University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chris Ketelsen

Lecture 3: Logistic Regression

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

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

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

Slightly easier this time

$$p(y = 0 \mid \mathbf{x}, \mathcal{D}) = 1 - p(y = 1 \mid \mathbf{x}, \mathcal{D})$$

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:

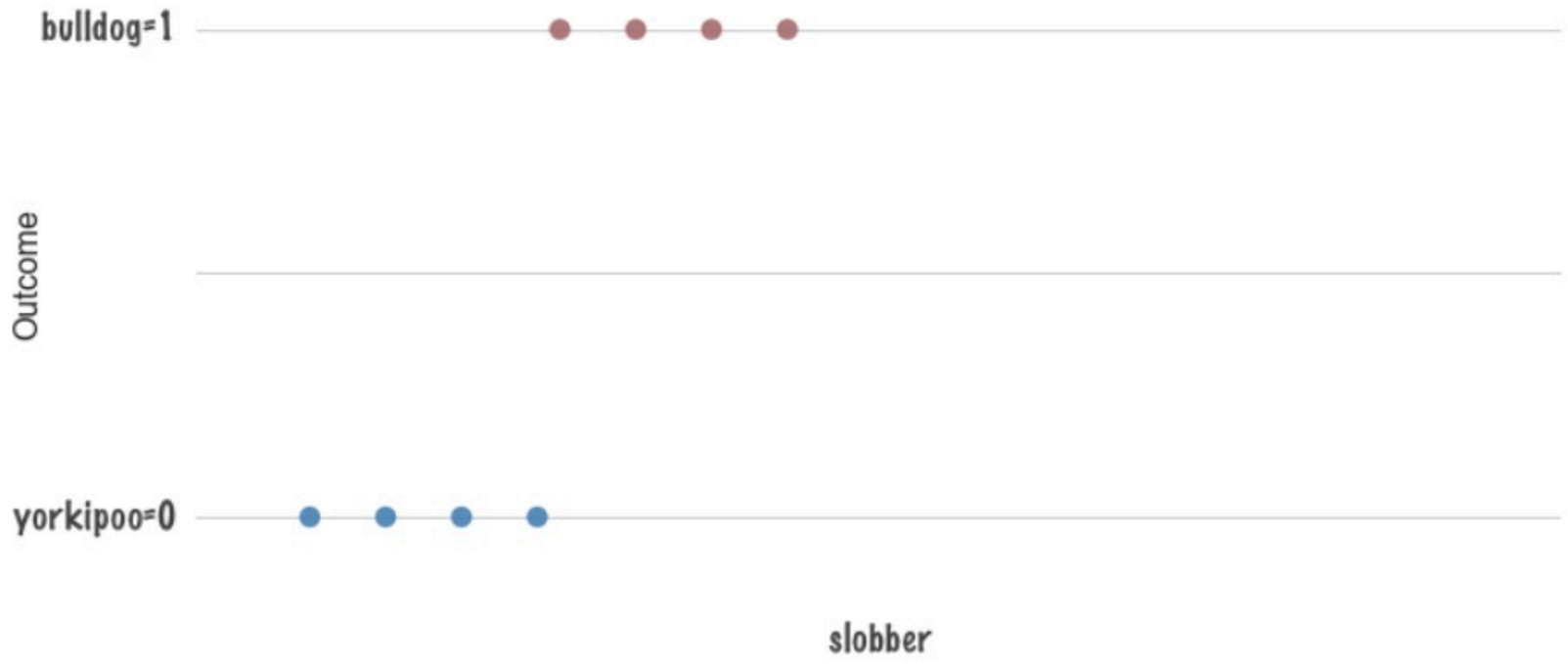
- Model only conditional relationship $p(y \mid \mathbf{x})$
- **Today**: Logistic Regression

