HackCU Episode IV

Learn New Skills
Build Something Cool
Meet Awesome People
Discover Job Opportunities

February 24th and 25th 2018

HackCU.org

Logistic Regression

Administrivia

- Homework 2 posted. Good Milestones:
  - **Problems 1 and 2 This Week**
  - Problems 3 and 4 Next Week
- I've added a better unit test for Problem 4 since last night. If you've already checked out the homework assignment, re-checkout tests/tests.py
- There is a Reading Quiz associated with today's lecture.

Previously on CSCI 4022

KNN:

- Find $\mathcal{N}_K(\mathbf{x})$: the set of K training examples nearest to $\mathbf{x}$ using $\|\mathbf{x}_i - \mathbf{x}\|^2$
- Predict $\hat{y}$ to be majority label in $\mathcal{N}_K(\mathbf{x})$
- Admits a probabilistic interpretation of class given data: $p(Y = k \mid \mathbf{x})$

The Perceptron:

- Learn weights $\mathbf{w}$ and b via the Perceptron Algorithm
- Predict $\hat{y}$ via $\hat{y} = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right)$
- Has no probabilistic interpretation

Logistic Regression

- Today we look at a classifier that is similar to the Perceptron, but more flexible
- It also admits a probabilistic interpretation

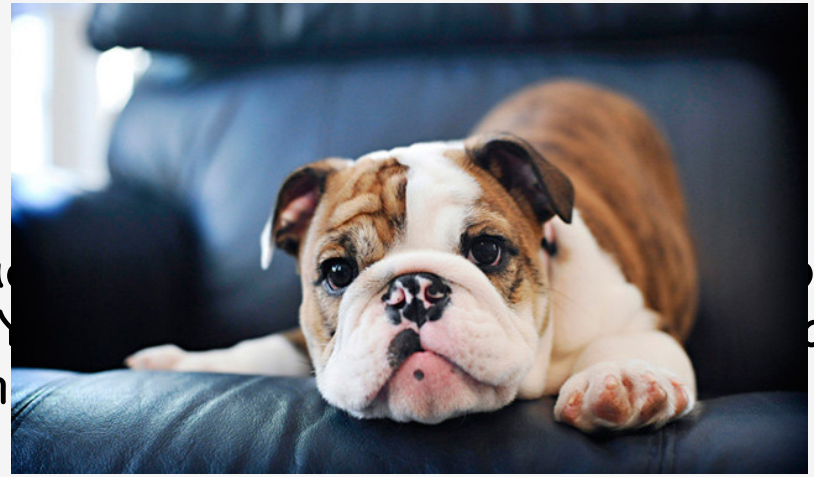
Logistic Regression

- Today we look at a classifier that is similar to the Perceptron, but more flexible
- It also admits a probabilistic interpretation

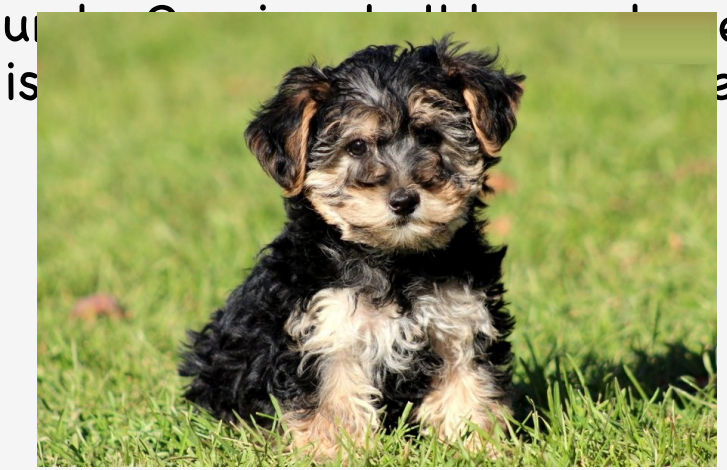
Example: Suppose that you have two dogs that frequently commit dog-crimes when you're not around. One is a bulldog and one is a yorkipoo. Your primary source of evidence of the culprit is how much slobber is left behind at the scene.

Logistic Regression

- Today we look at a classifier that is similar to the Perceptron, but more flexible
- It also admits a probabilistic interpretation



Example: Suppose that you have two dogs that frequ u're not arou e is a yorkipoo. Y f the culprit is ehind at the scen



Logistic Regression

- Today we look at a classifier that is similar to the Perceptron, but more flexible
- It also admits a probabilistic interpretation

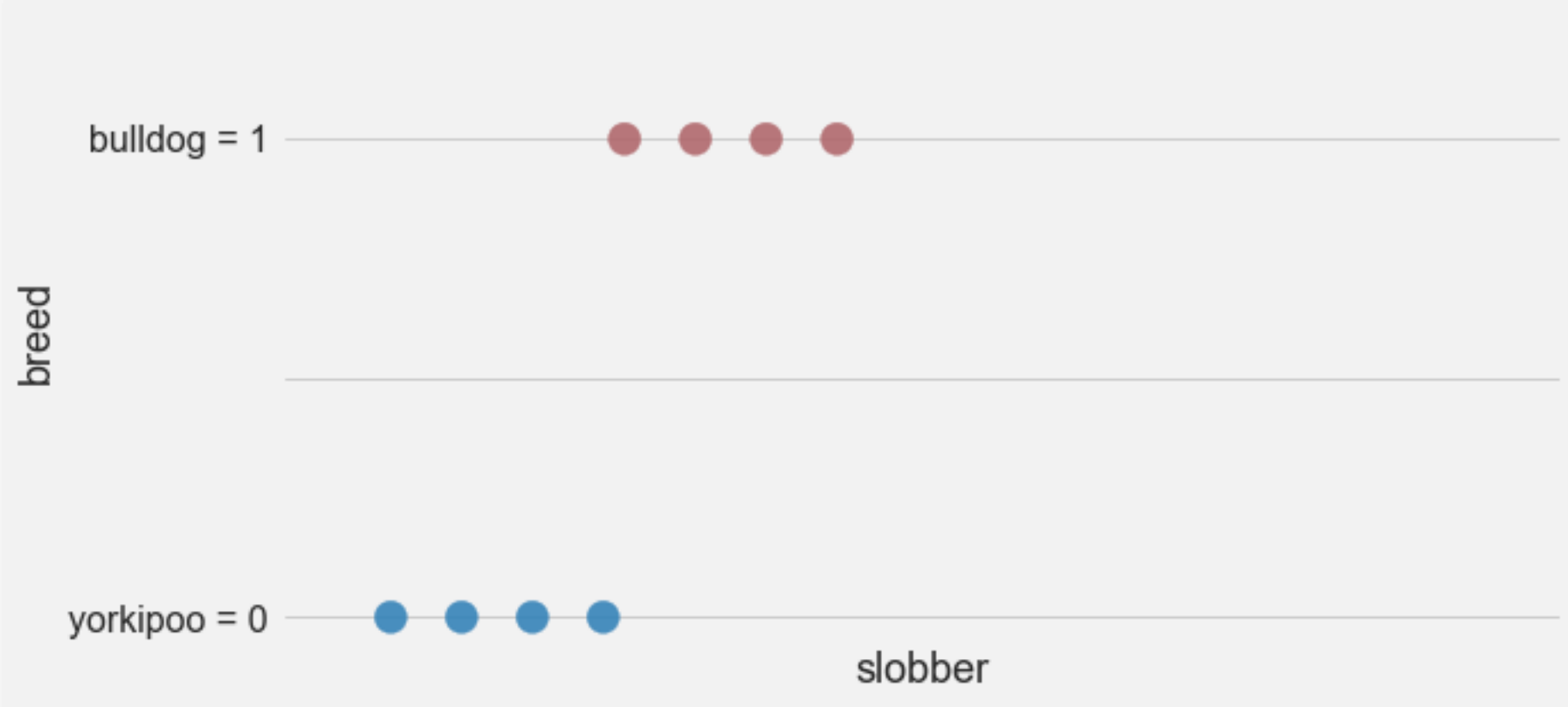
Example: Suppose that you have two dogs that frequently commit dog-crimes when you're not around. One is a bulldog and one is a yorkipoo. Your primary source of evidence of the culprit is how much slobber is left behind at the scene.

Having caught them each in the act multiple times, you have collected some training data where the feature is the amount of slobber and the class is encoded as $y = 0$ when the yorkipoo is guilty and $y = 1$ when the bulldog is guilty.

Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

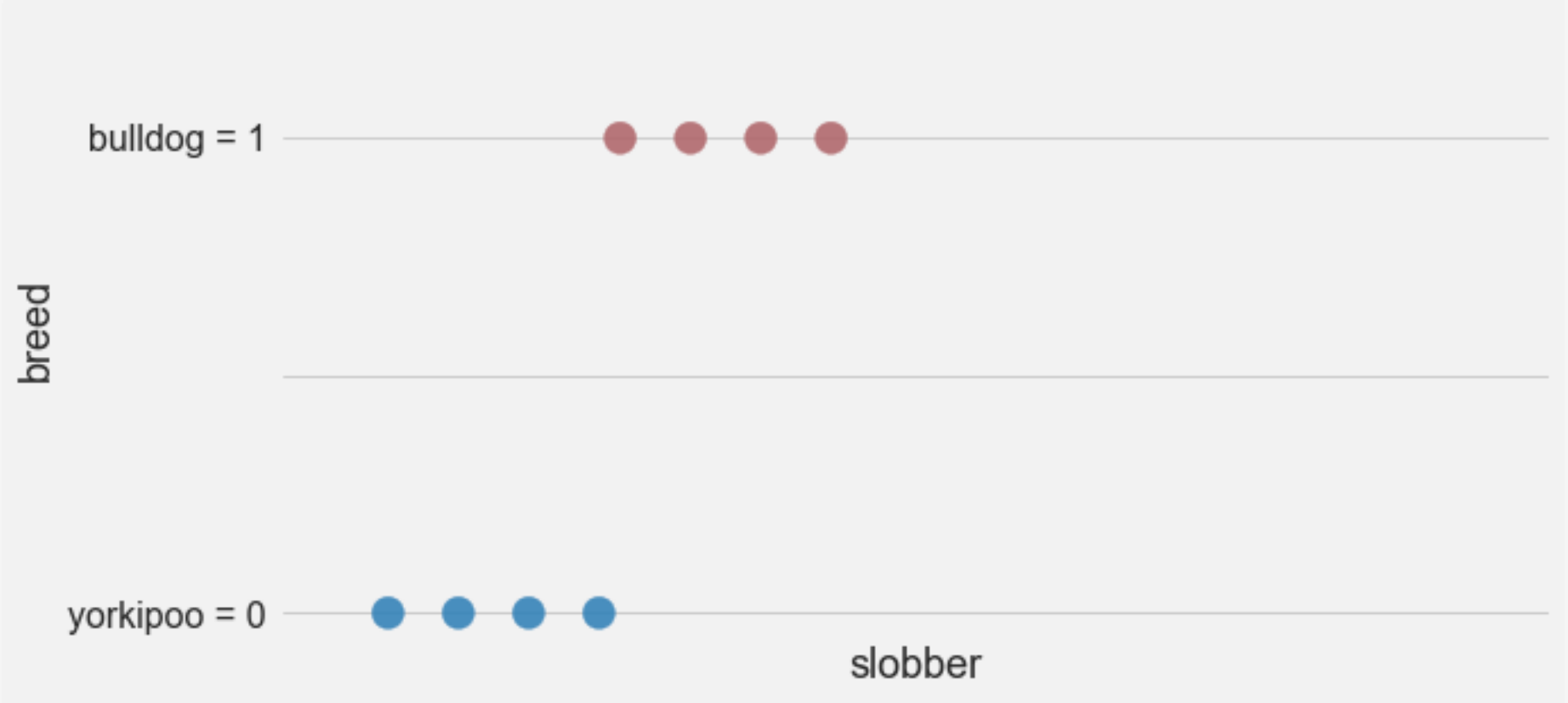
The output from the model is the prediction: $y = \{0, 1\} =$ {yorkipoo, bulldog}



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

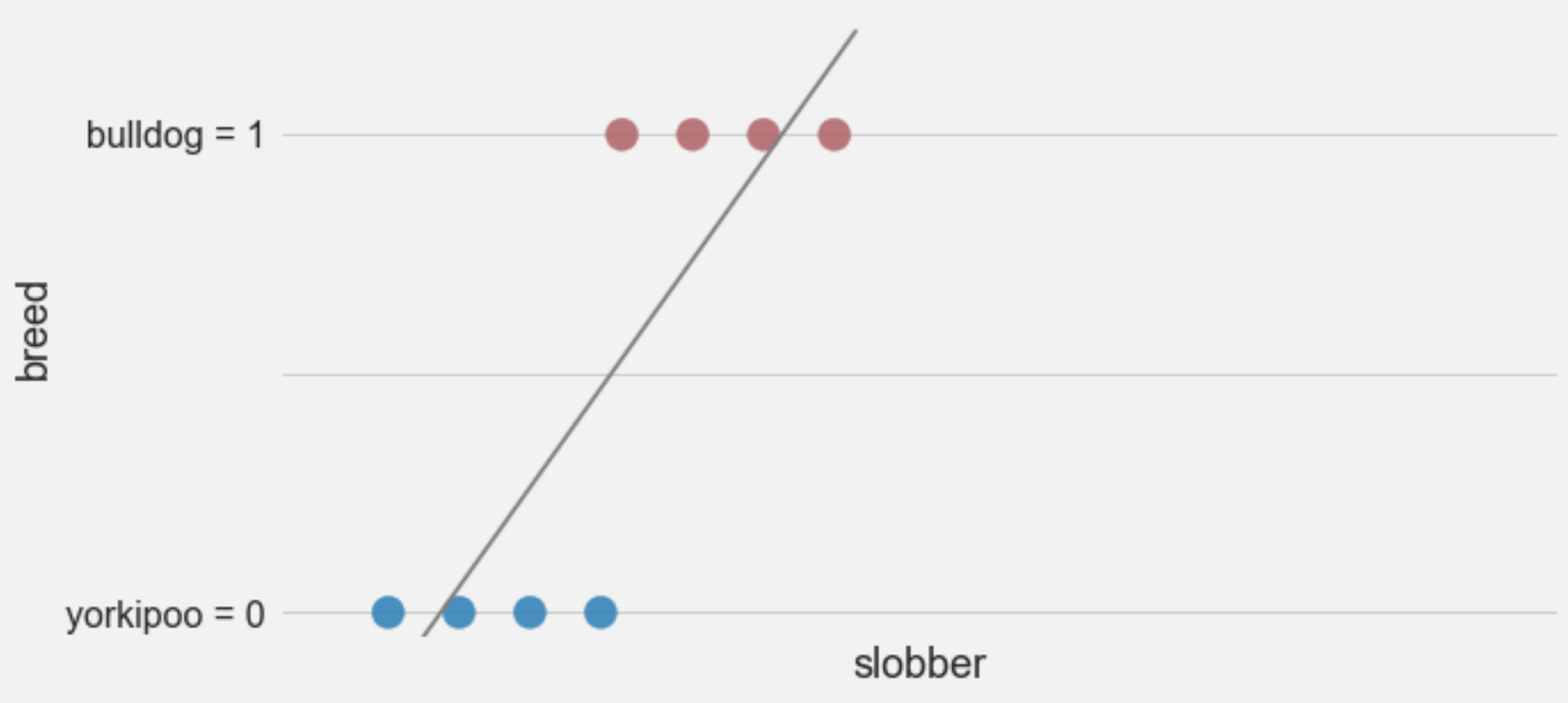
Question: How should we model the relation between feature and response?



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

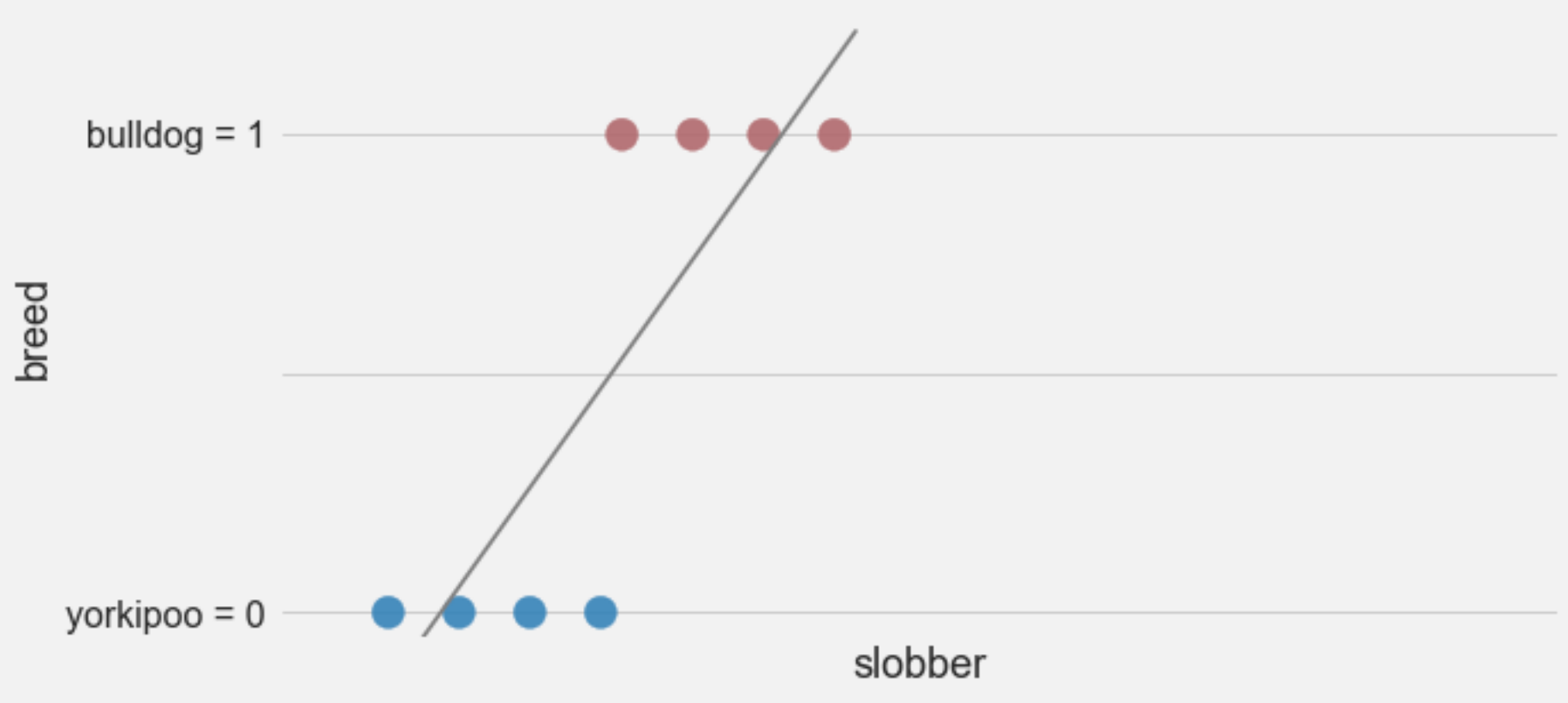
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

