Learning Objectives

Learn about the **Logistic Regression** classifier

Compare **Logistic Regression** and **Naive Bayes**

Probabilistic Binary Classification

Given: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ training examples

$$\mathbf{x}_i \in \mathbb{R}^D \qquad y_i \in \{0, 1\}$$

Goal: Given new data $\mathbf{x}$, predict its label y

Probabilistic Binary Classification

Given: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ training examples

$$\mathbf{x}_i \in \mathbb{R}^D \qquad y_i \in \{0, 1\}$$

Goal: Given new data $\mathbf{x}$, predict its label y

For each class c, estimate

$$p(y = c \mid \mathbf{x}, \mathcal{D})$$

Assign $\mathbf{x}$ to the class with highest probability

$$\hat{y} = \arg\max_c p(y = c \mid \mathbf{x}, \mathcal{D})$$

Generative vs. Discriminative Models

How do we model/estimate these conditional probabilities?

Generative:

- Model the joint probability distribution $p(\mathbf{x}, y)$.
- Make assumptions about relationship between $\mathbf{x}$ and y
- Make assumptions about data itself
- **Last Time**: Naive Bayes

Discriminative:

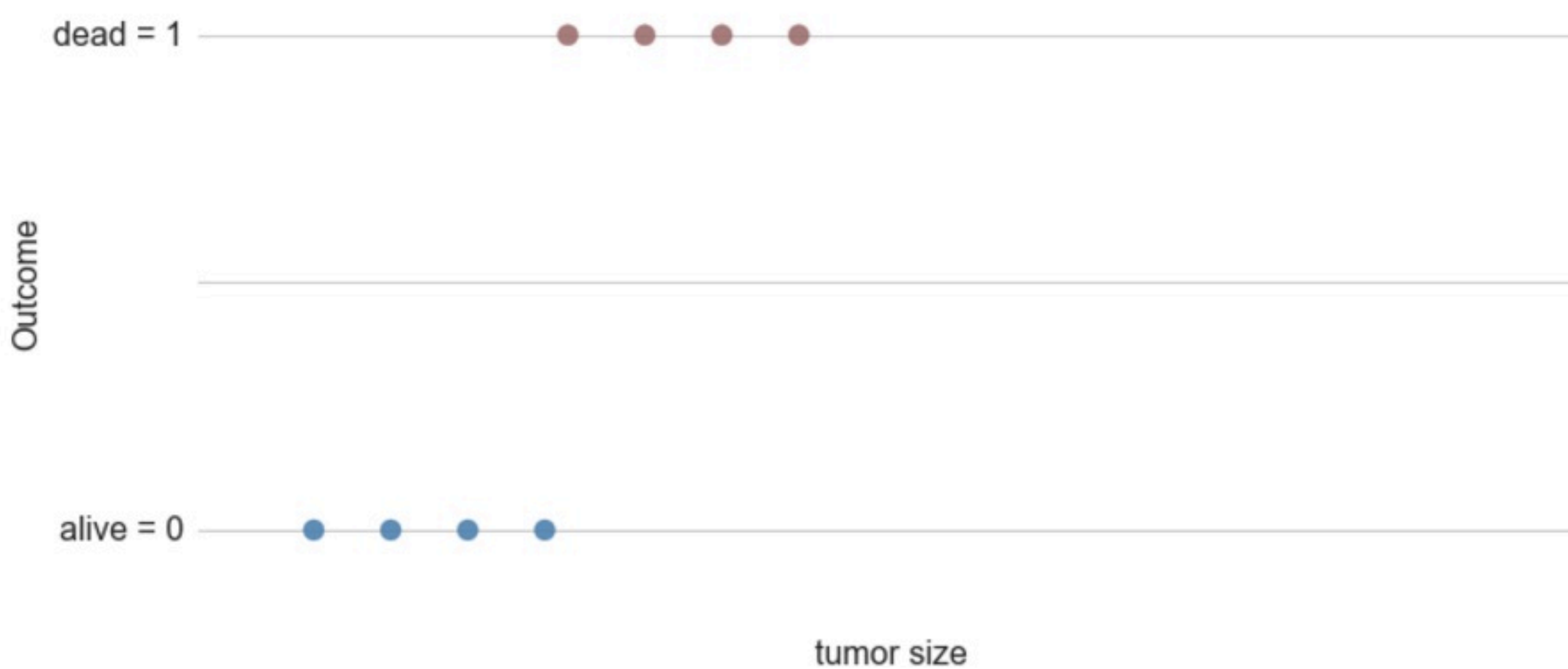
- Model only conditional relationship $p(y \mid \mathbf{x})$
- **Today**: Logistic Regression

Logistic Regression

- Simplest discriminative model
- Does not make strong assumptions about data
- Works well on medium size data sets
- Fairly easy to train

A Simple Example

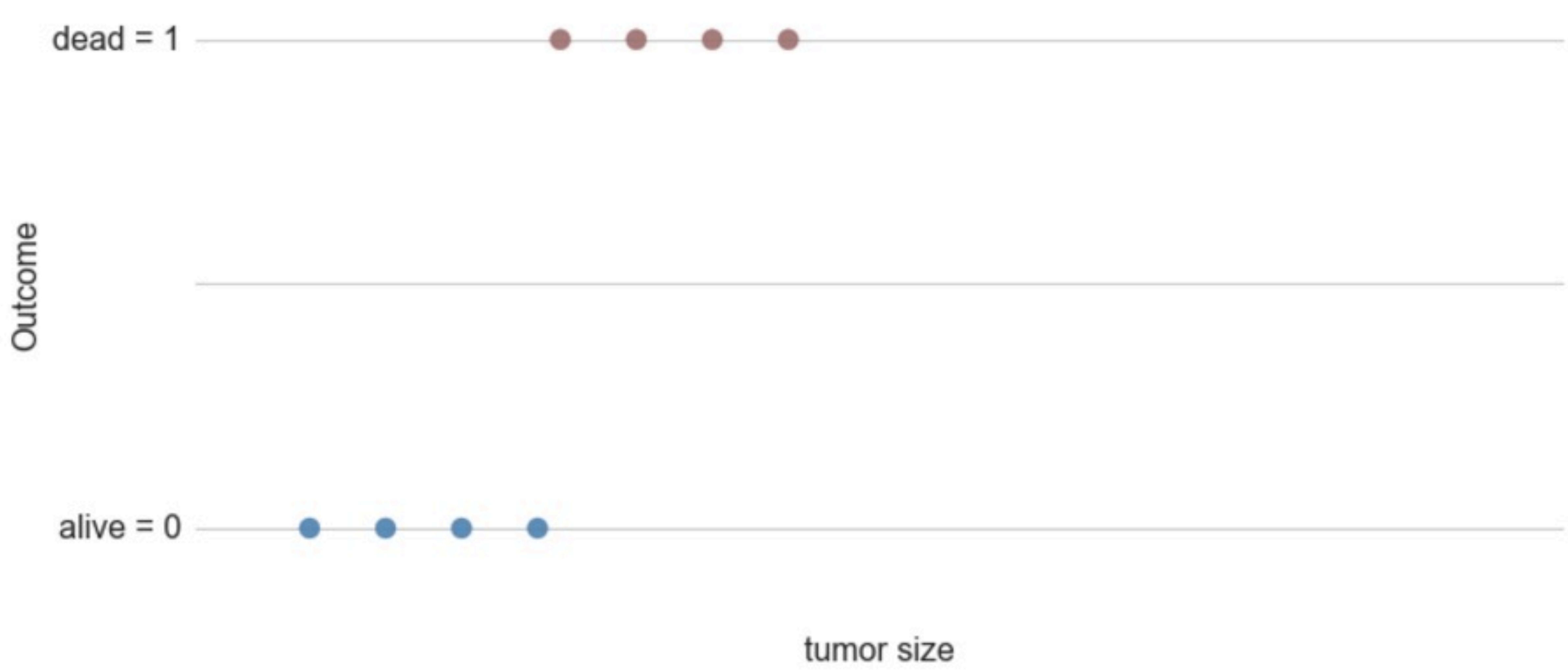
Suppose you track patients in a cancer study and record, among other things, the size of their tumor at the start of the study and whether they were alive at the end of the study



