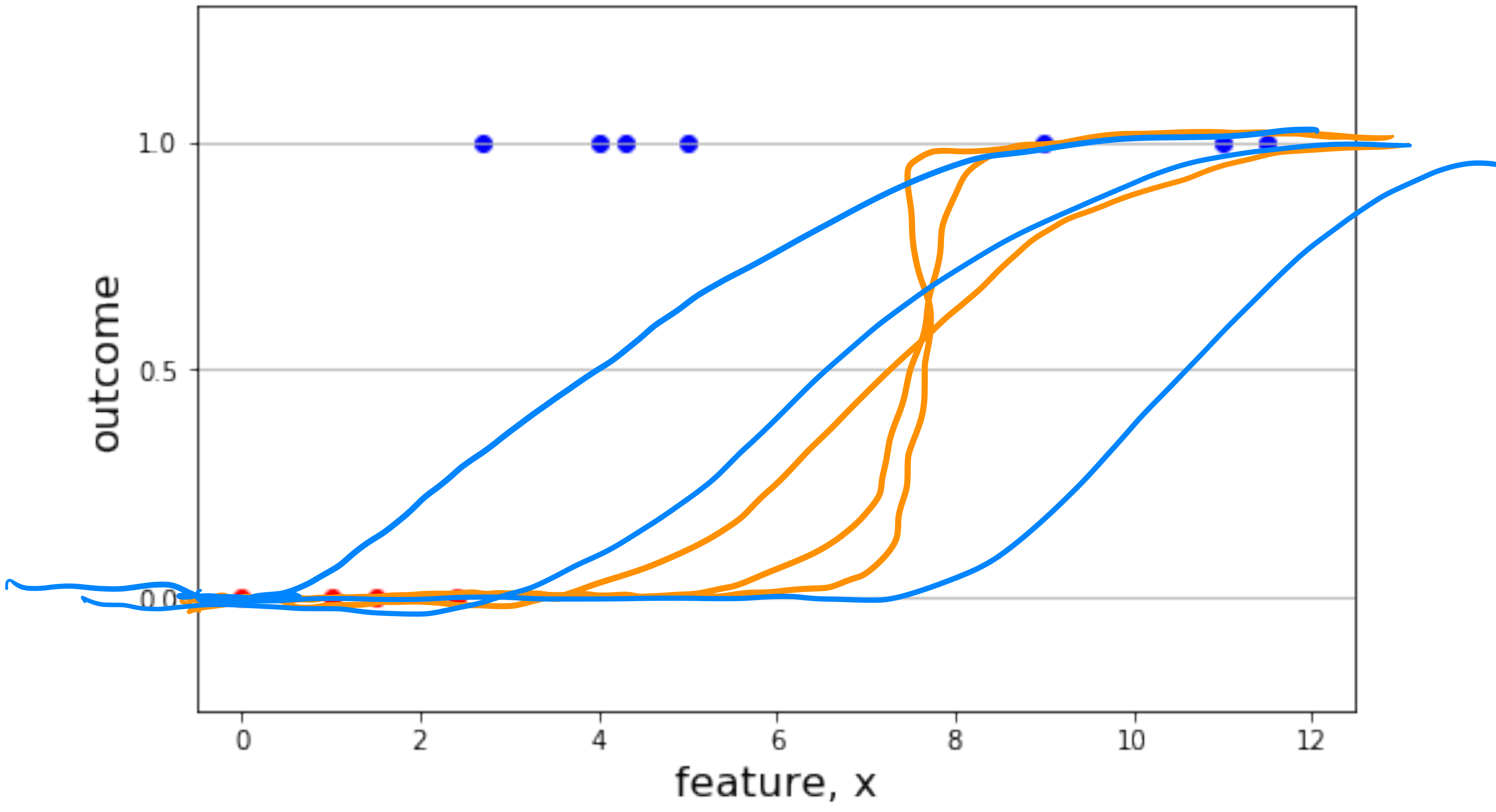
- The **sigmoid function** has the properties that we want for the conditional probability of y, given x.