Logistic Regression

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

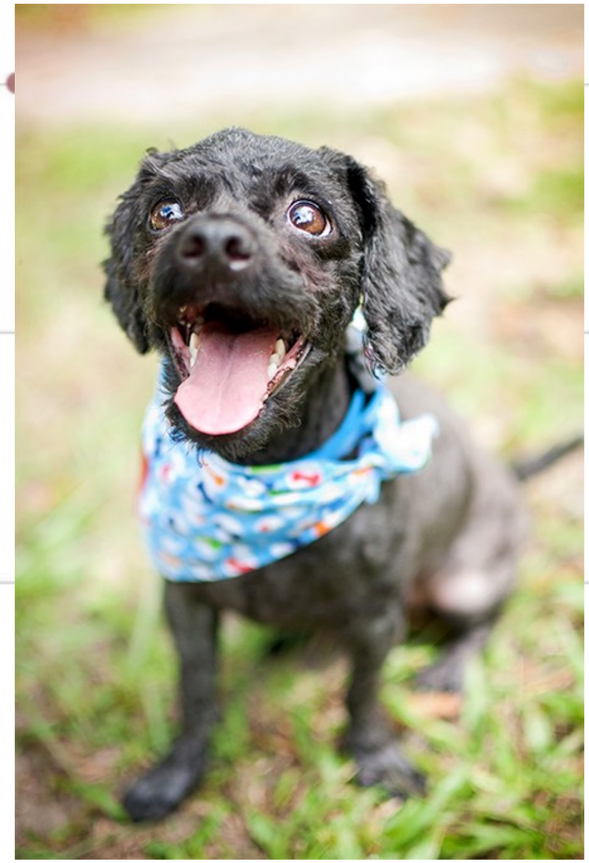
A Simple Example

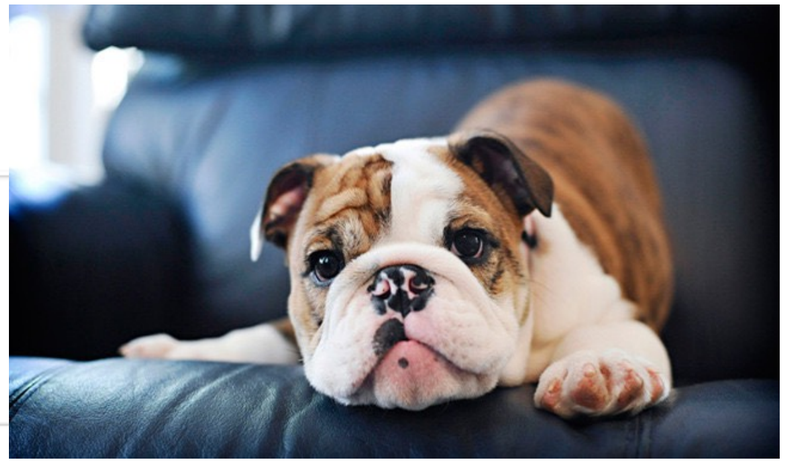
Suppose ~~you track patients in a cancer study~~ you have two dogs that frequently commit dog-crimes when you're not around. One is a bulldog and the other is a yorkipoo. Your primary source of evidence of the culprit is how much slobber is left behind at the scene.



A Simple Example

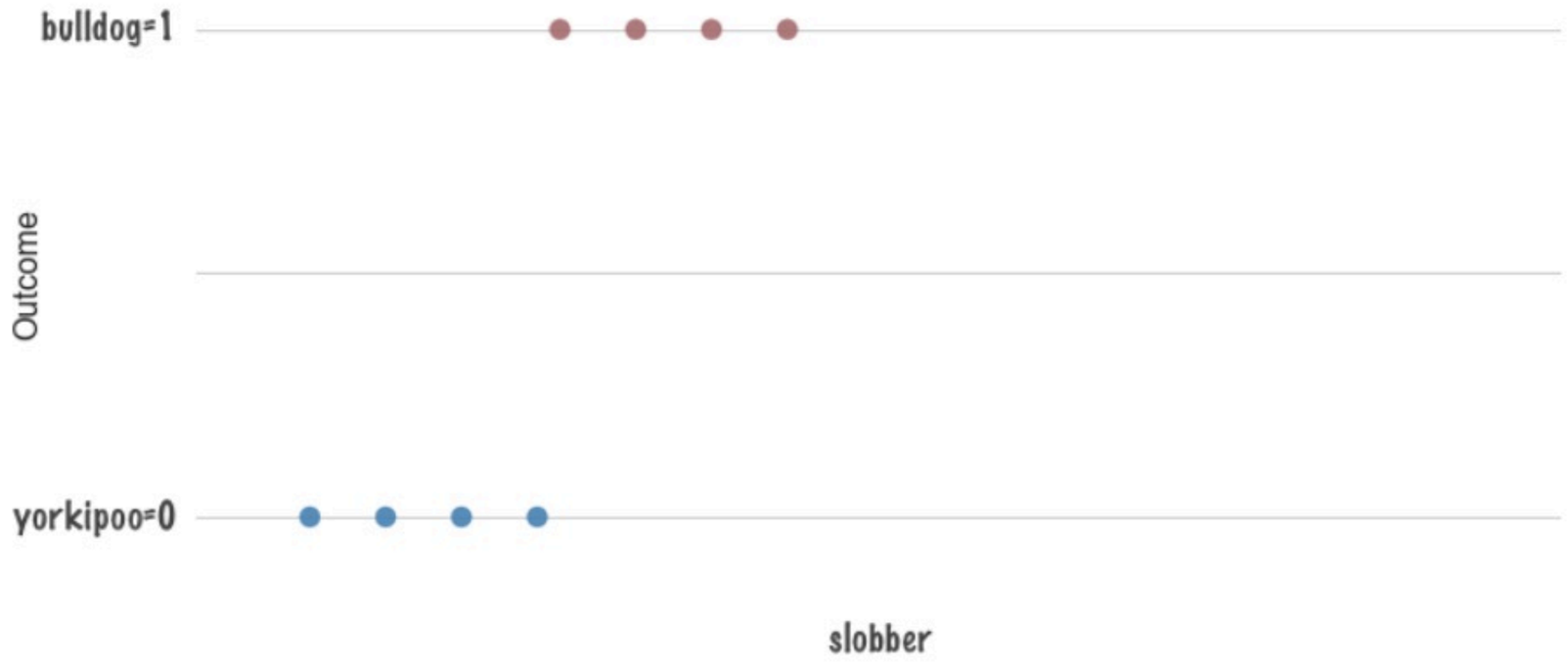
Suppose ~~you track patients in a cancer study~~ you have two dogs that frequently commit dog-crimes when you're not around. One is a bulldog and the other is a yorkipoo. Your primary source of evidence of the culprit is how much slobber is left behind at the scene.