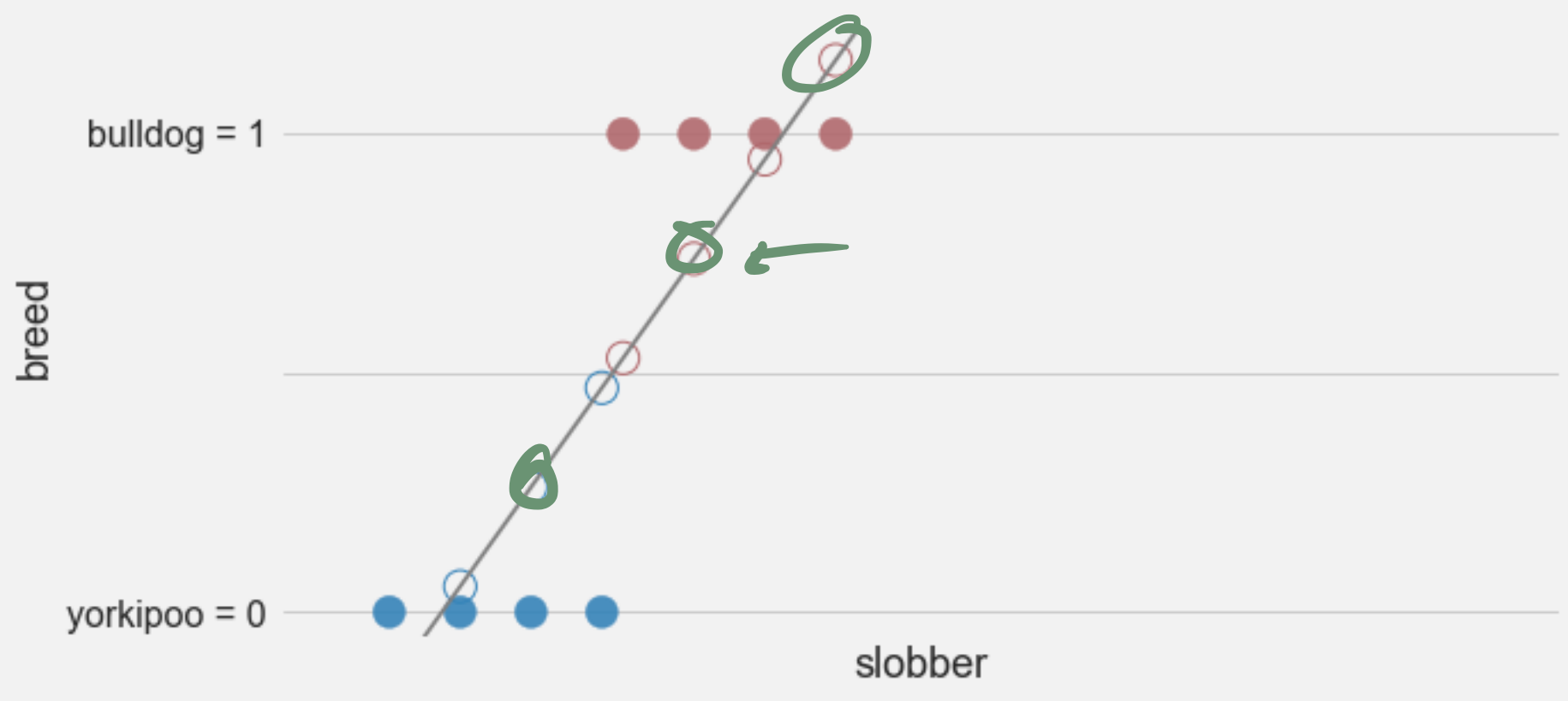
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

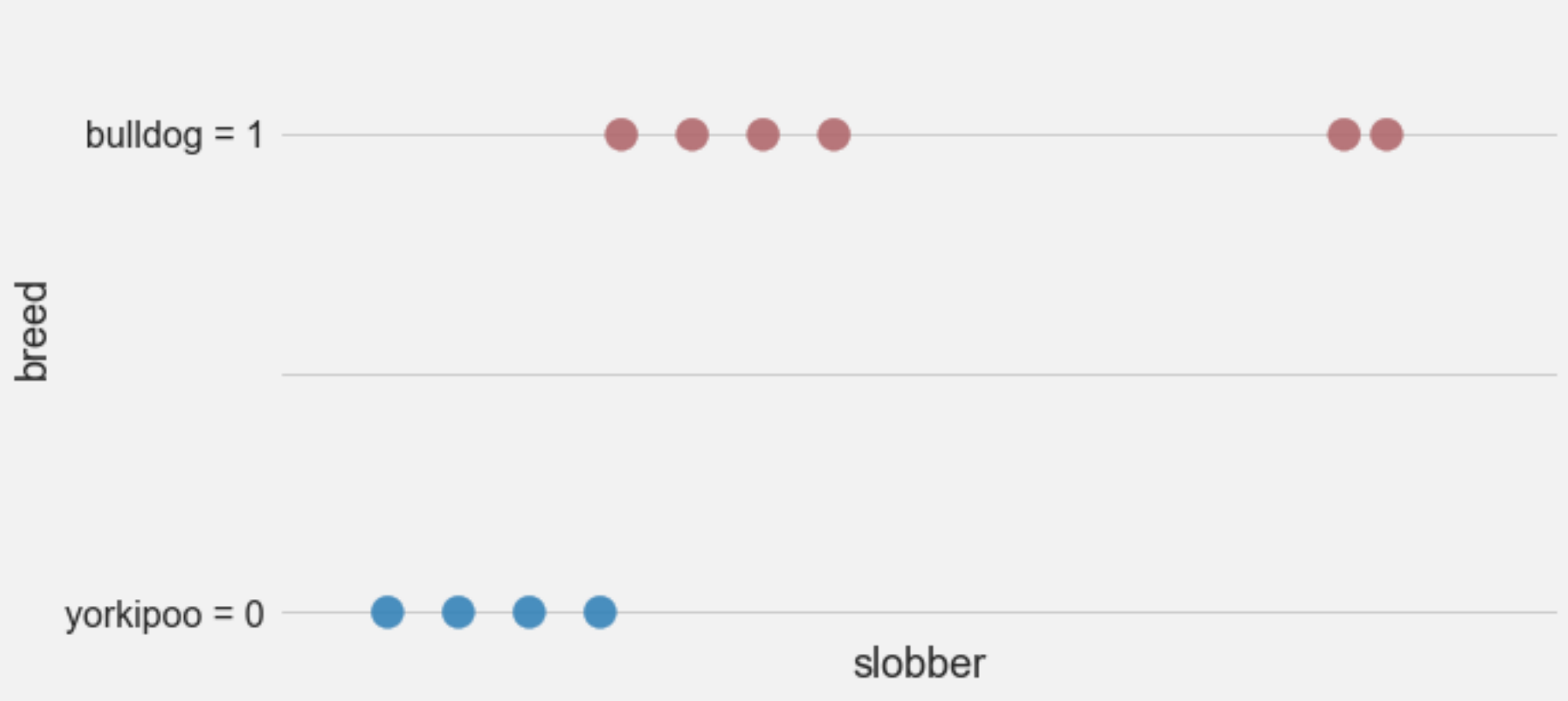
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