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

$$= \frac{1}{1+e^{-\beta_0}e^{-\beta_1 x}}$$

$$k = e^{-\beta_0}$$

$$= \frac{1}{1+k e^{-\beta_1 x}}$$

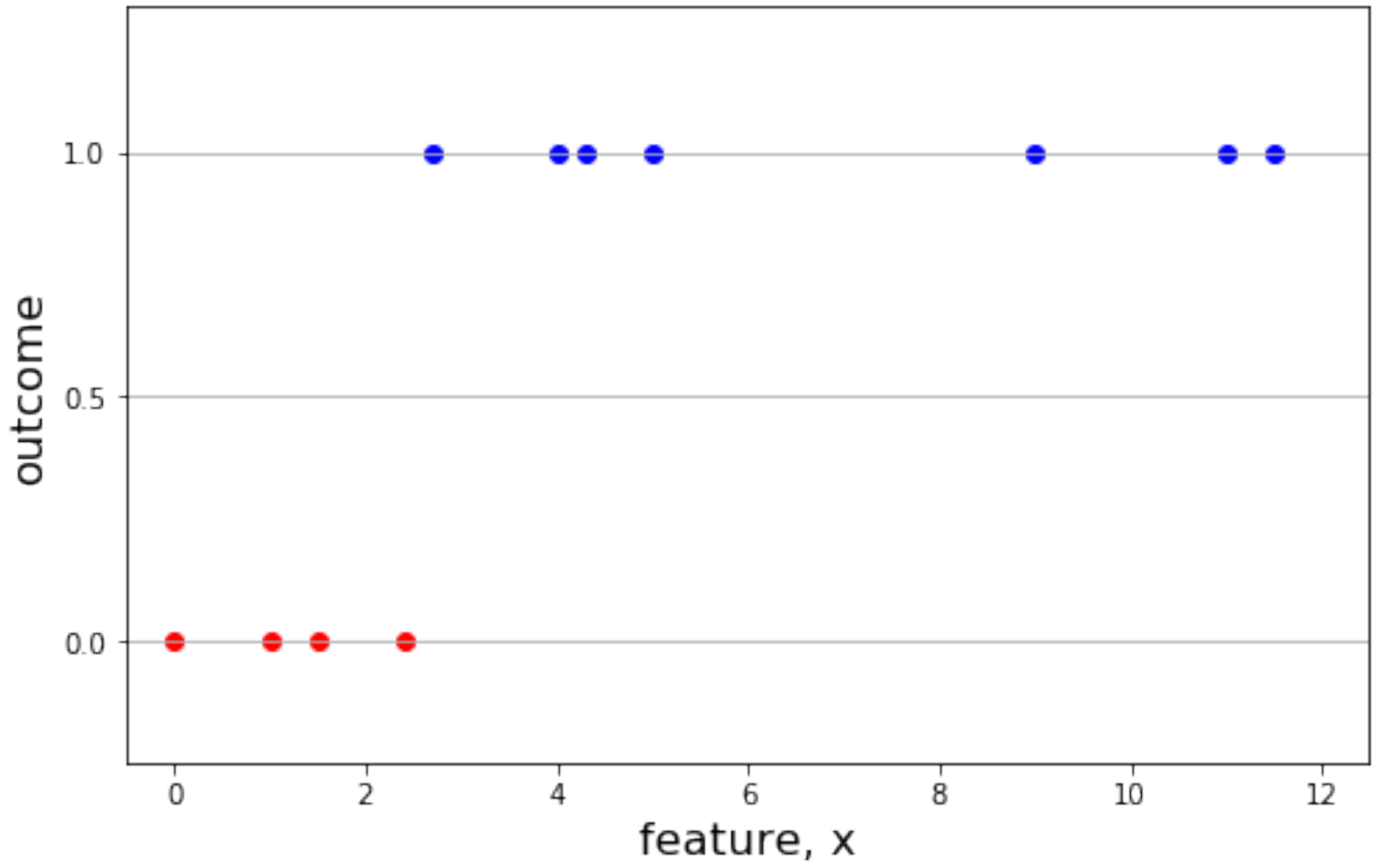


Logistic regression

- The **sigmoid function** has the properties that we want for the conditional probability of y, given x.

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

- Let's group up, grab an in-class notebook, and let's see for ourselves.



Logistic regression

- The **sigmoid function** helps us with the classification problem.