A Simple Example

Single feature: $x_1 = \text{tumor size}$

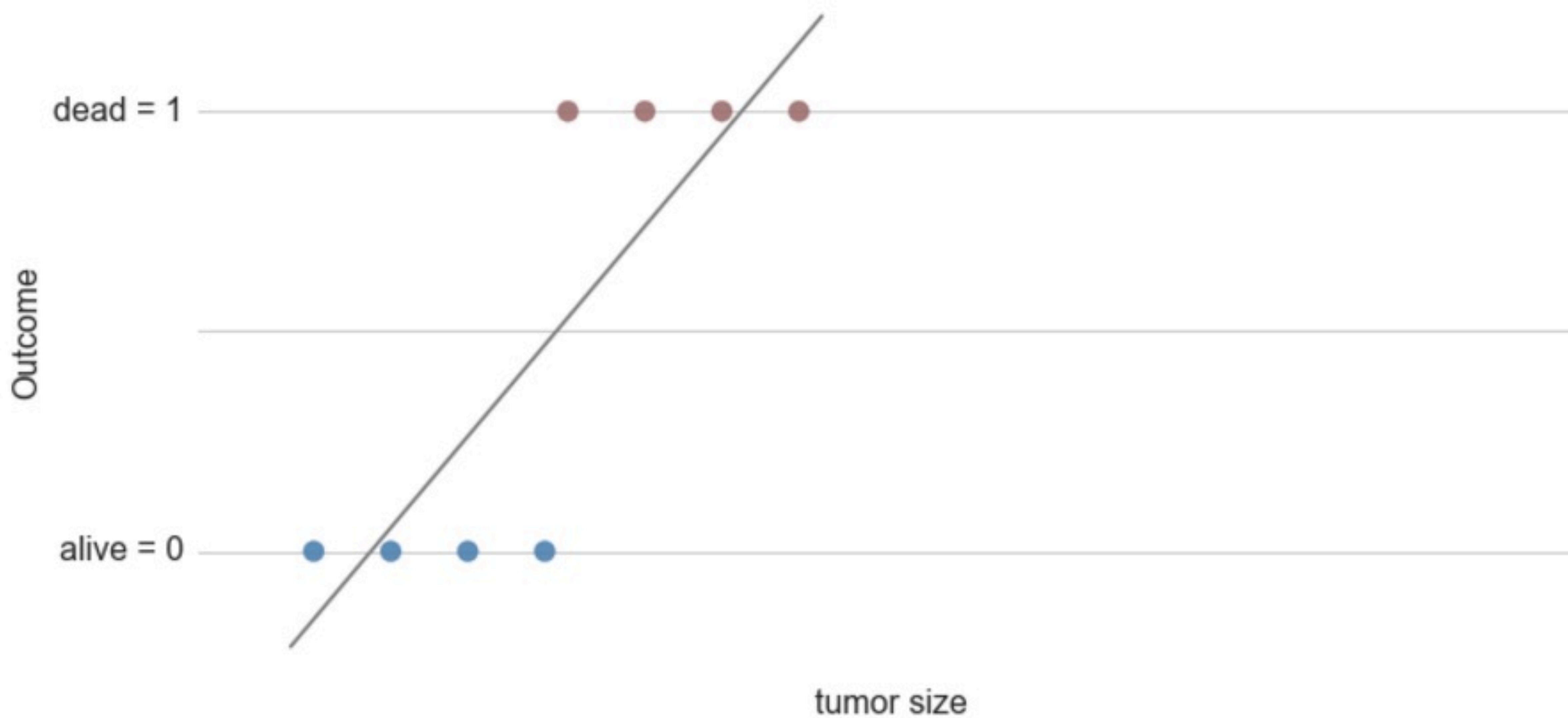
How do we model $p(y \mid x_1, \mathcal{D})$?



A Simple Example

Single feature: x_1 = tumor size

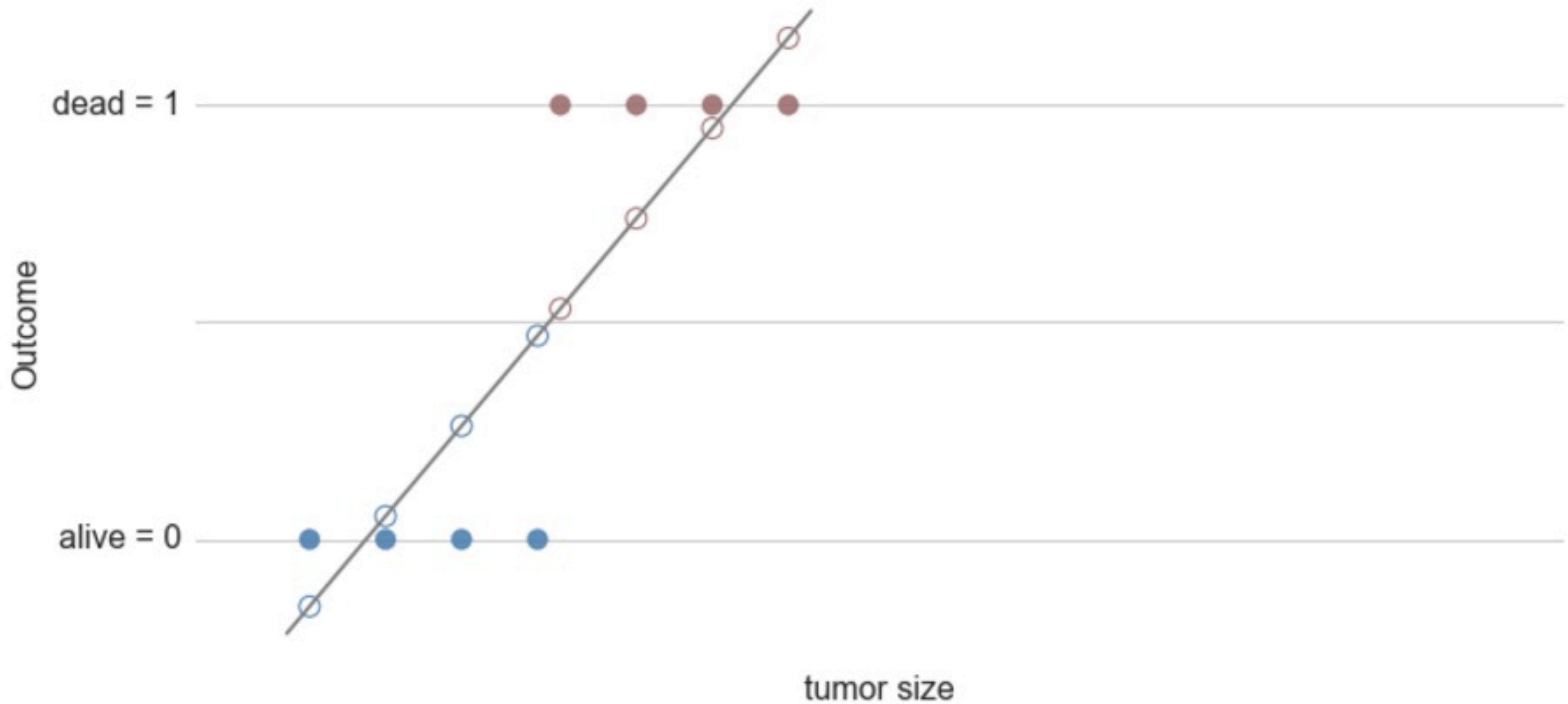
Idea: Linear Regression $$p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$$