A Simple Example

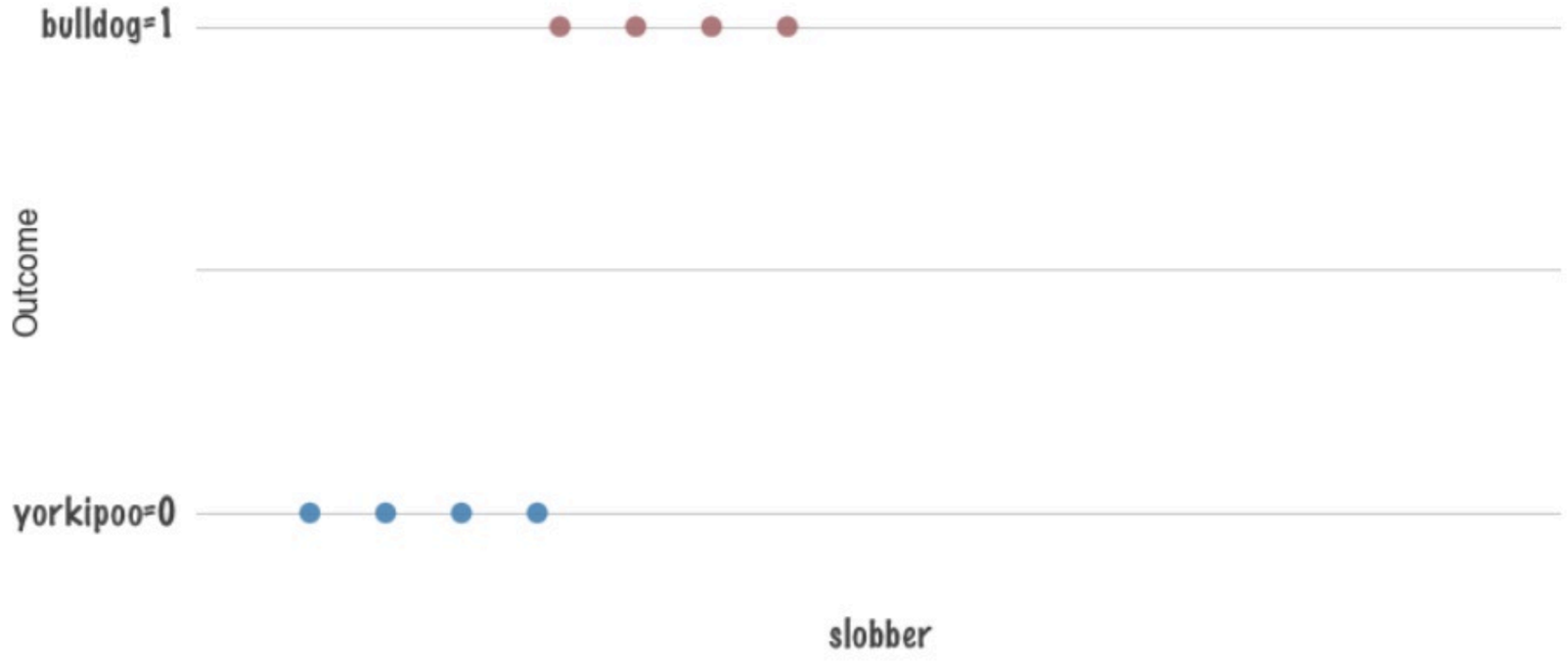
Suppose ~~you track patients in a cancer study~~ you have two dogs that frequently commit dog-crimes when you're not around. One is a bulldog and the other is a yorkipoo. Your primary source of evidence of the culprit is how much slobber is left behind at the scene.



A Simple Example

Single feature: x_1 = slobber

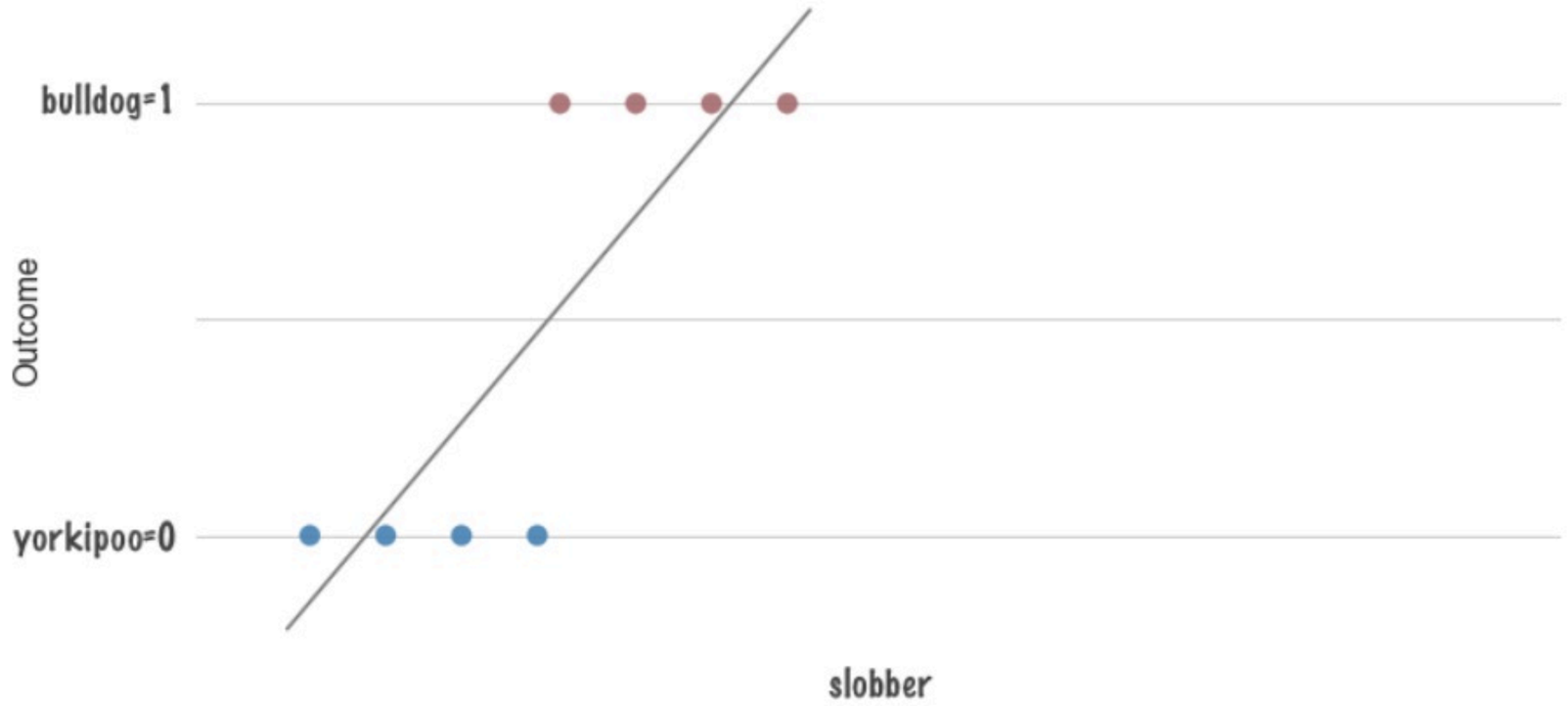
How do we model $p(y \mid x_1, \mathcal{D})$?