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

"S curve"

- Now, if you give me an *x*, I'll simply compute the value of this function!
  - If P>0.5, the class is *more likely* to be a 1 than a 0. Therefore, classify as a 1.
  - If P<0.5, the class is *less likely* to be a 1 than a 0. Therefore classify as a 0.

- In the notebook we found *good* parameters for the coefficients (betas).

  … but how should we find the *best values*?

sklearn!

Odds

- In statistics, **the odds of an event occurring** are the ratio of the probability that the event will occur to the probability that it will not occur, and then generally flipped to get a value bigger than 1.
- In math: $\frac{p}{1-p} = \text{odds}$
- **Example**: If p=0.75, then odds = $\frac{0.75}{0.25} = 3$.
- We would say the odds are three-to-one in favor
- **Example**: If p=0.1, then odds = $\frac{0.1}{0.9} = \frac{1}{9}$.
- We would say the odds are nine-to-one against.

- **Note**: p is constrained to the interval [0,1], but *odds* can range from 0 to infinity!

Odds

- Previously, we modeled the *probability* that the classification was a 1.

$$P(y=1|x) = \frac{1}{1+e^{-\underbrace{(\beta_0+\beta_1 x)}_{z = \beta_0 + \beta_1 x}}}$$

- What if we consider the odds?

$$\text{odds} = \frac{p}{1-p} = \frac{\frac{1}{1+e^{-z}}}{1 - \frac{1}{1+e^{-z}}}$$

$$= \frac{\frac{1}{1+e^{-z}}}{\frac{\cancel{1}+e^{-z}}{1+e^{-z}} - \frac{\cancel{1}}{1+e^{-z}}}$$

$$= \frac{\frac{1}{\cancel{1+e^{-z}}}}{\frac{e^{-z}}{\cancel{1+e^{-z}}}} \cdot \frac{\cancel{1+e^{-z}}}{\cancel{1+e^{-z}}}$$

$$= \frac{1}{e^{-z}}$$

$$= e^{z} \qquad \text{odds} = e^{z}$$

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

Odds…or log odds?

- Previously, we modeled the *probability* that the classification was a 1.

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

- What if we consider the odds?

$$\text{odds} = \frac{P(y=1|x)}{1-P(y=1|x)} = \cdots = e^{\beta_0+\beta_1 x}$$

- What if we consider the [natural] **log odds**?

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


ARE YOU
NOT
ENTERTAINED

Log odds

- We now have this nice view of the problem: $\text{log odds} = \beta_0 + \beta_1 x$
- There was a **regression problem** hiding in there **the whole time**!
- We have implicitly been doing linear regression *even when we're doing logistic regression.*
- It's a linear regression for the log odds, not for the original probabilities!

Log odds

intercept slope

- We now have this nice view of the problem: $\text{log odds} = \beta_0 + \beta_1 x$

feature

- There was a **regression problem** hiding in there **the whole time**!

- We have implicitly been doing linear regression *even when we're doing logistic regression.*
- It's a linear regression for the log odds, not for the original probabilities!

- Let's go back into the notebooks and learn how to actually fit a logistic regression in Python.
- To do this, we'll introduce a new package, *sci kit learn* or *sklearn*.

done. nbd.

I heard you like features…

- We know that a classification problem based on a single feature *x* looks like $\text{log odds} = \beta_0 + \beta_1 x$
- My [actual research] goal [with Joel Kralj in MCDB]:
  - **Predict** whether each bacterium will live or die when we add antibiotics!
  - **Features**: motility, rate of division, genetic mutation counts, calcium concentration…
- More features? Not a problem! $\text{log odds} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
- What does this mean for the probability?

$$P(y=1 \mid \underset{\uparrow}{\vec{x}}, \vec{\beta}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}} = \frac{1}{1+e^{-\vec{\beta}\cdot\vec{x}}} = \frac{1}{1+e^{-\beta^T x}}$$

vector of features

vector of coeffs

Logistic Regression with Many Features

- Multiple feature logistic regression: $p(y=1 \mid x) = \mathrm{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$