A Simple Example

Single feature: x_1 = tumor size

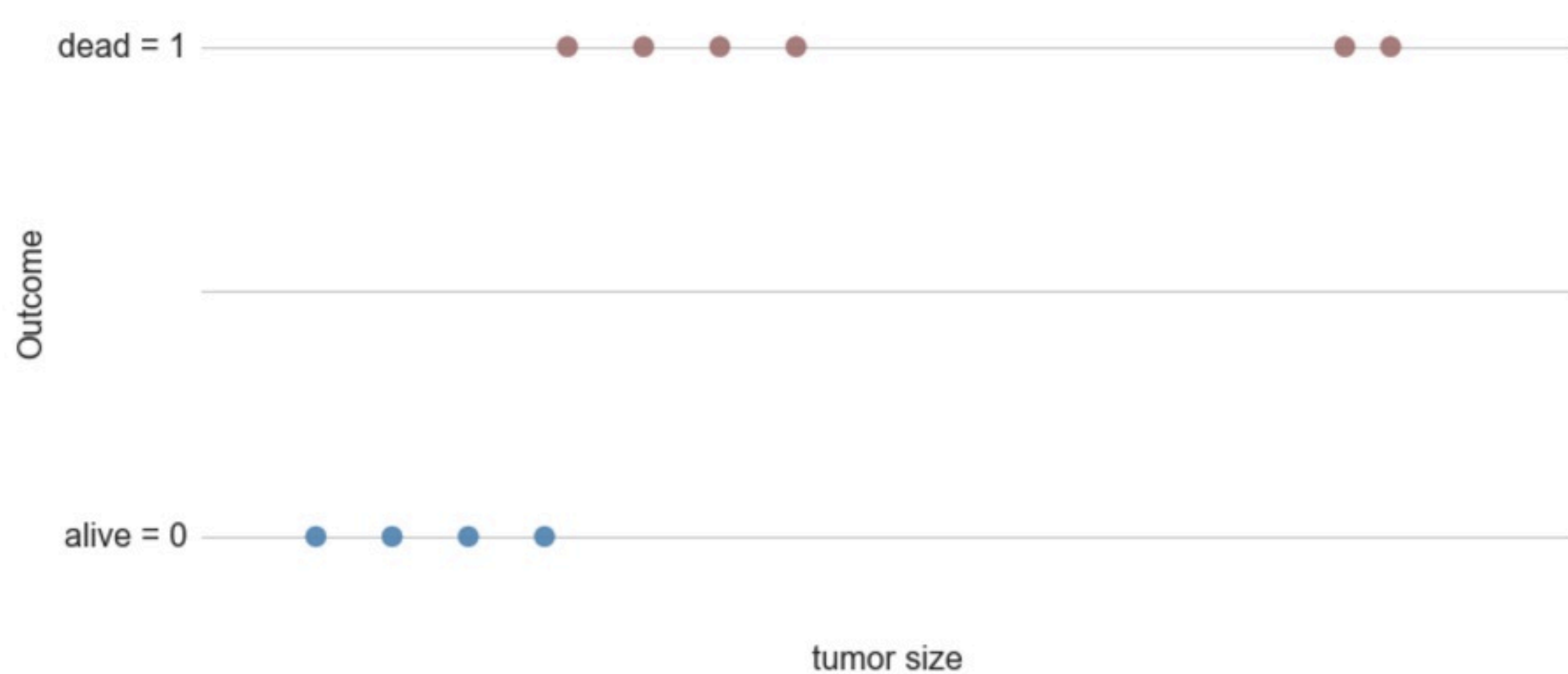
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

Single feature: x_1 = tumor size

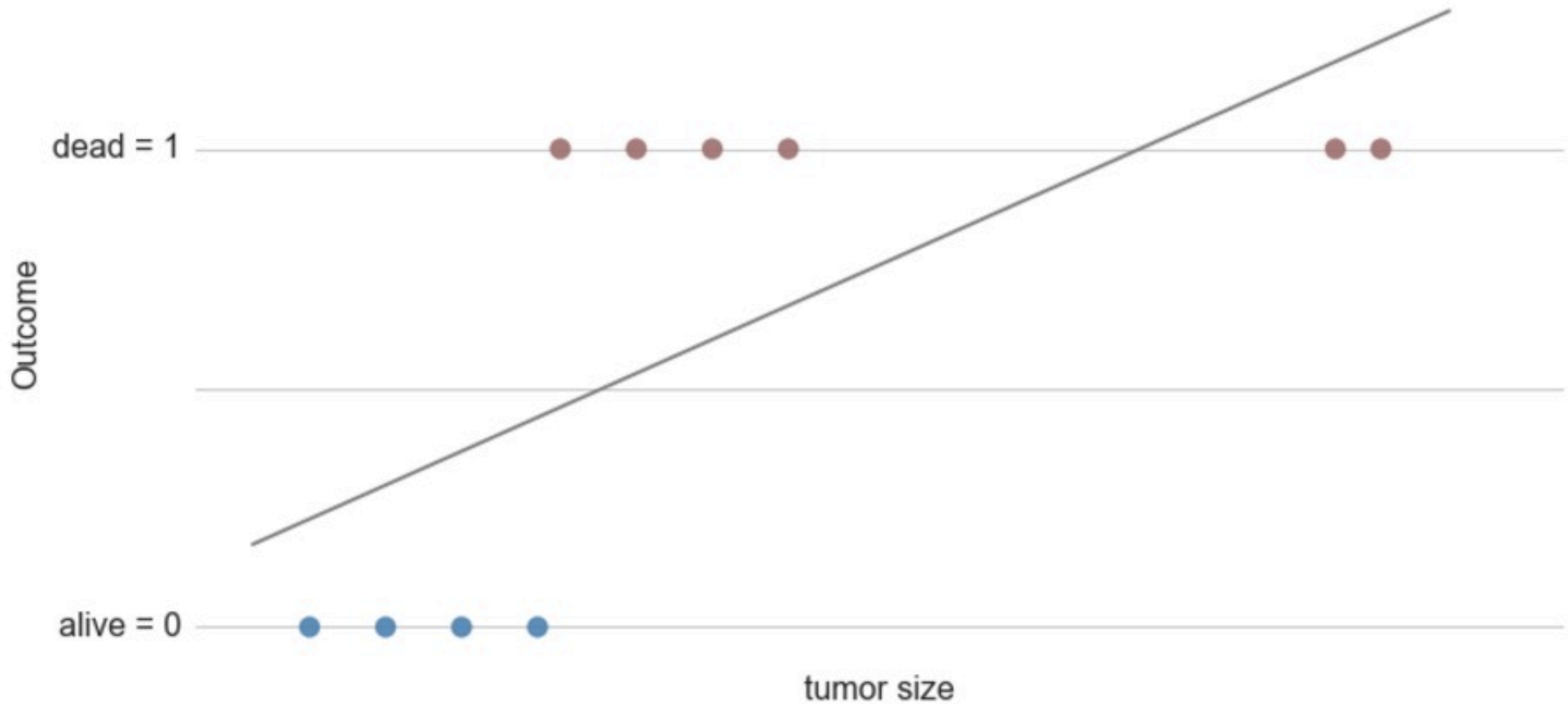
Idea: Linear Regression $$p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$$