A Simple Example

Single feature: x_1 = slobber

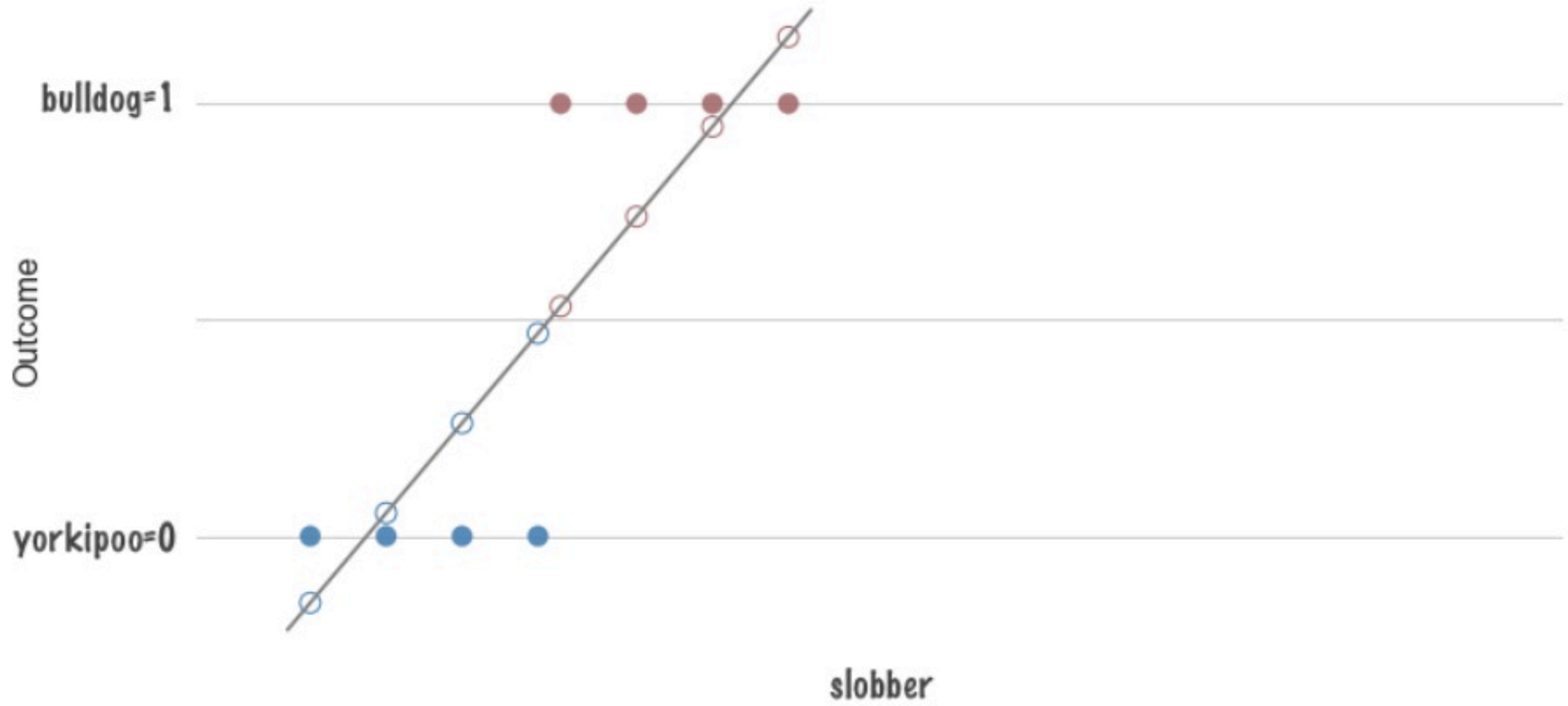
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

Single feature: x_1 = slobber

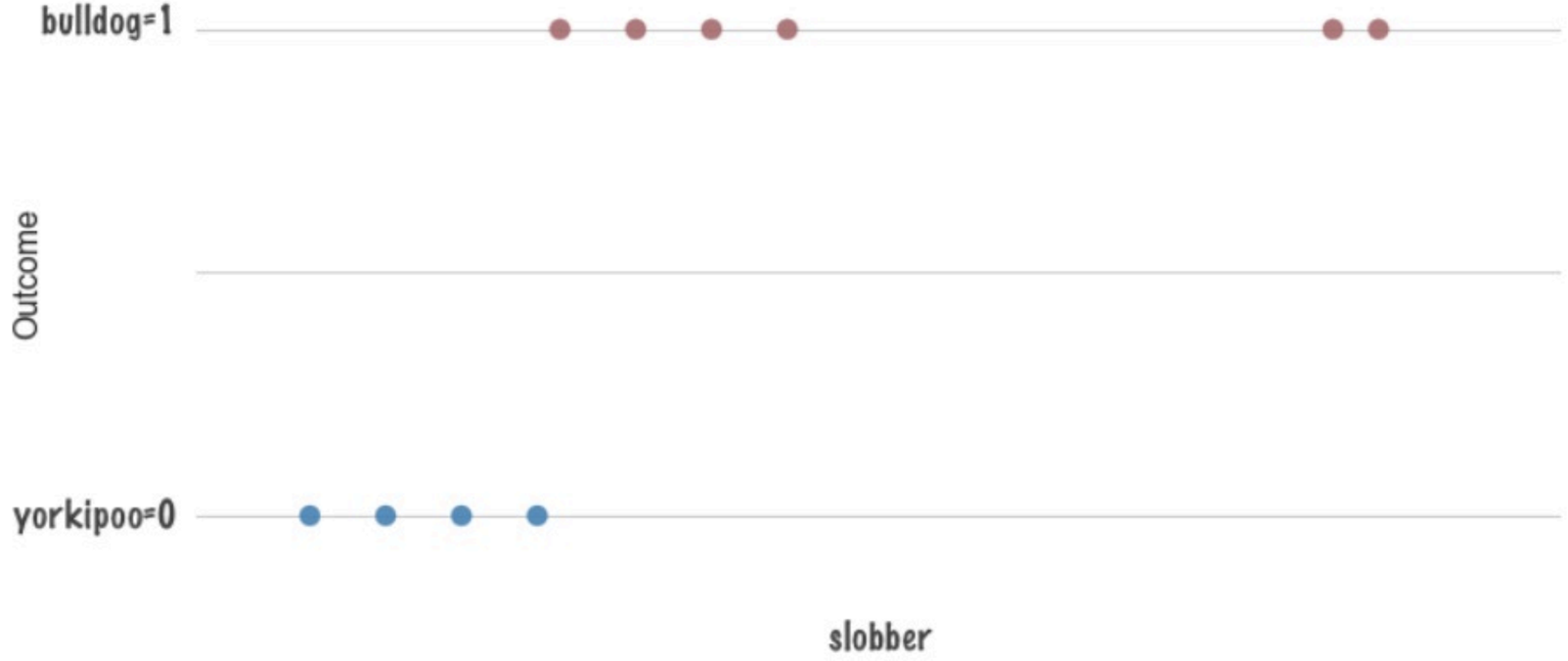
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