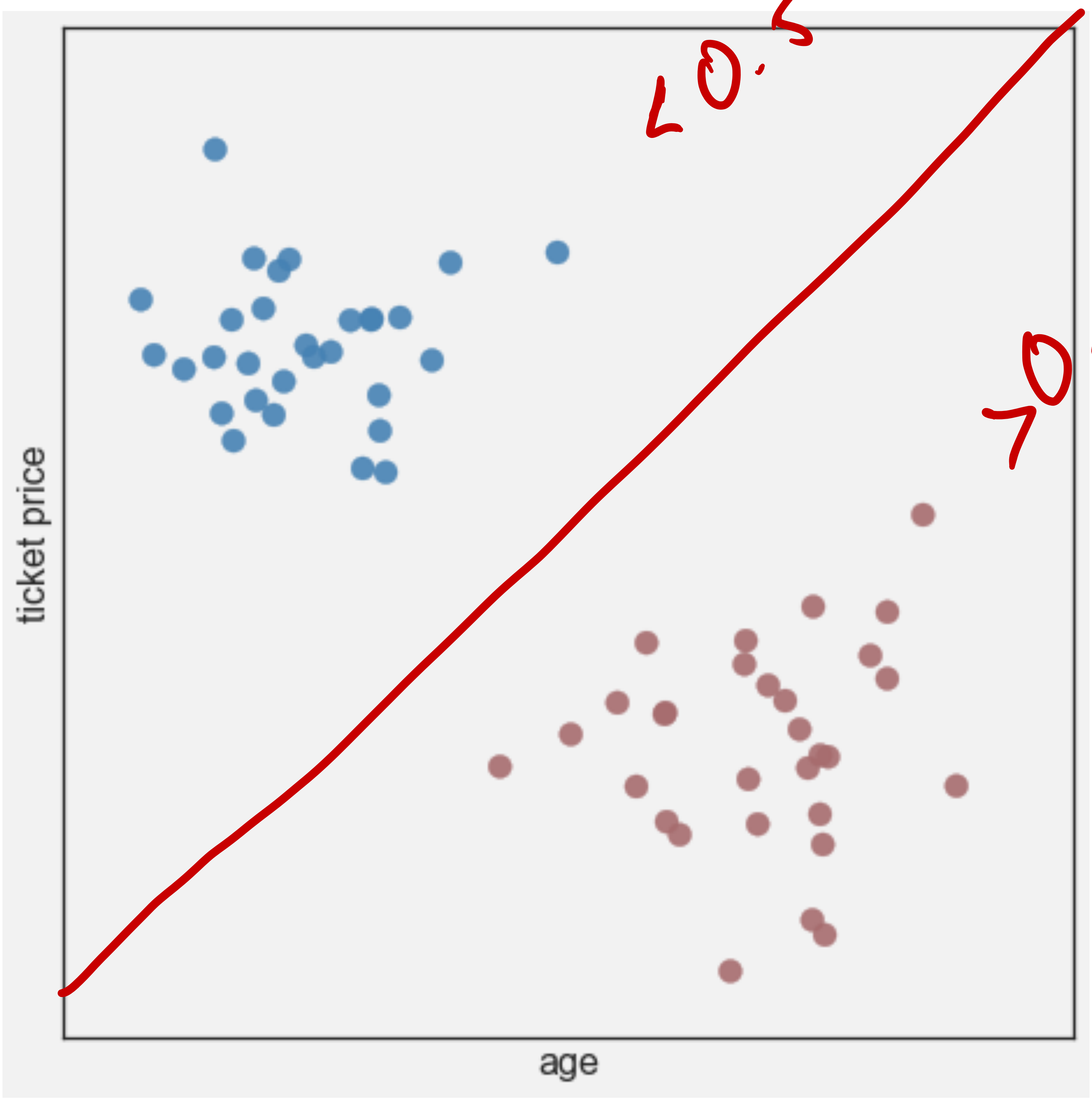
2 features

x_2

<0.5

0.5 line

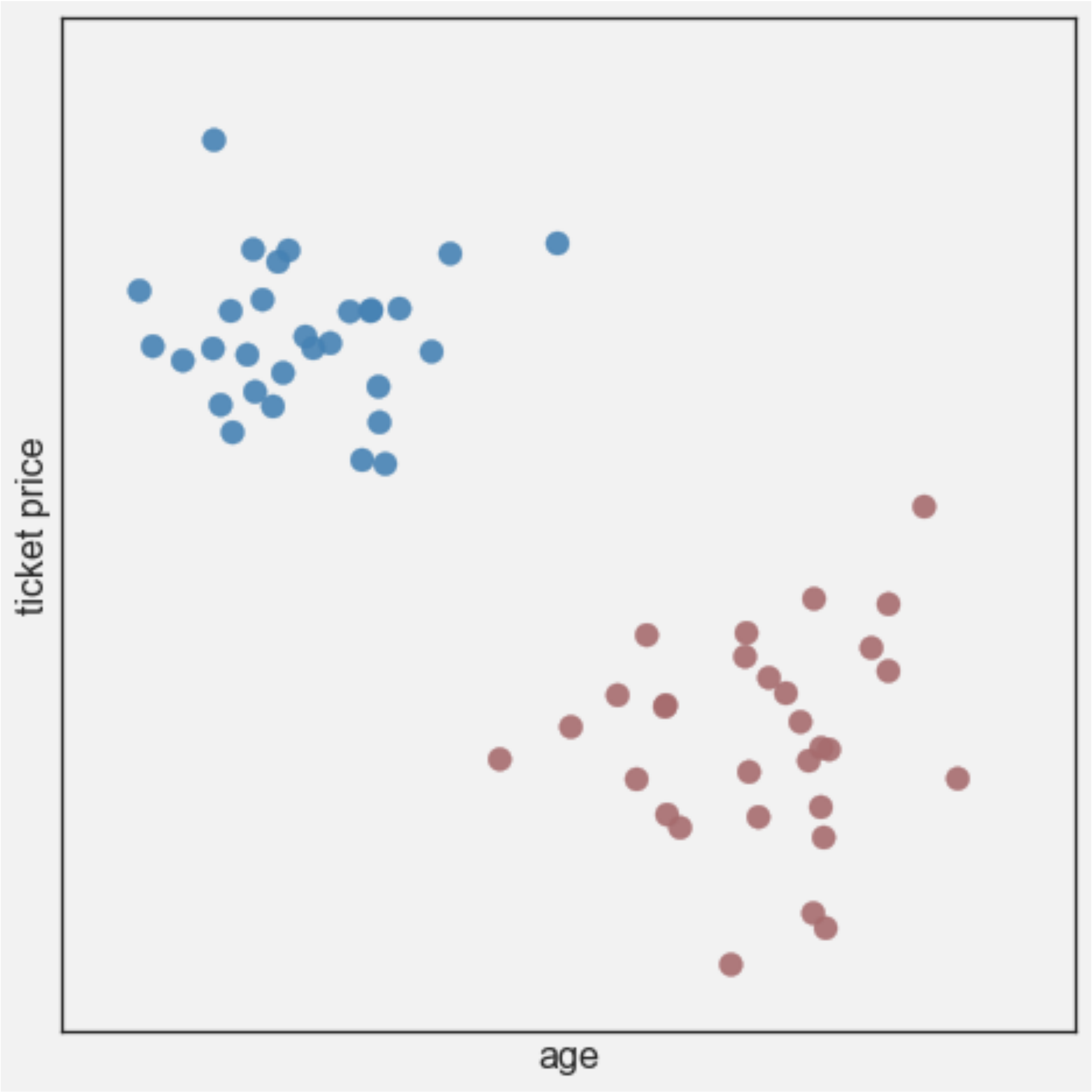
>0.5

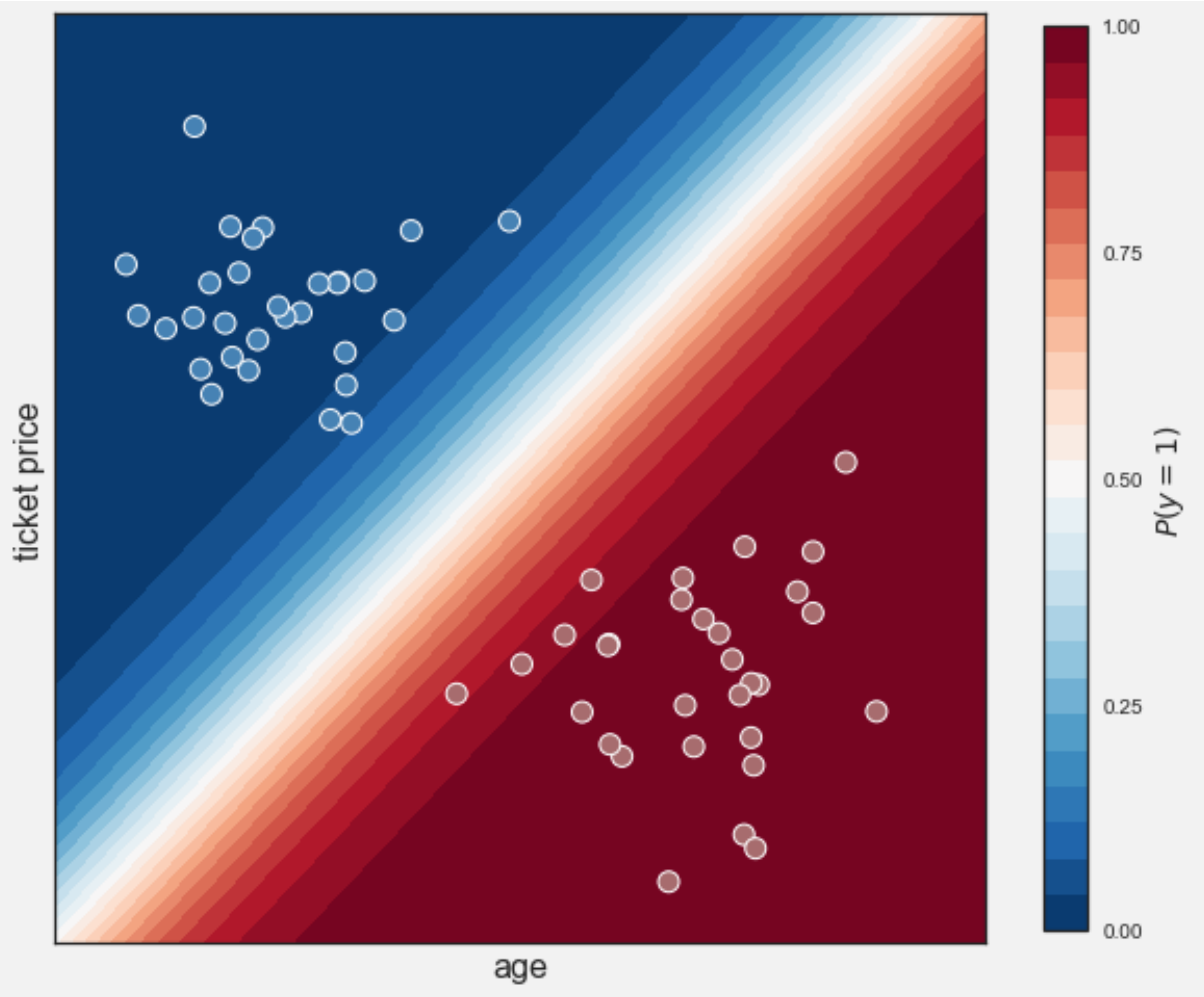


x_1

Logistic Regression with Many Features

- Multiple feature logistic regression: $p(y = 1 \mid x) = \mathrm{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$





Get stoked for CSCI 4831 (ML)

- Turns out we usually write this in a different way!

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \vec{\beta} \cdot \vec{x} = \beta^T x$$

$1 \times p \; X \; p \times 1 \rightarrow 1 \times 1$

- Prereqs: Algorithms, 3022, and Lin. Alg.

Final thoughts 1:

- Definition: A **Decision boundary** is a boundary that divides the feature space into the part that predicts one class and the part that predicts the other class. **Example**:

Decision Boundary $p = 0.5$

$odds = \frac{p}{1-p} = \frac{0.5}{1-0.5} = \frac{0.5}{0.5} = 1$ ✓

$\log odds = \log 1 = 0$

Find Dec. Boundary.

$\log odds = 0 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ defines a plane.

ex. 2 features $\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$ solve for x_2! $x_2 = \frac{-\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_1$

line

3 feature ... plane

4 features + hyperplane.

Final thoughts 2:

- **Cool properties** of the sigmoid function's derivative: $f'(z) = f(z)(1 - f(z))$

$$\frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right) = \frac{+e^{-z}}{(1+e^{-z})^2}$$

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

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

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

$$= f(z)\left(\frac{e^{-z}}{1+e^{-z}}\right) = f(z)(1 - f(z))$$