A Simple Example

Single feature: x_1 = tumor size

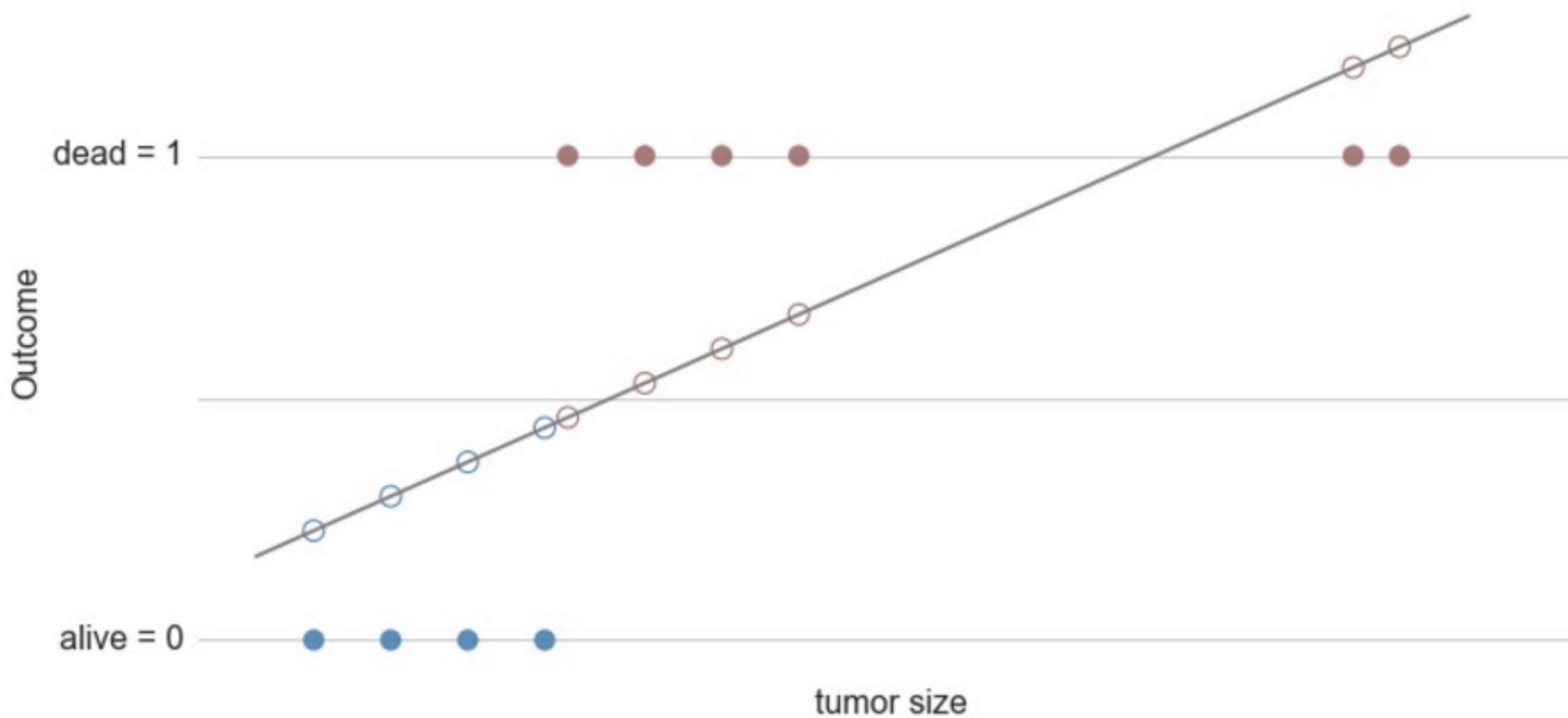
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

Single feature: x_1 = tumor size

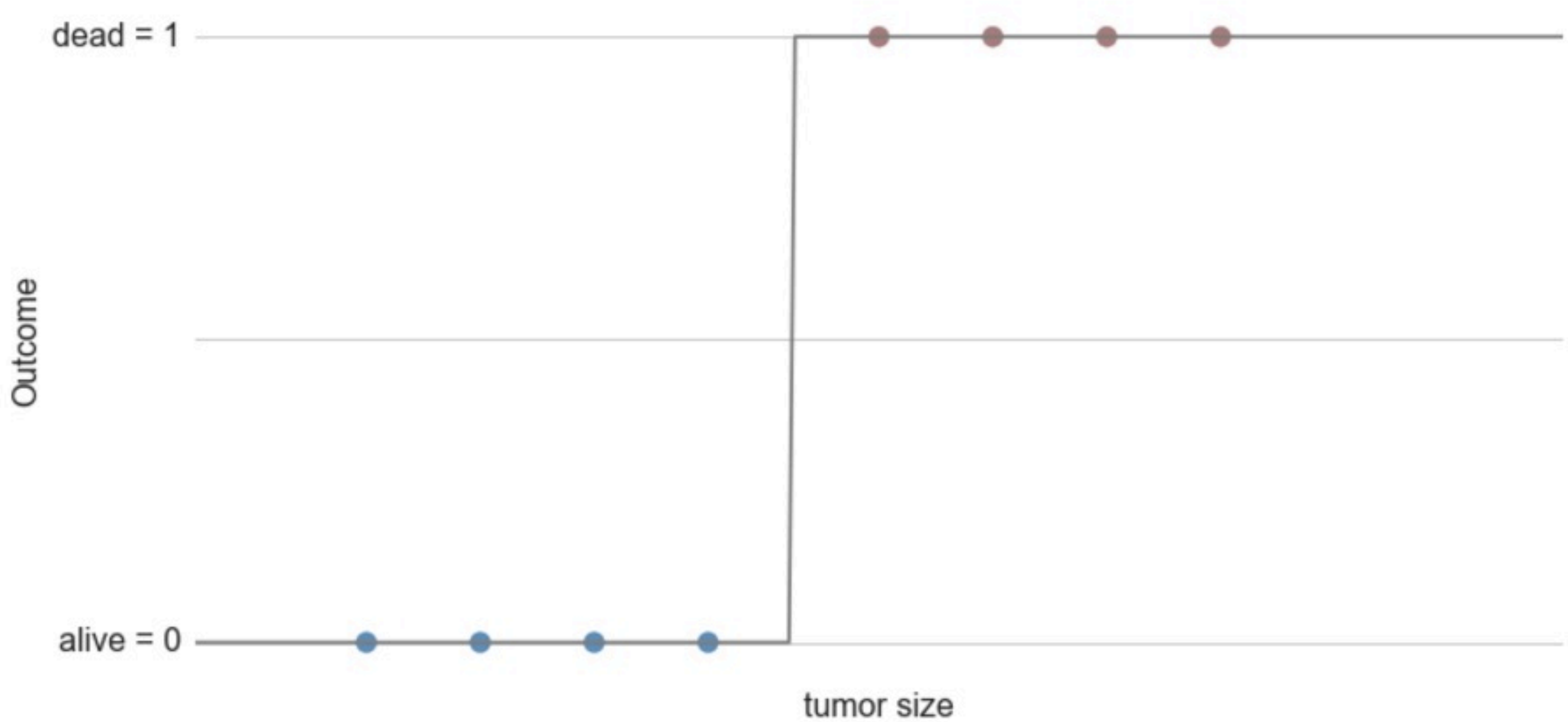
Idea: Linear Regression $$p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$$