Single feature: x_1 = slobber

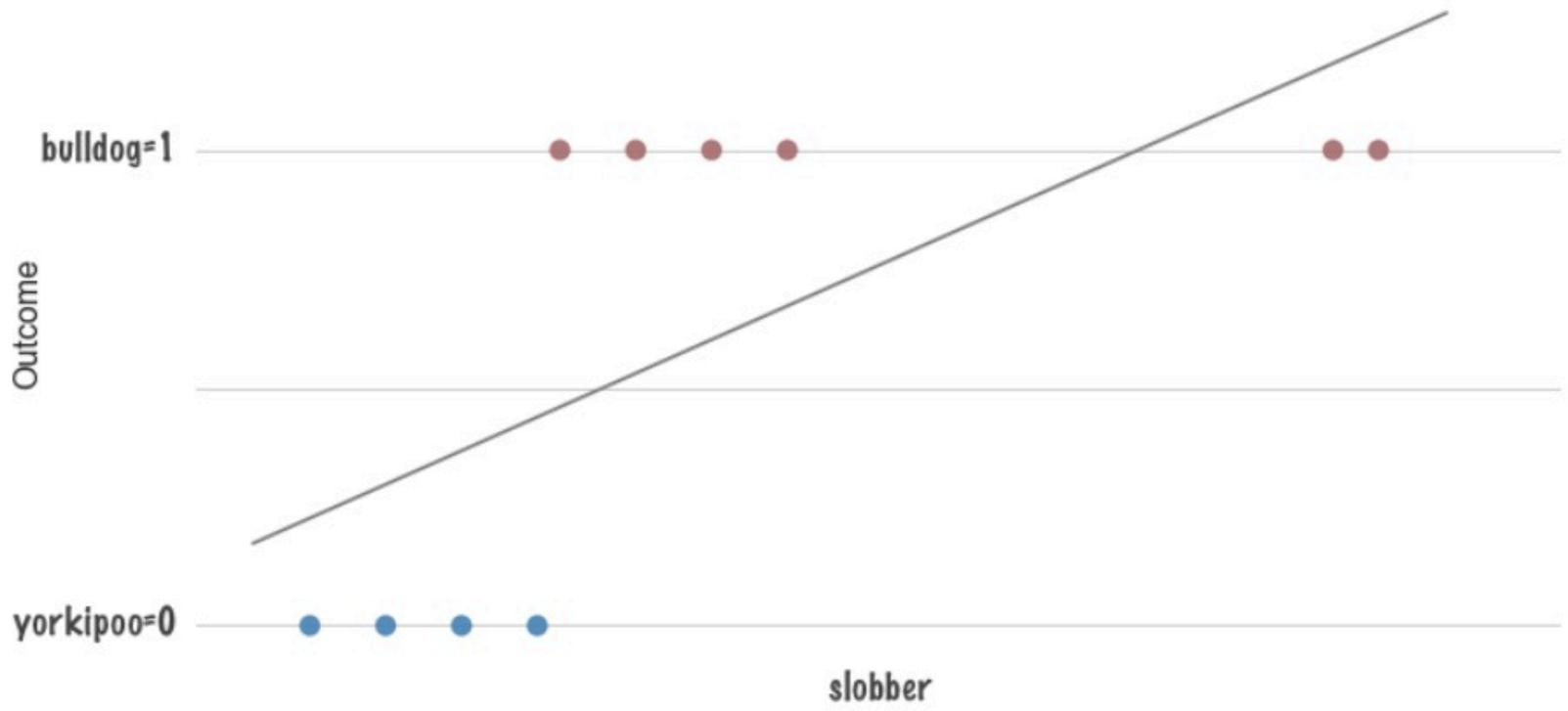
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