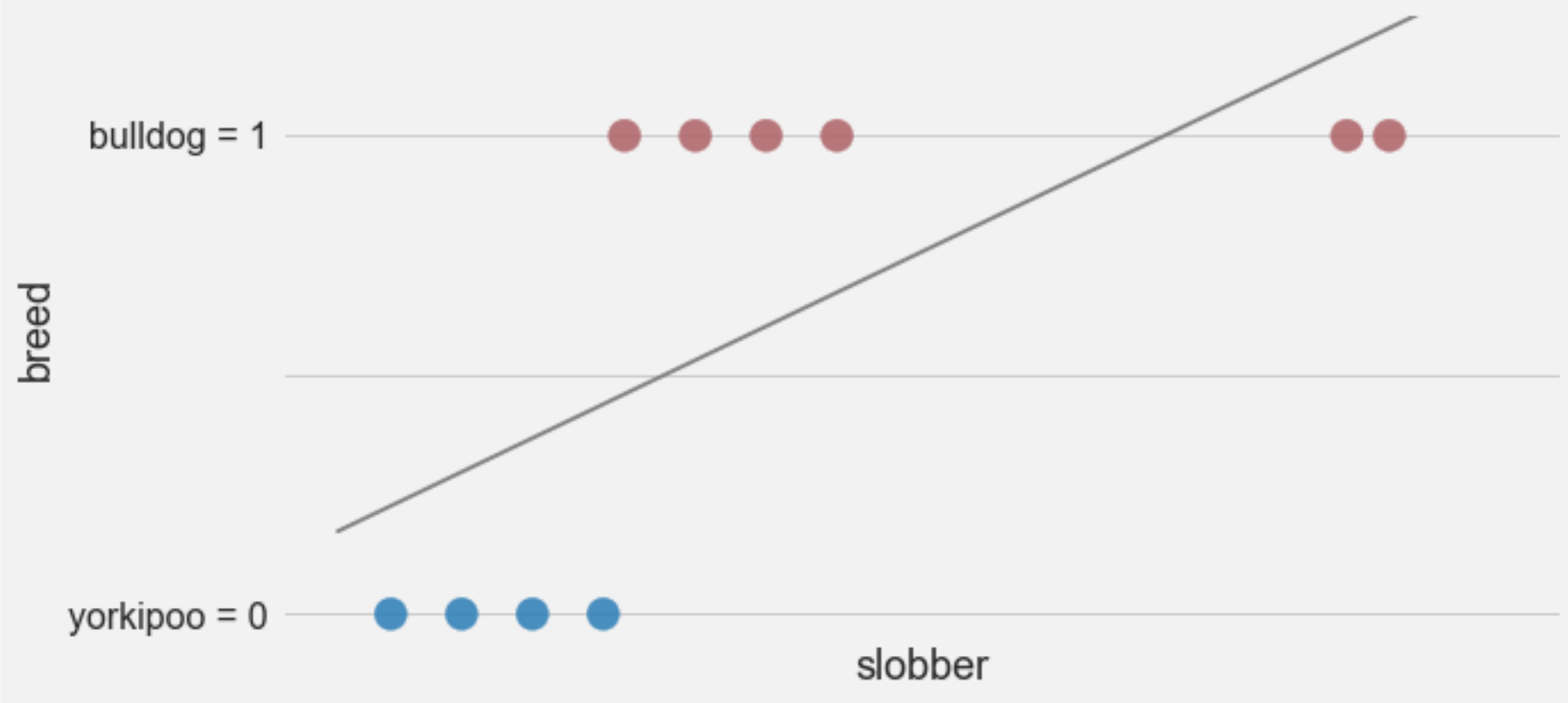
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

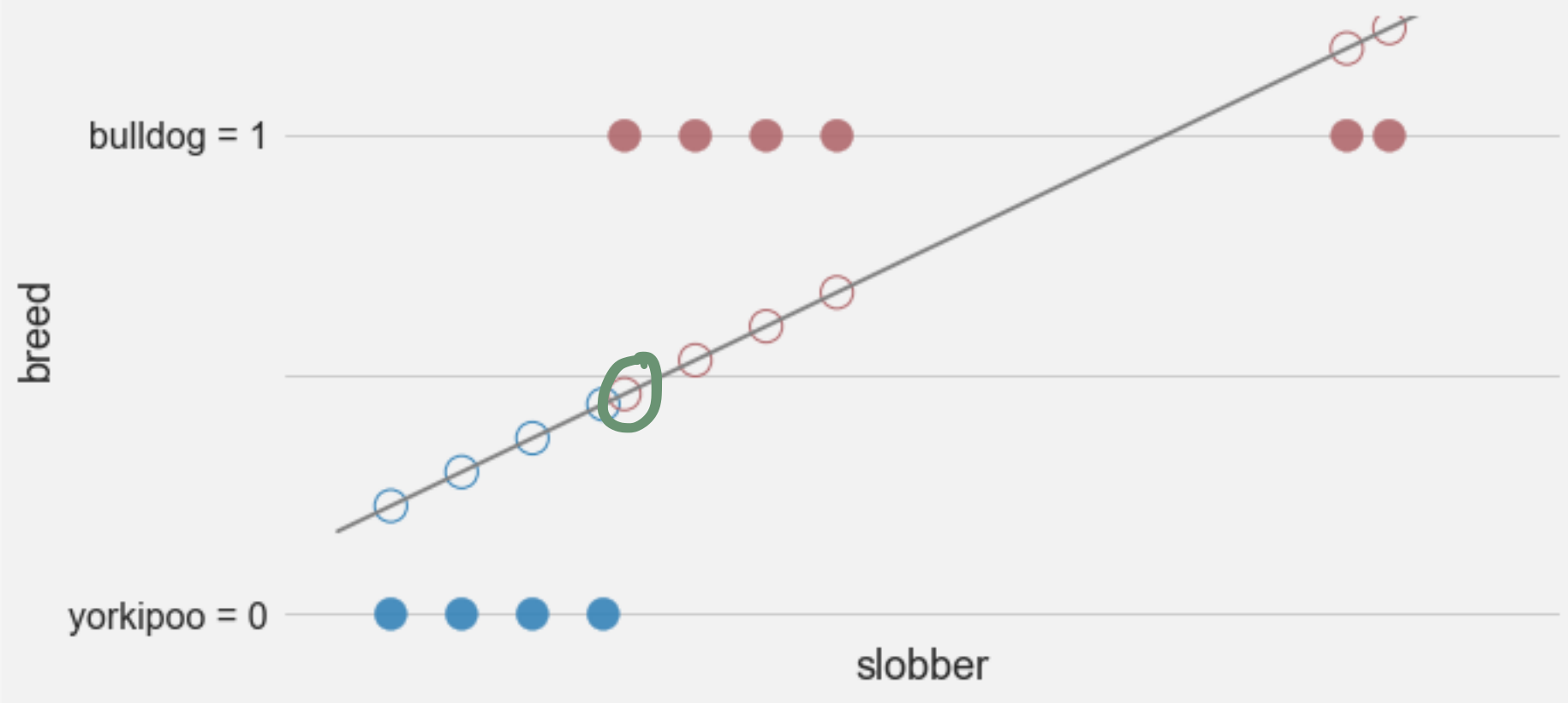
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

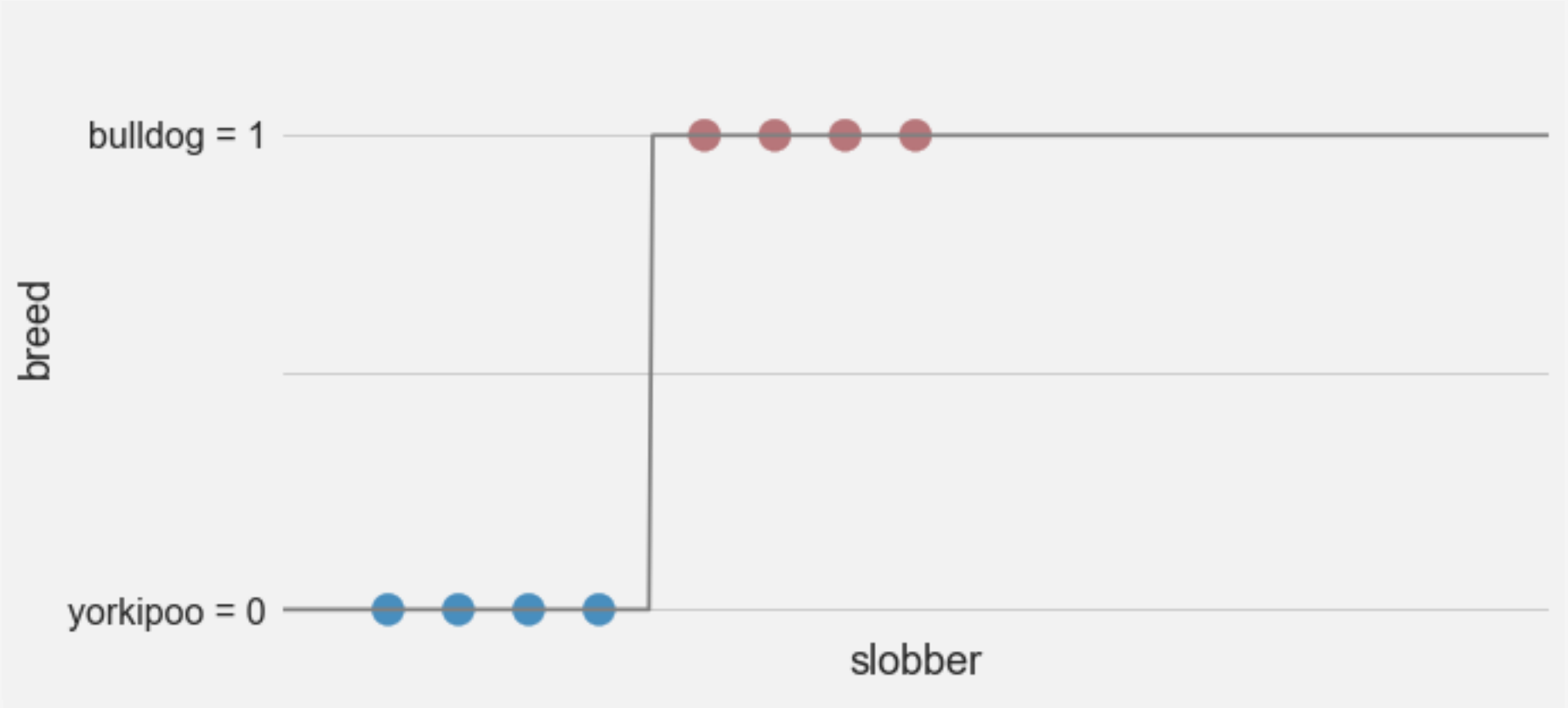
Idea: Linear Regression $y = \beta_0 + \beta_1 x_1$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