A Simple Example

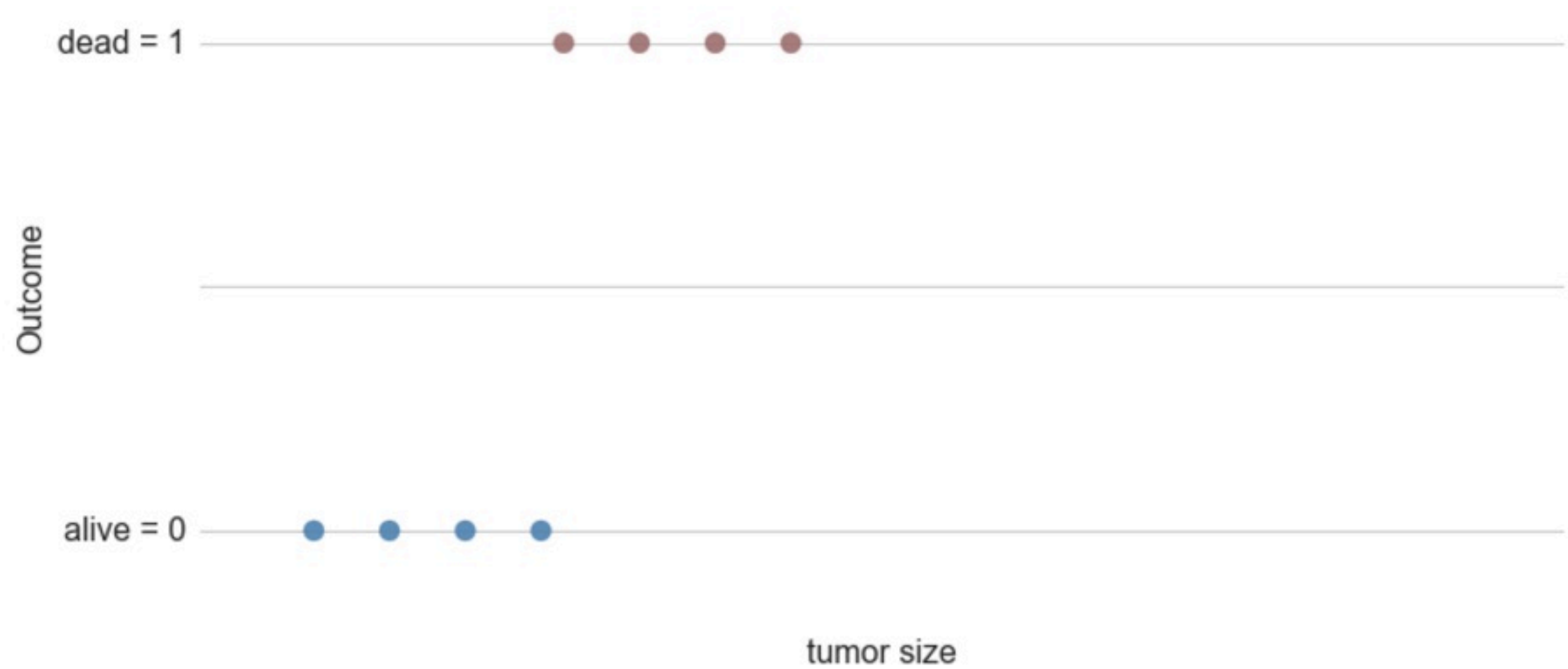
Single feature: x_1 = tumor size

Idea: Perceptron $$p(y = 1 \mid x_1; \mathbf{w}) = \begin{cases} 1 & \text{if } \; w_0 + w_1 x_1 > 0 \\ 0 & \text{else} \end{cases}$$



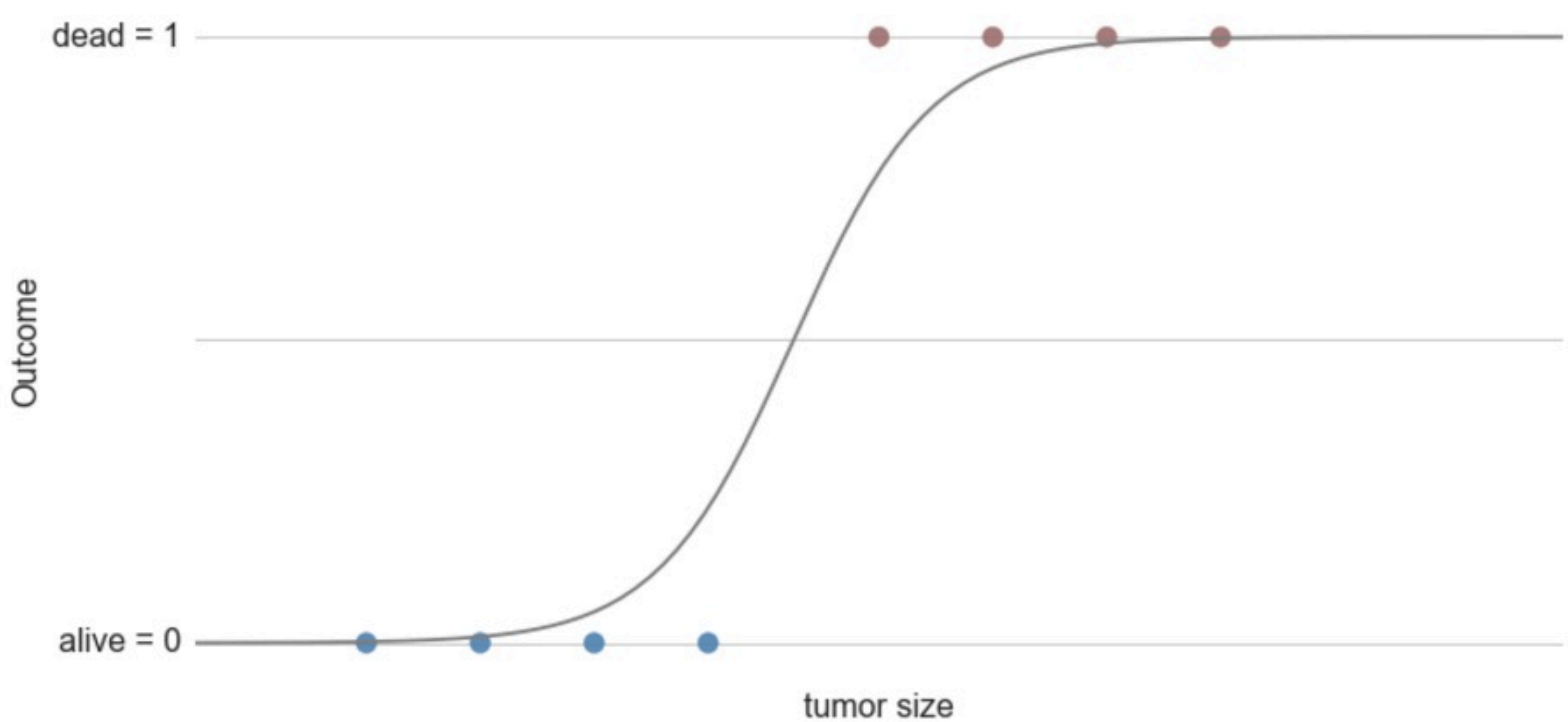
A Simple Example

Need something that behaves more like a probability ...