Single feature: x_1 = slobber

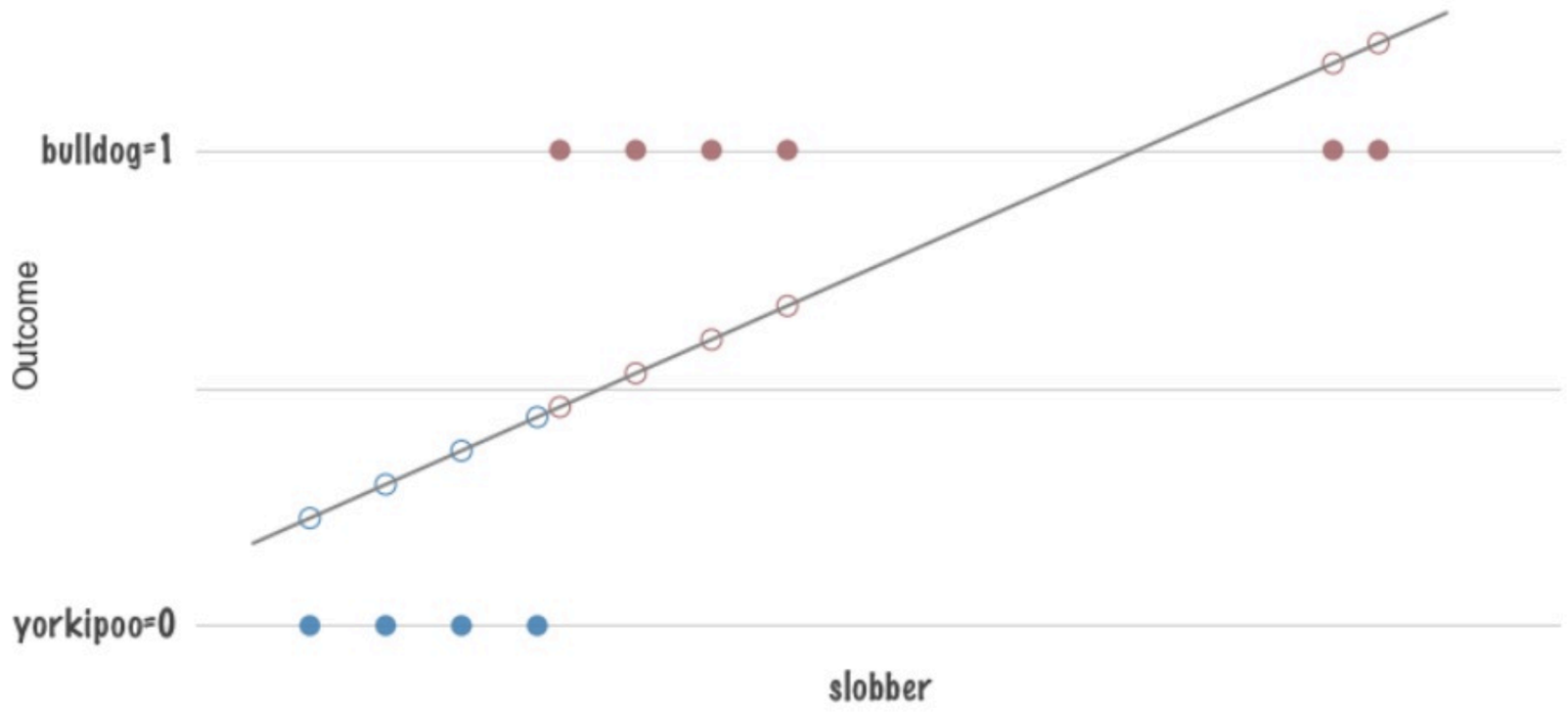
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

Single feature: x_1 = slobber

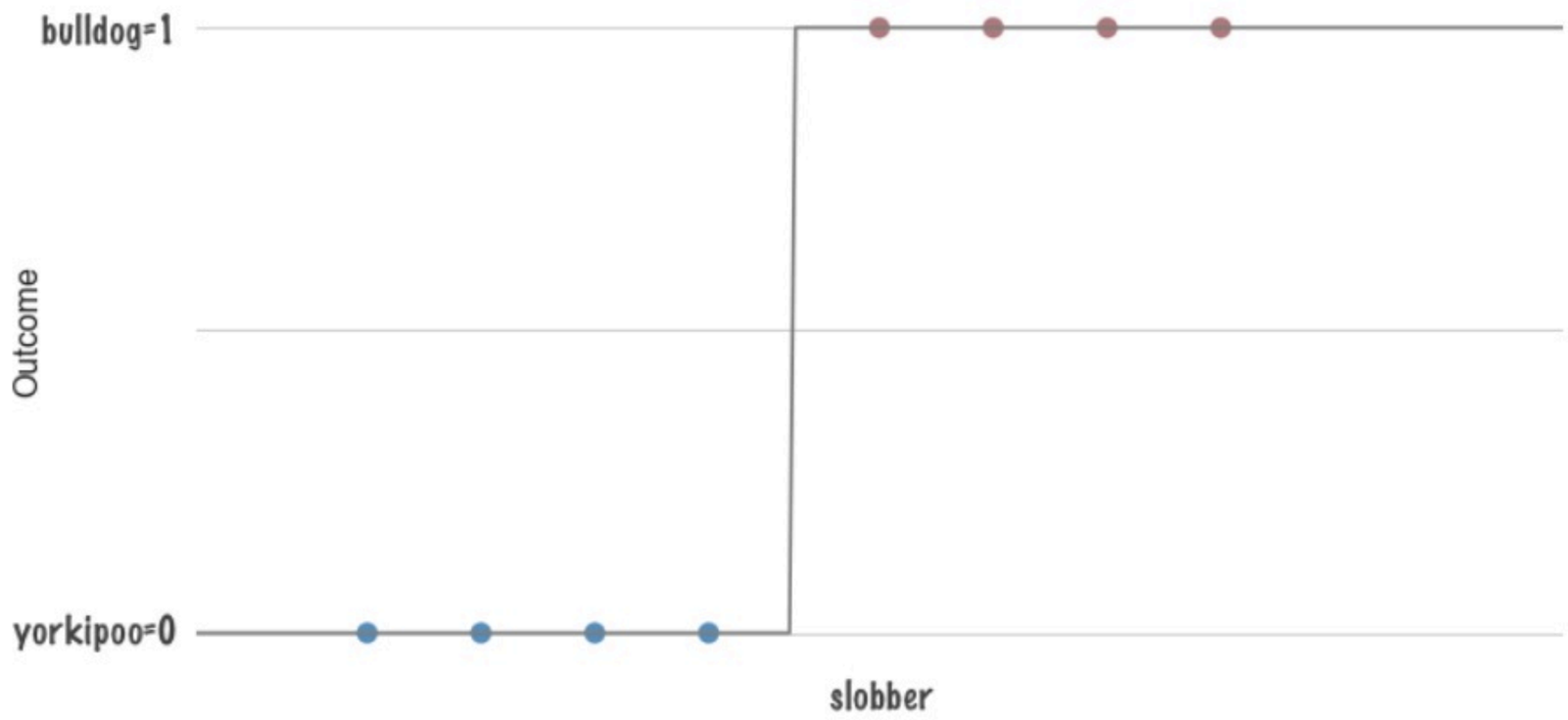
Idea: Linear Regression $p(y = 1 \mid x_1; \mathbf{w}) = w_0 + w_1 x_1$