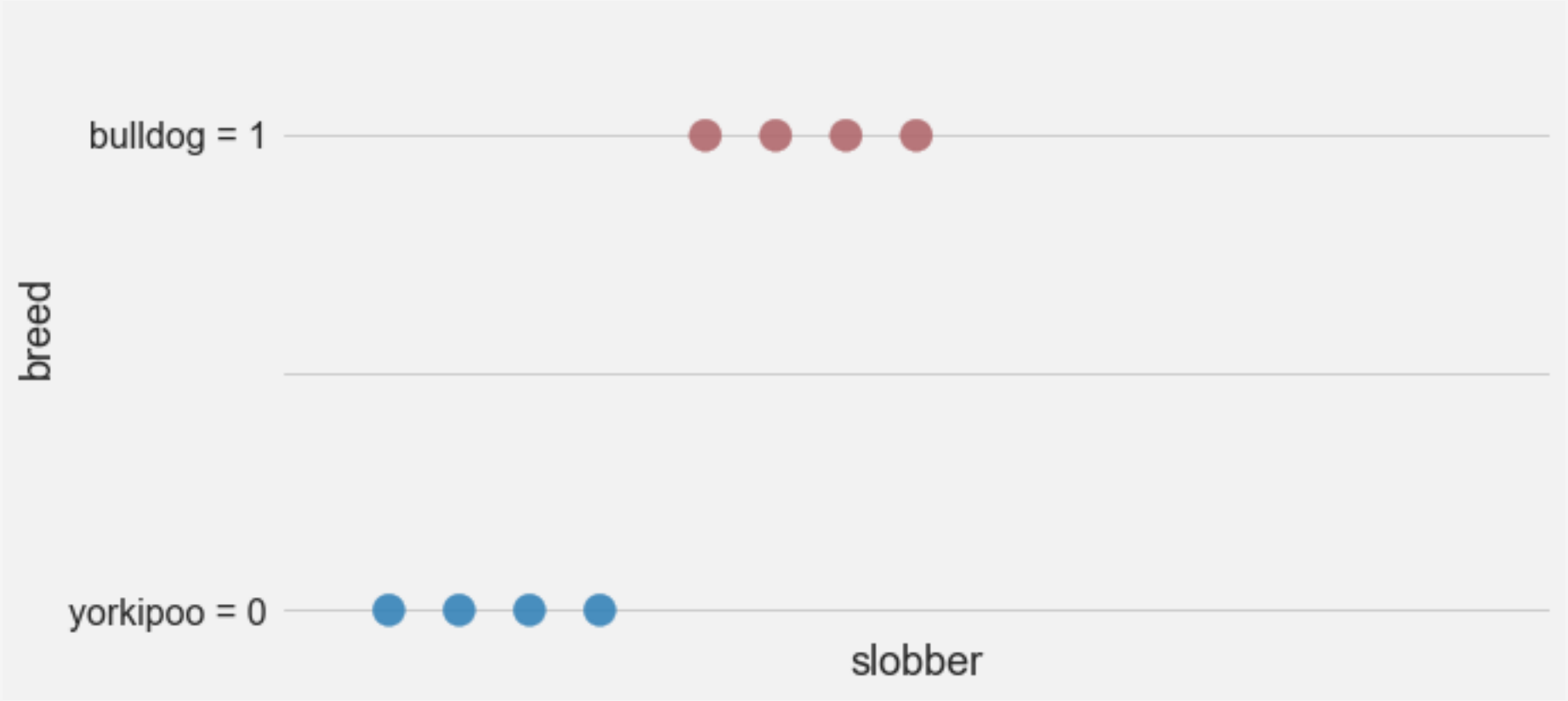
Idea: Use a piecewise function (modified perceptron) $y = \begin{cases} 1 & \text{if } \beta_0 + \beta_1 x_1 > 0 \\ 0 & \text{otherwise} \end{cases}$



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

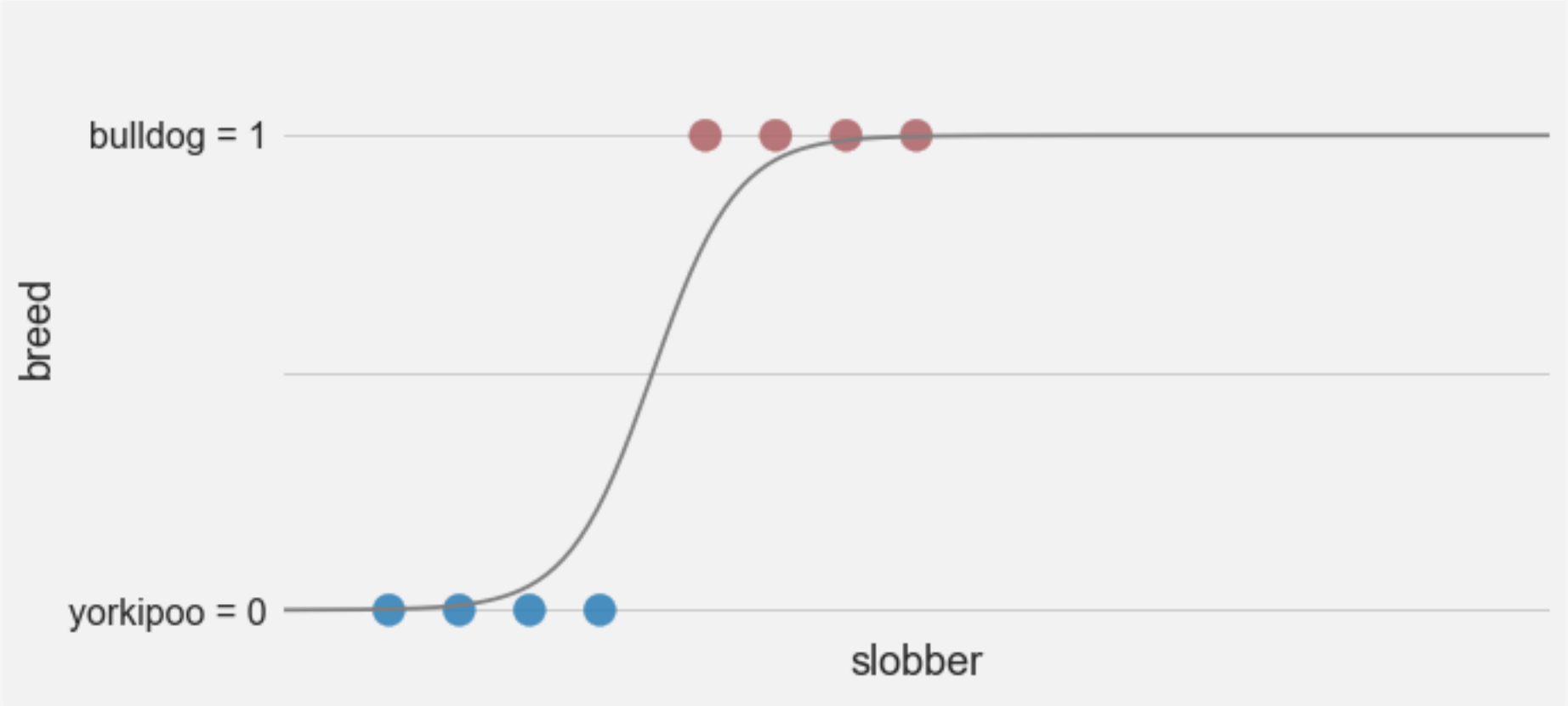
Idea: Need something that behaves more like a probability …