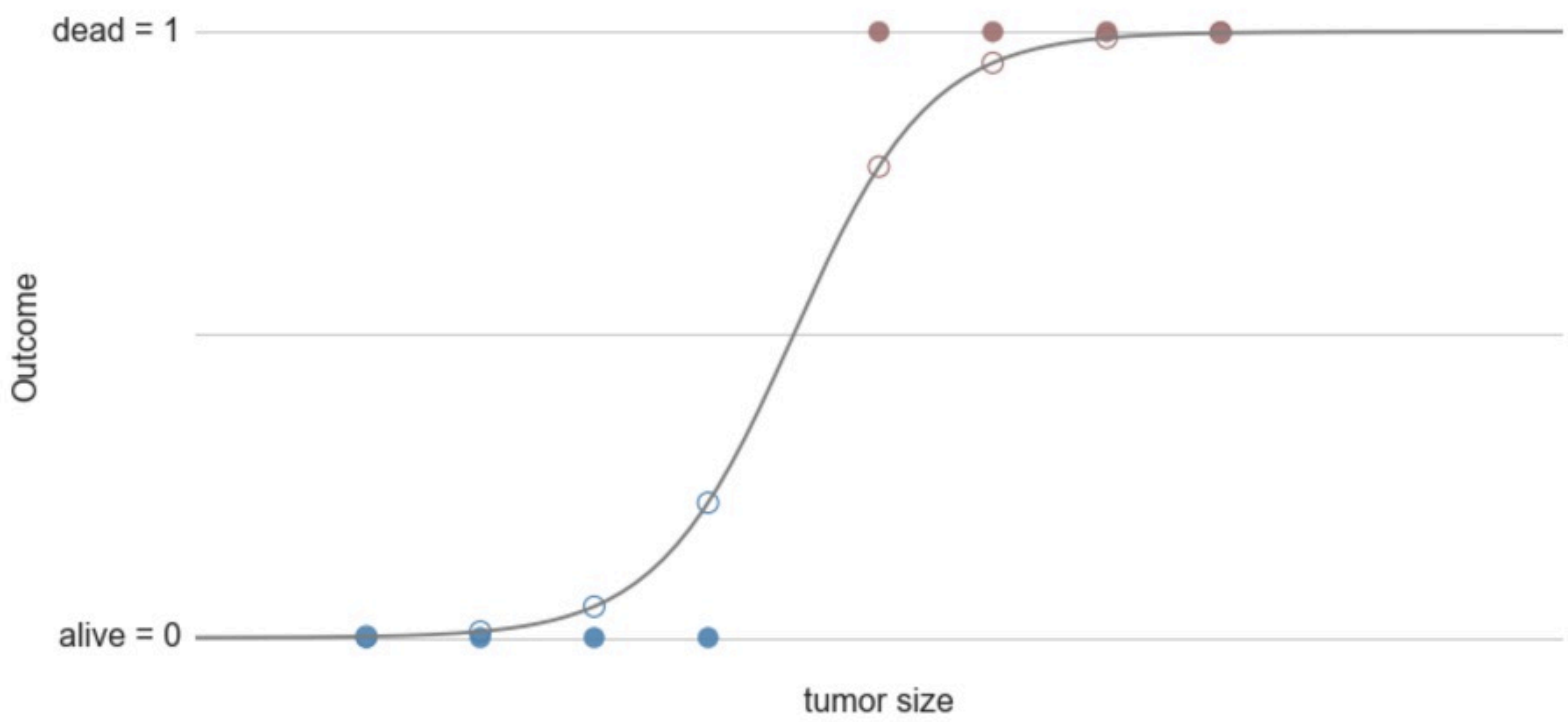
Enter the sigmoid Function

$$p(y = 1 \mid x_1; \mathbf{w}) = \mathrm{sigm}(w_0 + w_1 x_1) = \frac{1}{1 + \exp[-(w_0 + w_1 x_1)]}$$



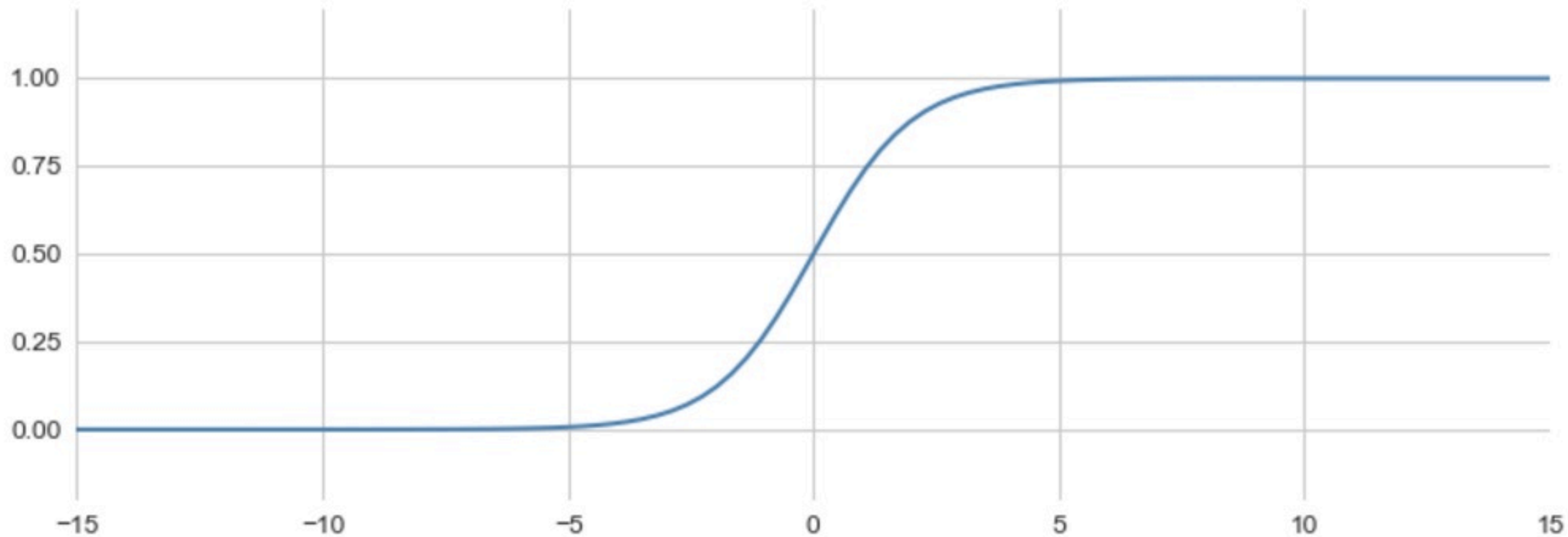
Enter the sigmoid Function

$$p(y = 1 \mid x_1; \mathbf{w}) = \mathrm{sigm}(w_0 + w_1 x_1) = \frac{1}{1 + \exp[-(w_0 + w_1 x_1)]}$$