A Simple Example

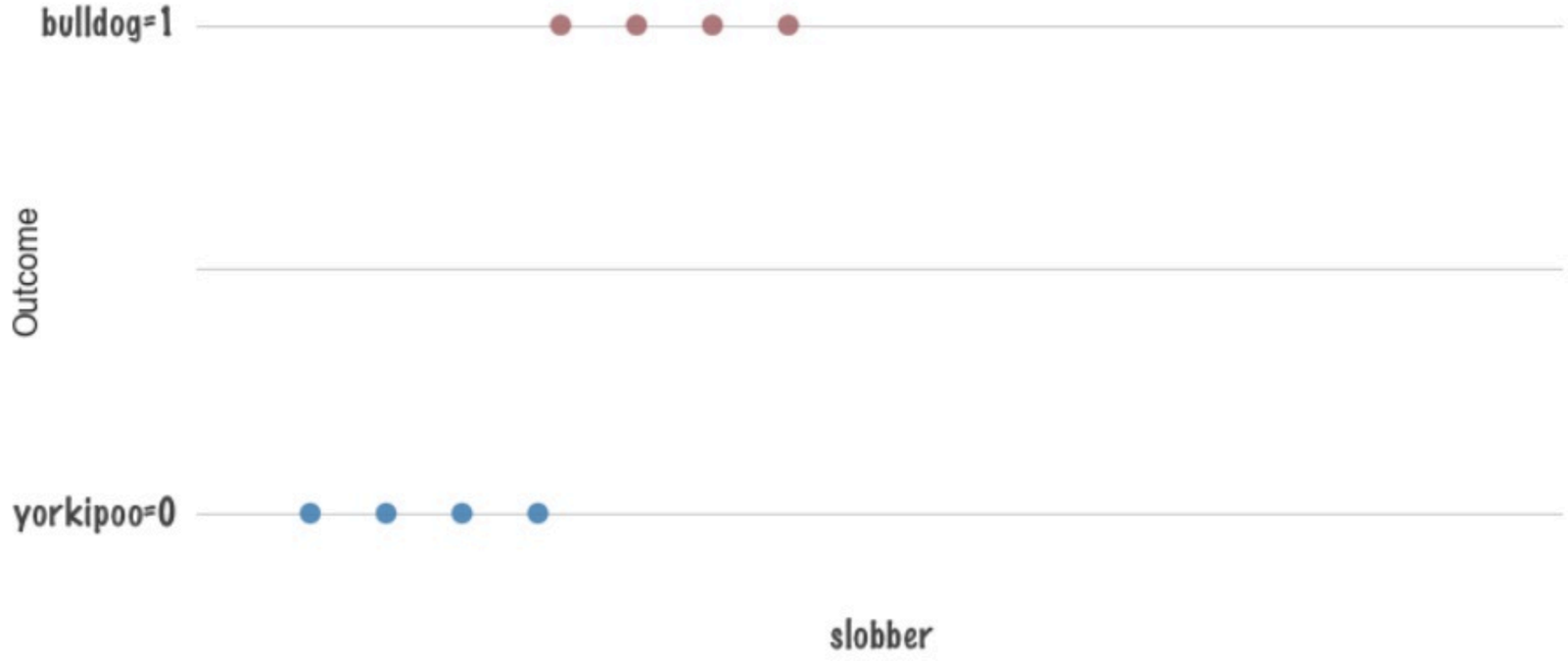
Single feature: x_1 = slobber

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

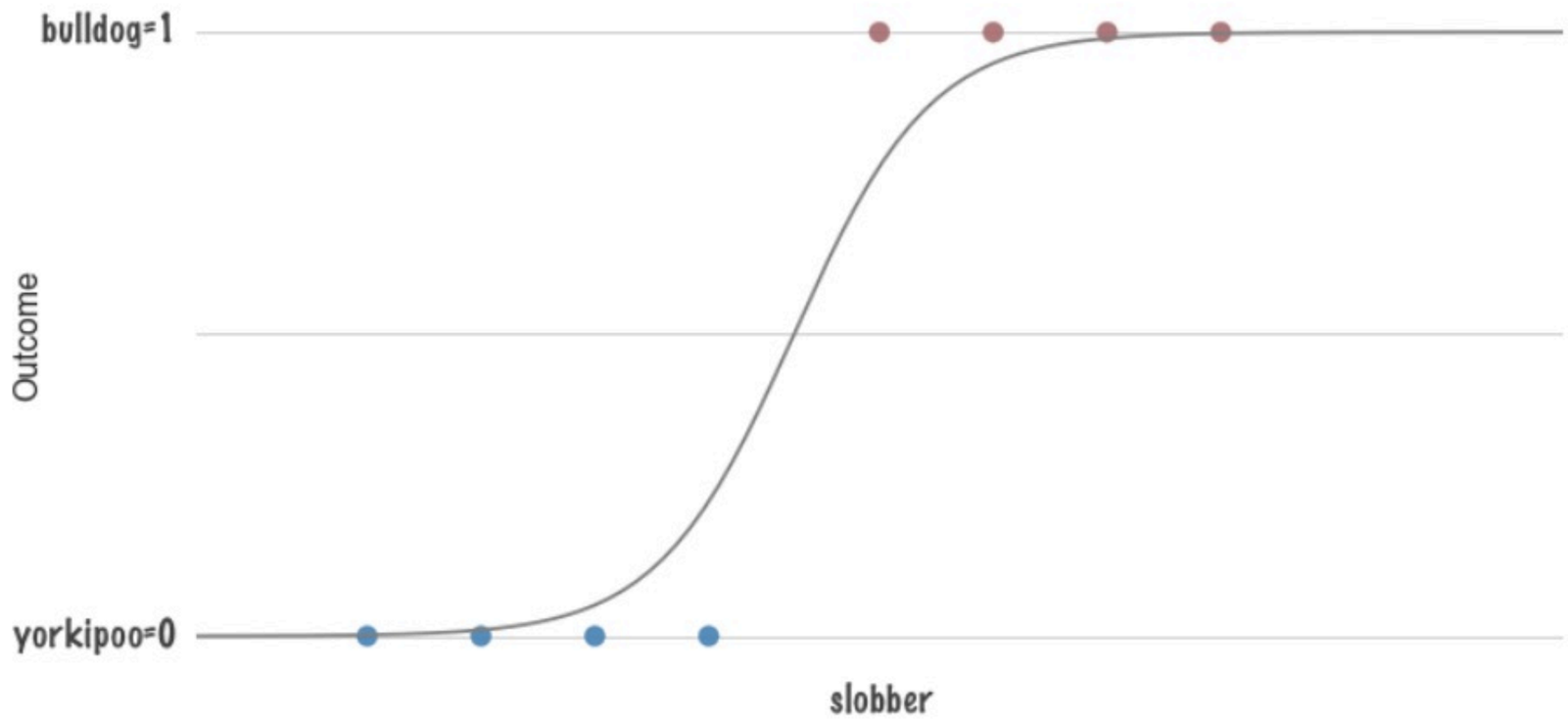
Need something that behaves more like a probability ...