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

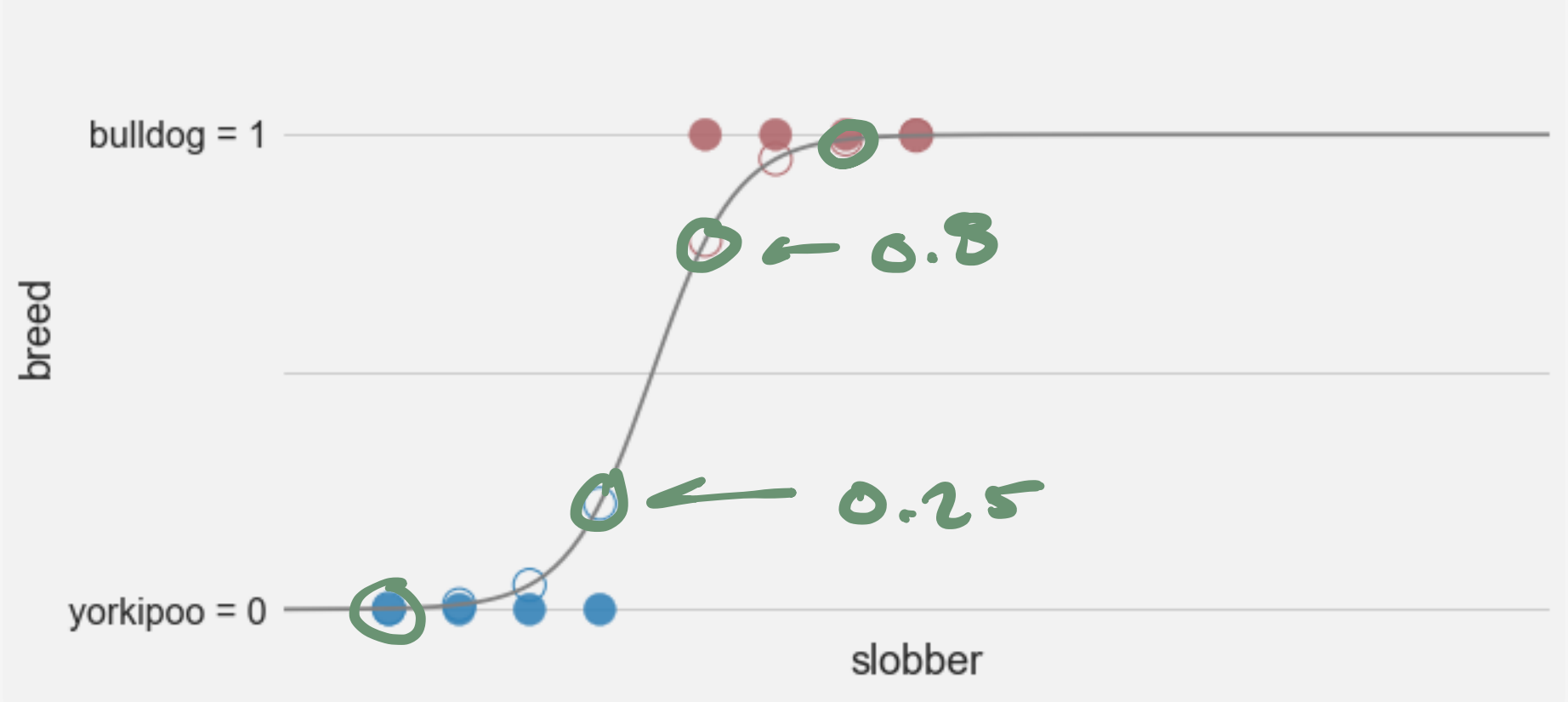
Idea: Need something that behaves more like a probability …



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

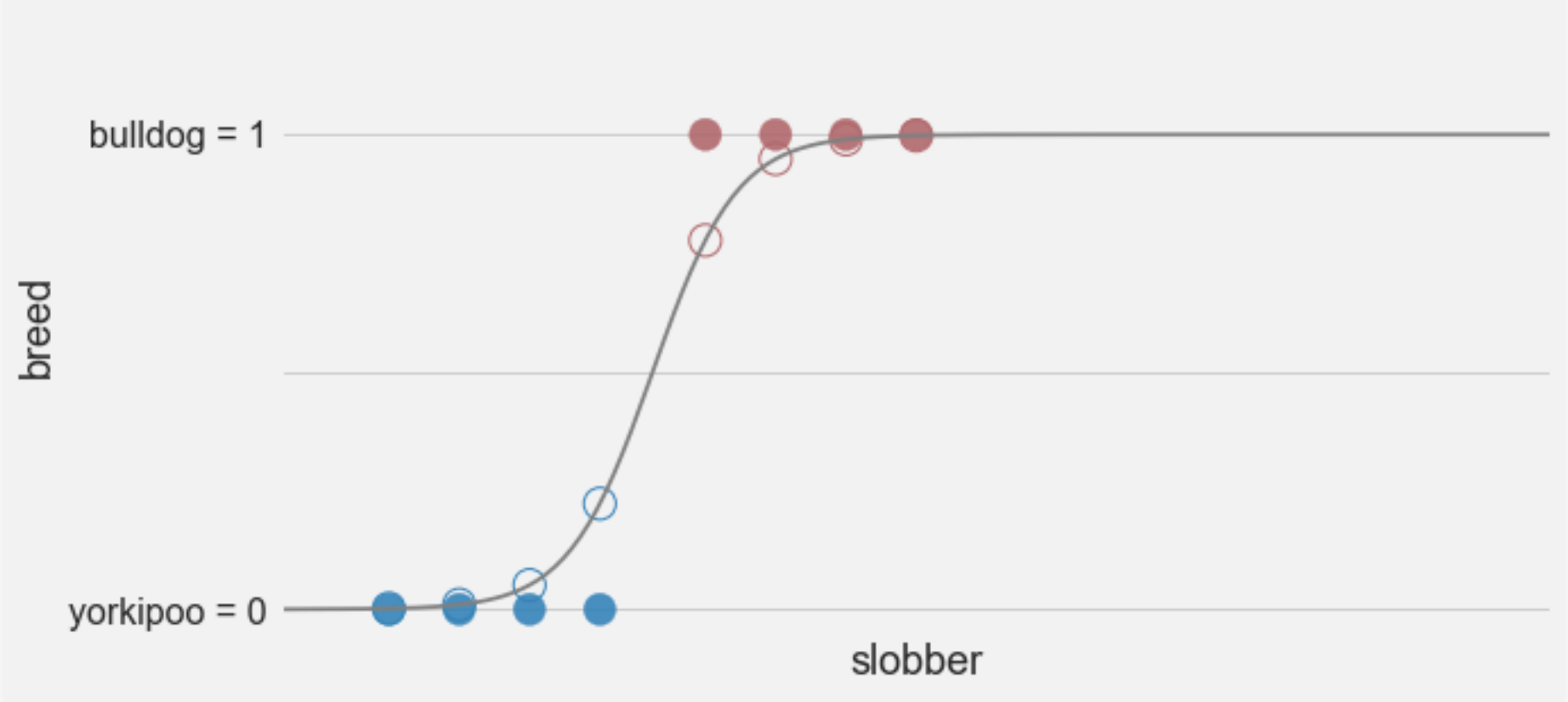
Idea: Need something that behaves more like a probability …



Predicting Dog Breed

The input to the model is a single feature: $x_1 =$ slobber

Question: What kind of function does this?



The Sigmoid Function

$z \to \infty \to 1$

$z \to -\infty \to 0$

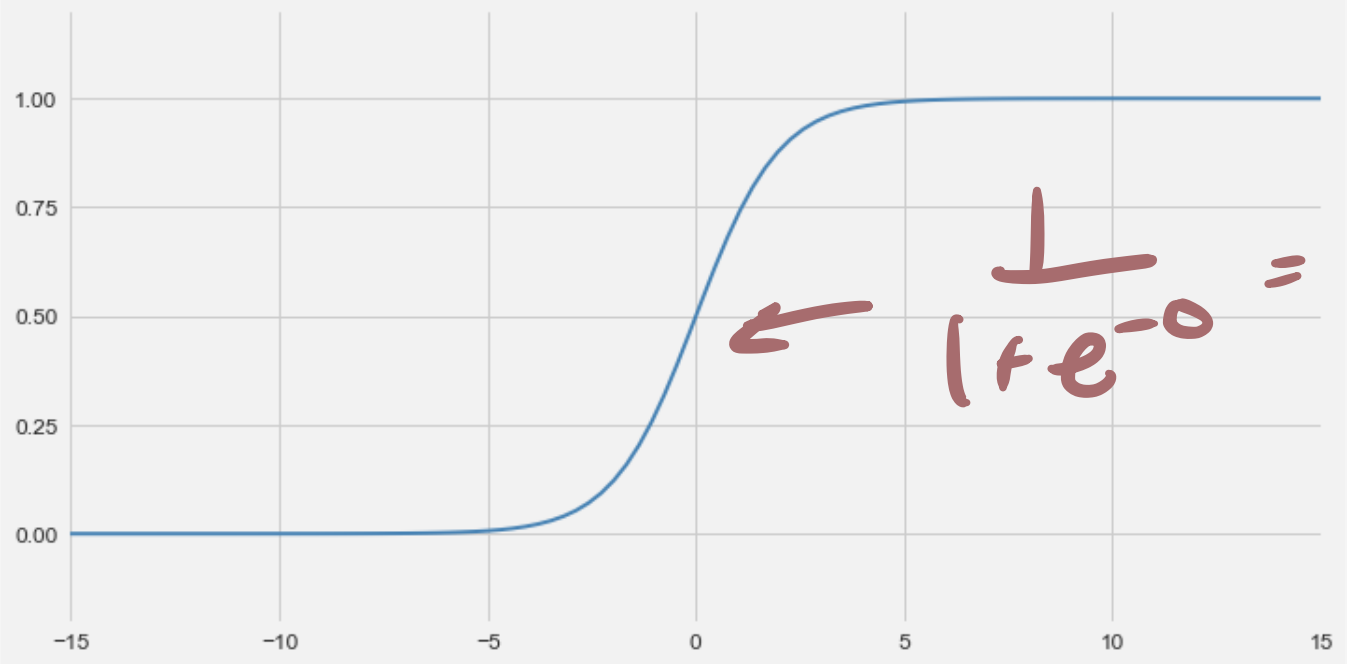
It has Everything!