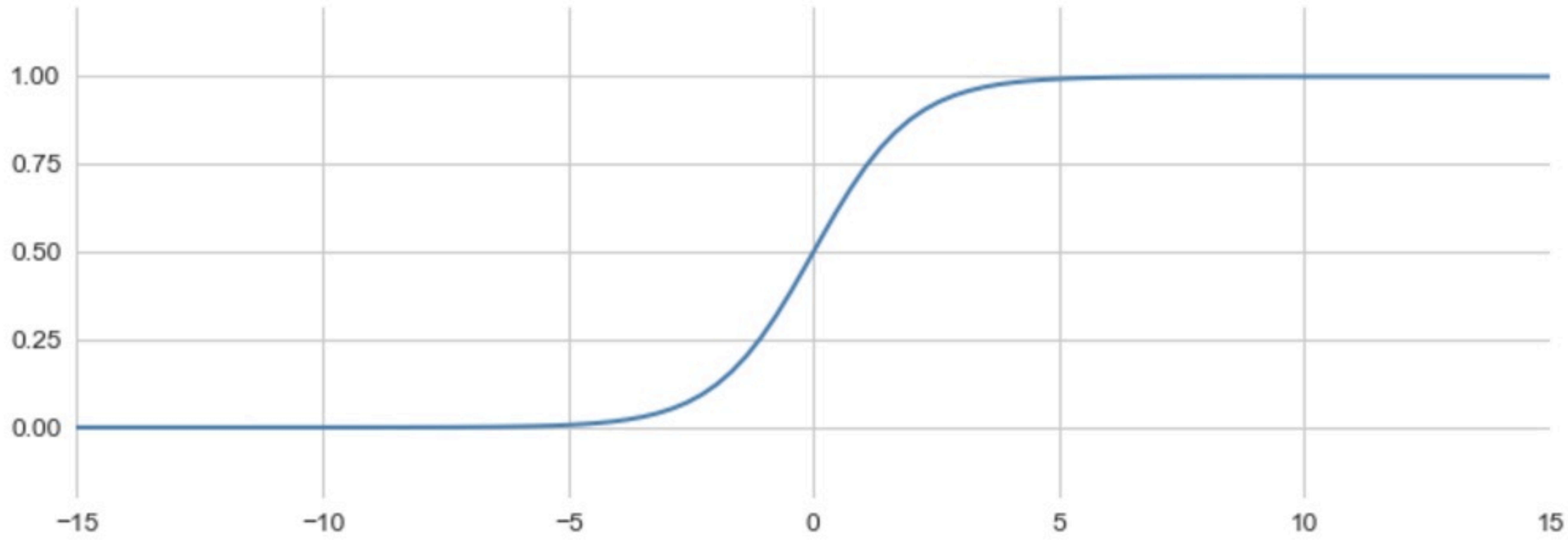
Enter the sigmoid Function

$$\text{sigm}(z) = \frac{1}{1 + \exp[-z]}$$



It Has Everything!

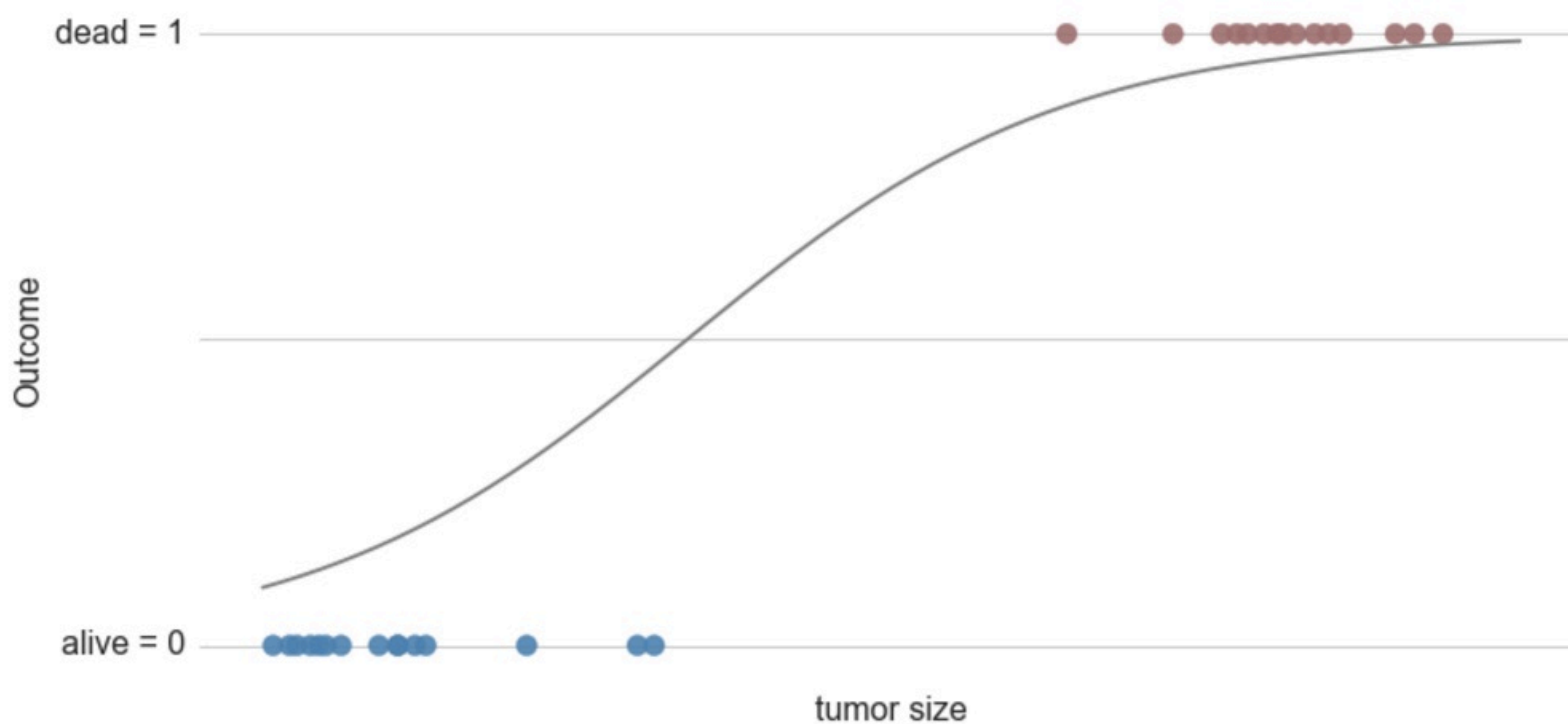
- Behaves like a probability ($0 < \mathrm{sigm}(z) < 1$)
- Distinguishes between points
- It's really smooth (important later)