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

Behaves like a Probability

Distinguishes Between Points

Really Smooth



$$\frac{1}{1+e^{-0}} = \frac{1}{1+1} = \frac{1}{2}$$

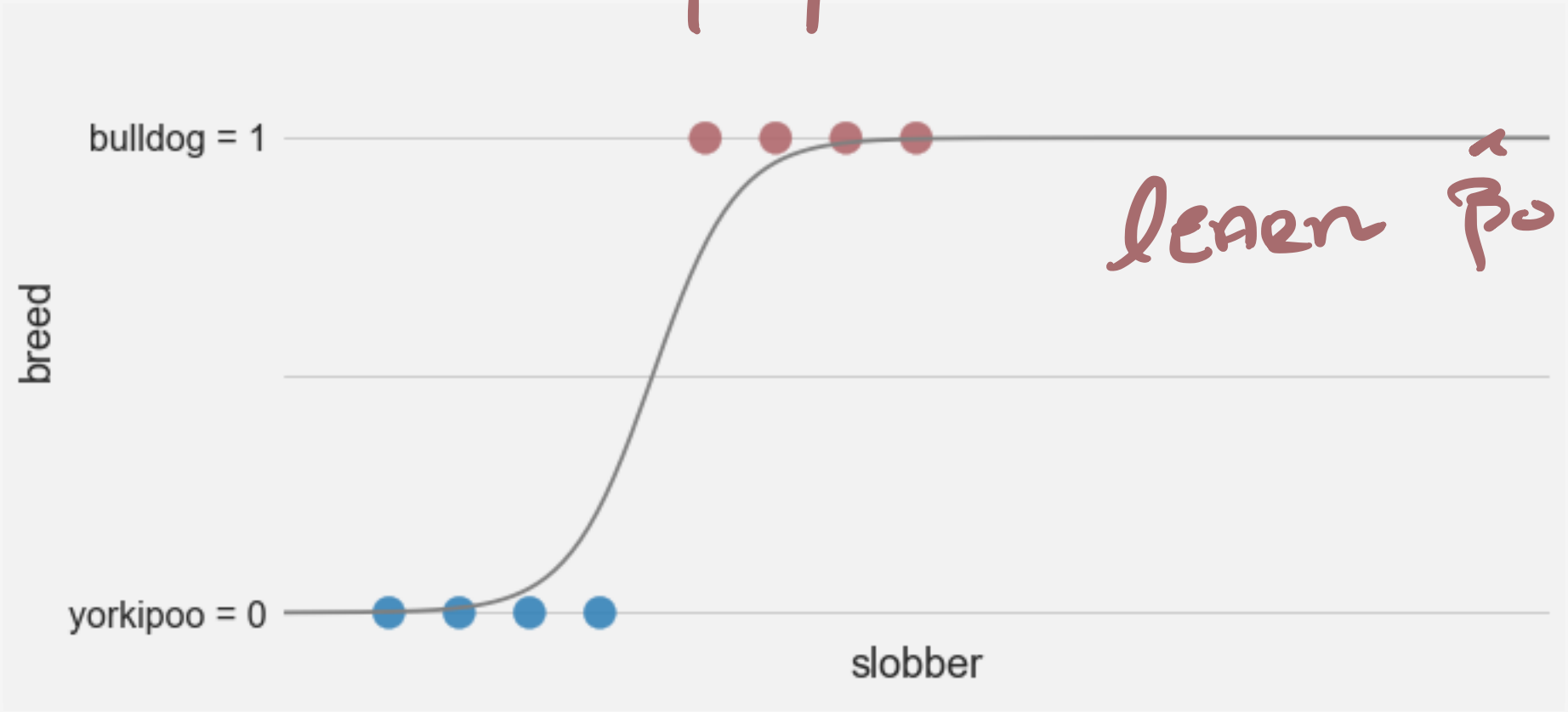
Logistic Regression

The Model:

Learn the weights from the data

$P(y=1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x)$

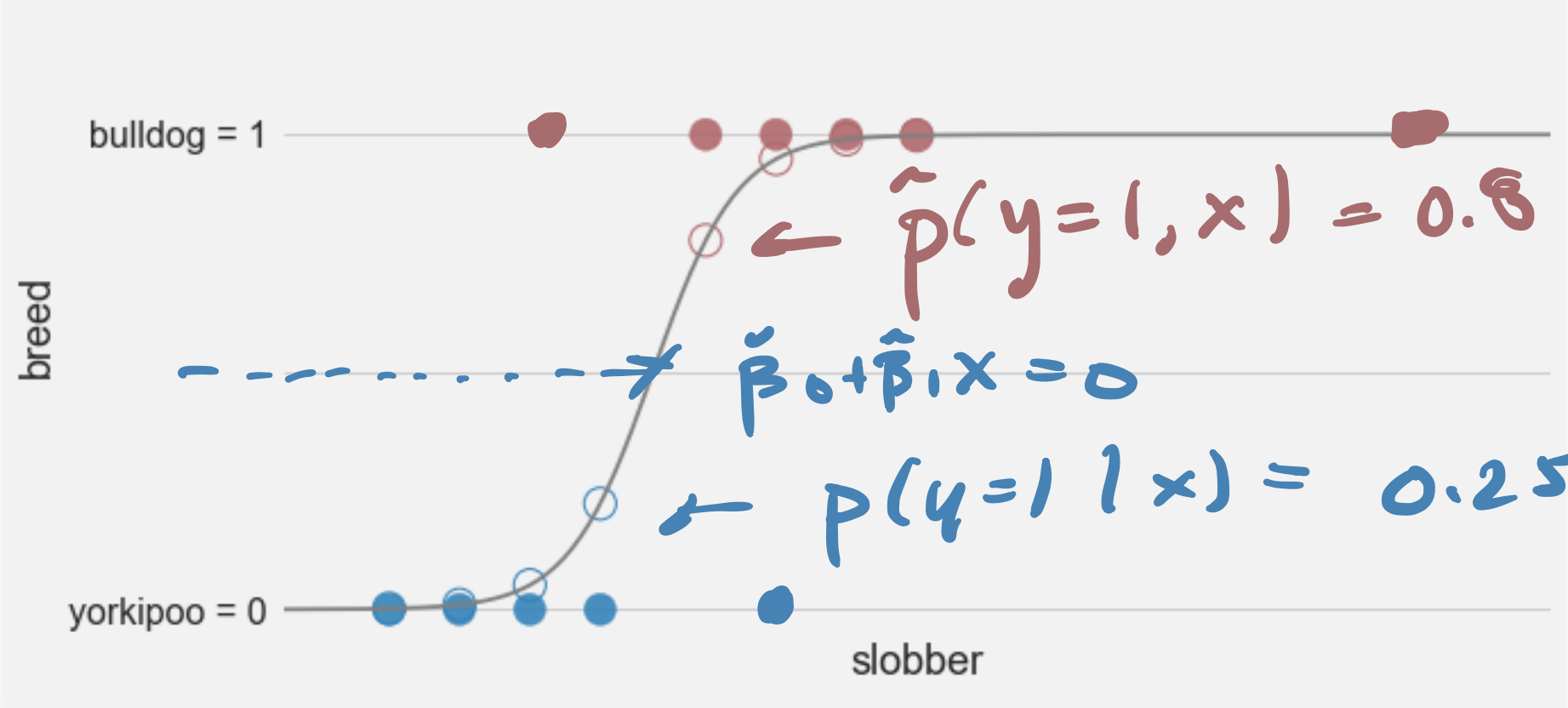
$P(y=0 \mid x) = 1 - \text{sigm}(\beta_0 + \beta_1 x)$



learn $\hat{\beta}_0$ & $\hat{\beta}_1$

Logistic Regression

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



An Odd(s) View of Logistic Regression

Our inevitable path to Logistic Regression and the sigmoid function began with our insistence on modeling on modeling the relationship between the features and the response as a bonafide probability.

It turns out that through some basic algebra we can arrive at an interpretation of Logistic Regression that is very regression-like.

But first we have to put on our gambling hats and talk about **odds**.

An Odd(s) View of Logistic Regression

In statistics, the odds of an event occurring 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 1: If $p = 0.75$ then $\text{odds} =$ 3

We would say the odds are 3 to 1 in favor

Example 2: If $p = 0.1$ then $\text{odds} =$ 1/9

We would say the odds are 9 to 1 against

$p = 0.75$

$\text{ODDS} = \frac{3/4}{1-3/4} = 3$

$p = 0.1 = 1/10$

$\text{ODDS} = \frac{1/10}{1-1/10} = \frac{1}{9}$

An Odd(s) View of Logistic Regression

In Logistic Regression we model $p = p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x)$

If instead we compute the odds that y = 1 given the data, we have

$$\text{ODDS} = \frac{p}{1-p} = \frac{\text{sigm}(\beta_0+\beta_1 x)}{1-\text{sigm}(\beta_0+\beta_1 x)}$$

... MIRACLE ...

An Odd(s) View of Logistic Regression

Taking the natural log of both sides, gives

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

So it turns out we **have** been doing linear regression all along, but for the **log-odds** instead of the probability!

Backing up a step, we had $\text{odds} = \exp(\beta_0 + \beta_1 x)$

This gives us a new interpretation of the Logistic Regression weight β_1

An Odd(s) View of Logistic Regression

$$\text{ODDS} = \exp[\beta_0 + \beta_1 x]$$

x INCREASES 1 UNIT $\quad x \to x+1$

$$\text{ODDS} \to \exp[\beta_0 + \beta_1(x+1)]$$

$$= \exp[(\beta_0 + \beta_1 x) + \beta_1]$$

$$= \exp[\beta_0 + \beta_1 x]\exp[\beta_1]$$

$$\frac{\exp[\beta_0 + \beta_1 x]\exp[\beta_1]}{\text{OLD ODDS}}$$

INCREASE / DECREASE BY FACTOR OF $\exp[\beta_1]$

Logistic Regression with Many Features

Logistic Regression with a single feature looks like: $p(y = 1 \mid x) = \text{sigm}(\beta_0 + \beta_1 x)$

But in real life we typically have many features:

- **Predict** which candidate a person will vote for
- **Features**: education, household income, zip code, religion, etc

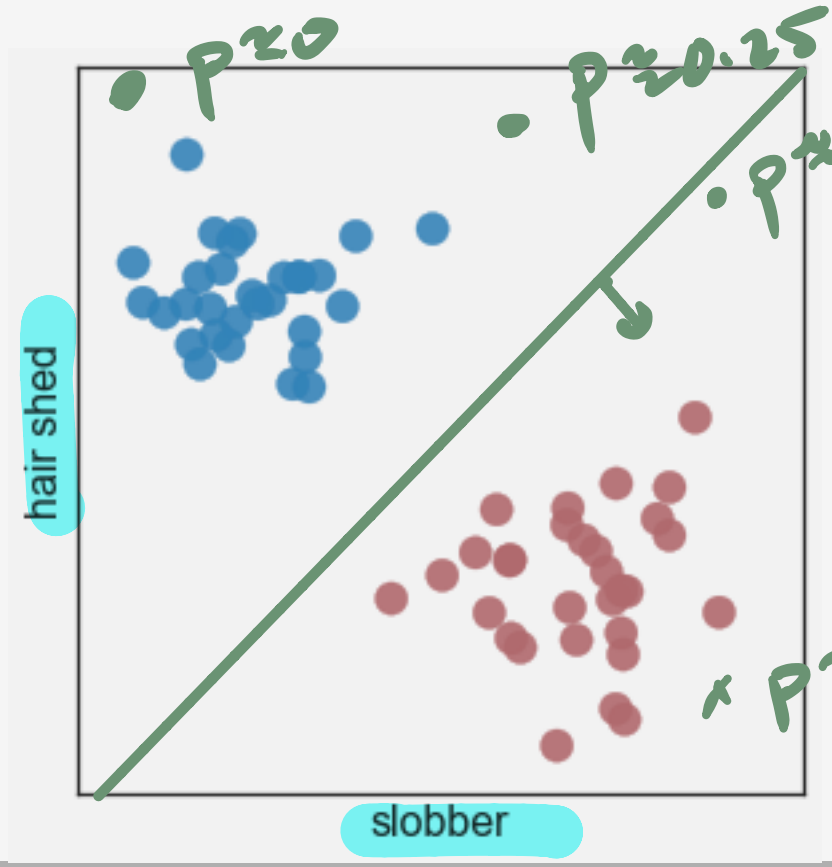
Multiple Feature Logistic Regression Model:

$$p(y = 1 \mid \vec{x}) = \text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

Logistic Regression with Many Features

Multiple Feature Logistic Regression Model:

$$p(y = 1 \mid \mathbf{x}) = \text{sigm}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right)$$



DECISION BDRY

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

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

$= \text{intercept}$ $= \text{slope}$

Logistic Regression with Many Features

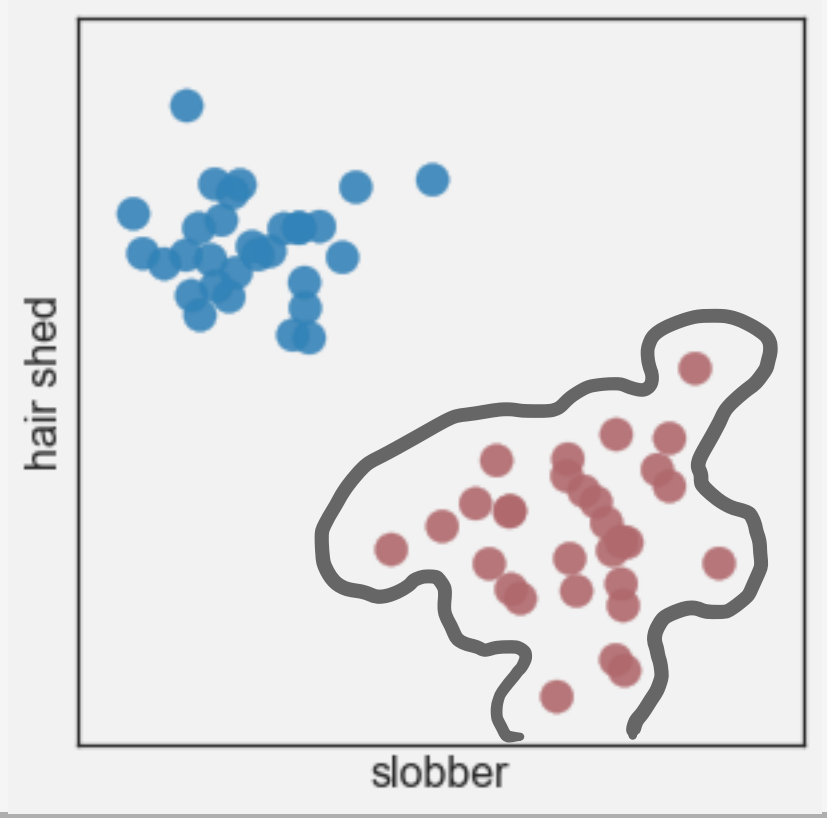
Multiple Feature Logistic Regression Model:

$$p(y = 1 \mid \mathbf{x}) = \text{sigm}\,(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

$\text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2)$

DB

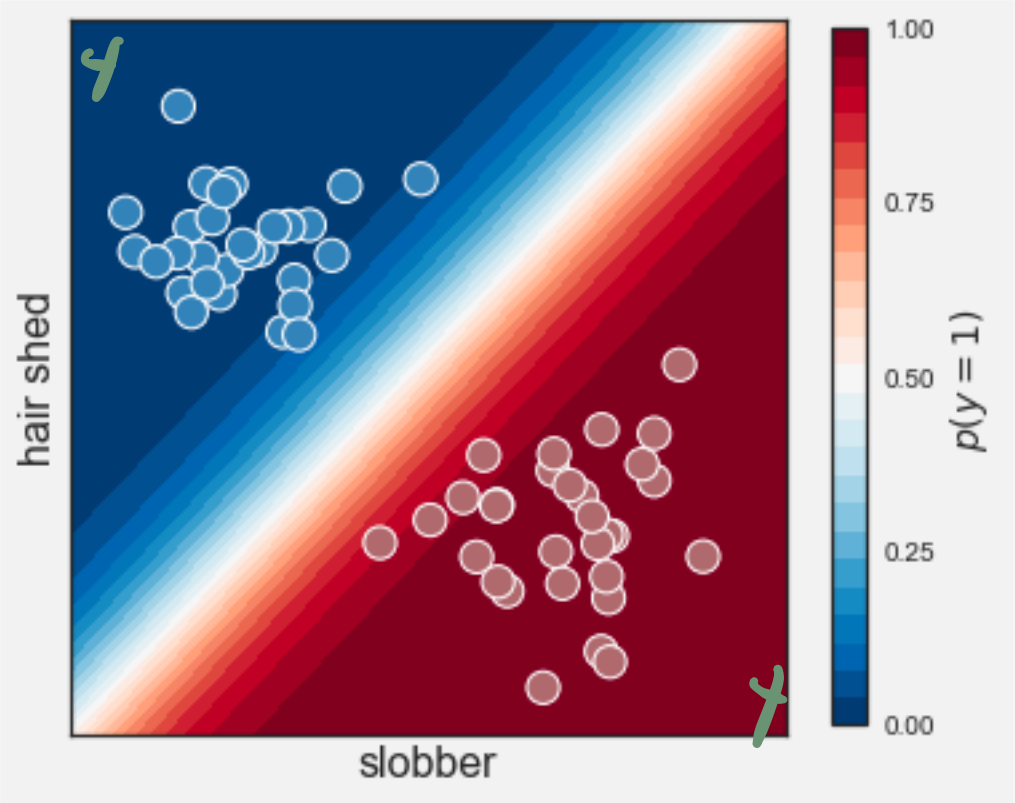
$\beta_0 + \beta_1 x_1 \cdots + \beta_4 x_2^2 = 0$



Logistic Regression with Many Features

Multiple Feature Logistic Regression Model:

$$p(y = 1 \mid \mathbf{x}) = \text{sigm}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right)$$



Logistic Regression Wrap-Up

- Logistic regression is a linear classifier with a probabilistic interpretation
- Can make it nonlinear by using derived nonlinear features
- Next week we'll train logistic regression models with Stochastic Gradient Descent

Next Time:

- How to derive features from text so we can classify documents

If-Time Bonus: Sigmoid Derivative

The Sigmoid function has some nice differential properties that we'll explore later

The most important of which, is that

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