The Plan

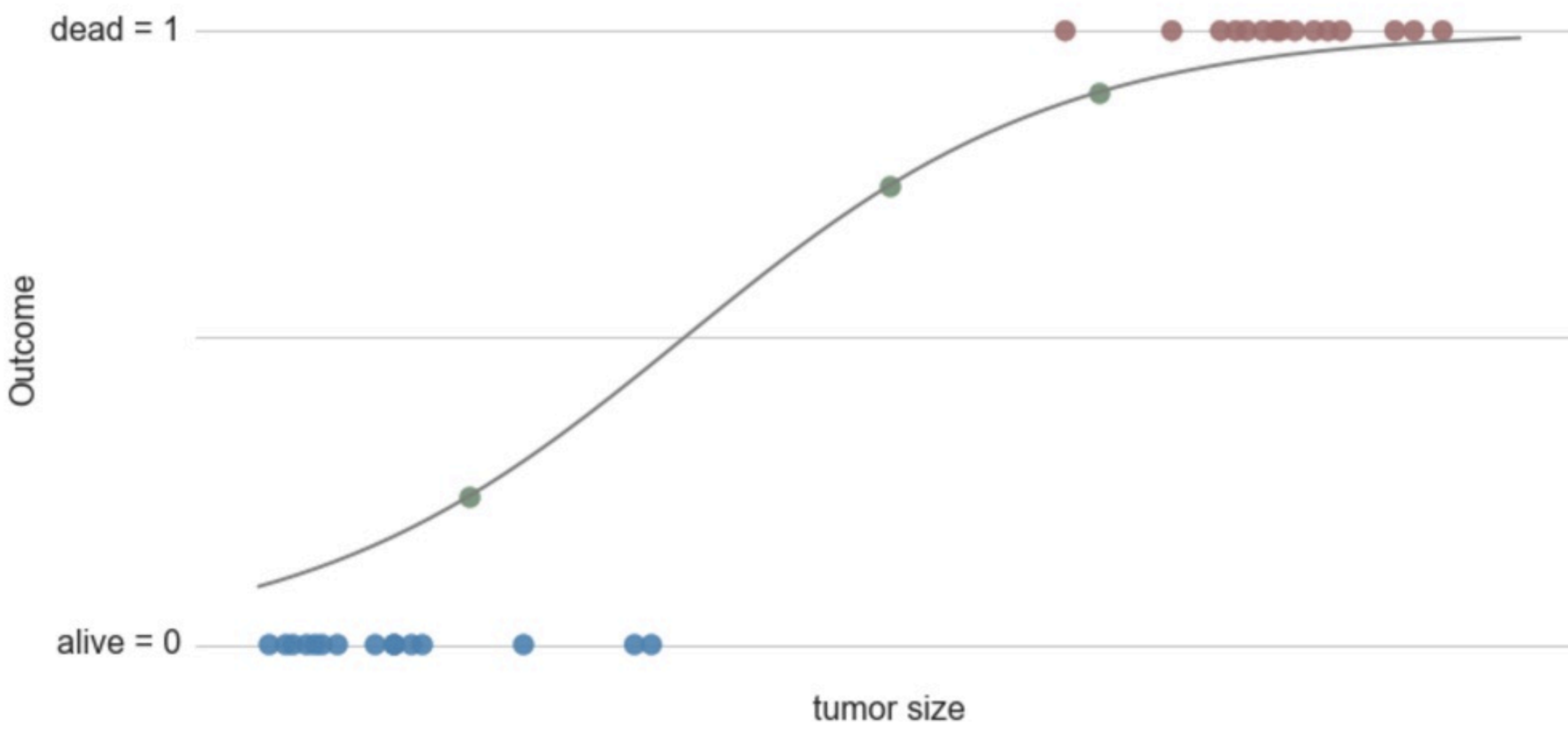
$$p(y = 1 \mid x_1; \hat{\mathbf{w}}) = \mathrm{sigm}(\hat{w}_0 + \hat{w}_1 x_1)$$

- Learn the weights $\hat{\mathbf{w}}$ from training data (next lecture!)



The Plan

Classify test sample x as $y = 1$ if
$\mathrm{sigm}(\hat{w}_0 + \hat{w}_1 x) > 0.5$
else classify as $y = 0$



The Plan

So far we've looked at a single-feature continuous example

Naturally generalizes to many features: $\mathbf{x} = [x_1, x_2, \dots, x_D]^T$

$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \mathrm{sigm}(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D)$$

New dot-product notation:

$$w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D = \mathbf{w}^T \mathbf{x} = \mathbf{w} \cdot \mathbf{x}$$

$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \mathrm{sigm}(\mathbf{w}^T \mathbf{x})$$

With this notation we prepend $\mathbf{x}$ with a 1, $\mathbf{x} = [1, x_1, \dots, x_D]^T$

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

$$
\begin{aligned}
z &= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \\
& \quad 0.1 + 2 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1 = 2.1
\end{aligned}
$$

$$
p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = \frac{1}{1 + \exp[-2.1]} = 0.89
$$

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

$$\begin{aligned} z &= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \\ & 0.1 + 2 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1 = 2.1 \end{aligned}$$

$$p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.89, \quad p(y = 0 \mid \mathbf{x}, \hat{\mathbf{w}}) = 1 - 0.89 = 0.11$$

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

$$
\begin{aligned}
z &= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \\
& \quad 0.1 + 2 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1 = 2.1
\end{aligned}
$$

$$p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.89, \quad p(y = 0 \mid \mathbf{x}, \hat{\mathbf{w}}) = 1 - 0.89 = 0.11$$

Since $p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.89 > 0.5$ predict SPAM!

Spam vs. Ham Example

$feature$	$bias$	$"\ viagra\ "$	$"\ mom\ "$	$"\ work\ "$	$"\ nigeria\ "$
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	-1.0	-0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

Pause and Ponder: What do the signs and magnitudes of the weights tell you about their associated features and how they affect the binary classification problem?

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

Intuition: Large weights mean associated features have large effect on overall classification. The signs on the weights tell you whether the feature is particularly important for the $y = 0$ or $y = 1$ class.

Caveat: Need to think about the relative sizes of the features before drawning meaningful conclusions.

Spam vs. Ham Example

$feature$	$bias$	$"\ viagra\ "$	$"\ mom\ "$	$"\ work\ "$	$"\ nigeria\ "$
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, nigeria\}$

Alternatively...

$$z = \mathbf{w}^T\mathbf{x} = [0.1\ \ 2.0\ \ -1.0\ \ -0.5\ \ 3.0]\begin{bmatrix}1\\0\\1\\0\\1\end{bmatrix} = 0.1 - 1.0 + 3.0 = 2.1$$

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, work, viagra, mom\}$

Text Model Interlude:

Binary Text Model: Feature $x_i = 1$ if word i is **present** in email

Bag-of-Words: Feature $x_i =$ # of times word i appears in message

EFY: Pause .... work example

Spam vs. Ham Example

feature	*bias*	" *viagra* "	" *mom* "	" *work* "	" *nigeria* "
coef	w_0	w_1	w_2	w_3	w_4
weight	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{mom, work, viagra, mom\}$

$$
\begin{aligned}
z \quad = \quad & w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \\
& 0.1 + 2 \cdot 1 - 1 \cdot 2 - 0.5 \cdot 1 + 3 \cdot 0 = -0.4
\end{aligned}
$$

$$p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.40, \quad p(y = 0 \mid \mathbf{x}, \hat{\mathbf{w}}) = 1 - 0.40 = 0.60$$

Since $p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.40 \leq 0.5$ predict HAM!

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	−1.0	−0.5	3.0

Email: $\mathbf{x} = \{\}$

Yes, this is an empty email

Spam vs. Ham Example

$feature$	$bias$	" $viagra$ "	" mom "	" $work$ "	" $nigeria$ "
$coef$	w_0	w_1	w_2	w_3	w_4
$weight$	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{\}$

$$\begin{aligned} z &= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \\ & 0.1 + 2 \cdot 0 - 1 \cdot 0 - 0.5 \cdot 0 + 3 \cdot 0 = 0.1 \end{aligned}$$

$$p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.52, \quad p(y = 0 \mid \mathbf{x}, \hat{\mathbf{w}}) = 1 - 0.40 = 0.48$$

Since $p(y = 1 \mid \mathbf{x}, \hat{\mathbf{w}}) = 0.52 > 0.5$ predict SPAM!

Spam vs. Ham Example

feature	*bias*	" *viagra* "	" *mom* "	" *work* "	" *nigeria* "
coef	w_0	w_1	w_2	w_3	w_4
weight	0.1	2.0	–1.0	–0.5	3.0

Email: $\mathbf{x} = \{\}$

Notice that when all of the features are zero, the only thing affecting the probability is the bias.

In a sense, the bias encodes something similar to a prior probability of a class.

Generative vs Discriminative Models Revisited

- Generative models tend to make much stronger assumptions, but when their assumptions are correct they tend to dominate
- Discriminative models are more robust because they don't rely on strong assumptions
- Generative models are usually cheaper to train
- Discriminative models do much better with engineered features