Enter the sigmoid Function

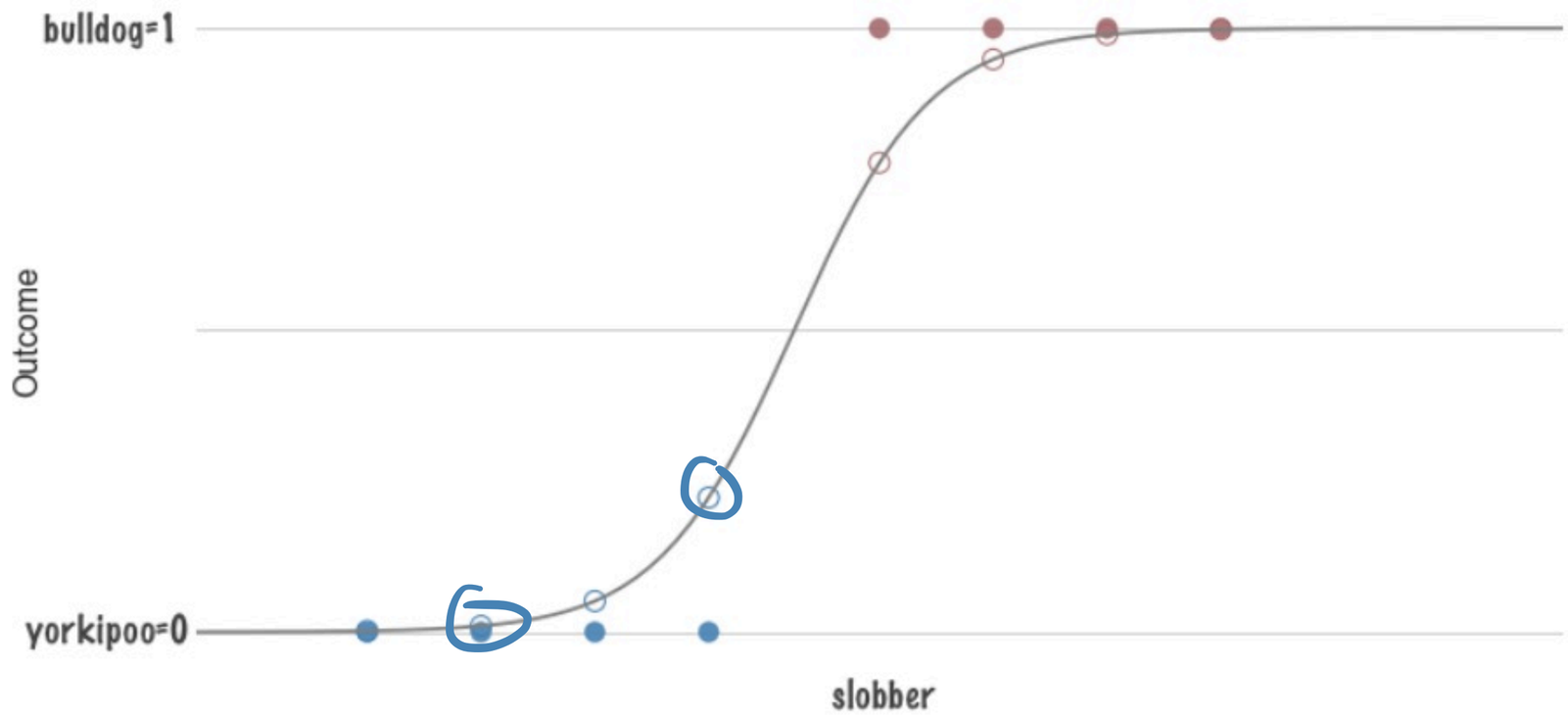
$$\text{sigm}(z) = \frac{1}{1+e^{-z}}$$

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



Enter the sigmoid Function

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

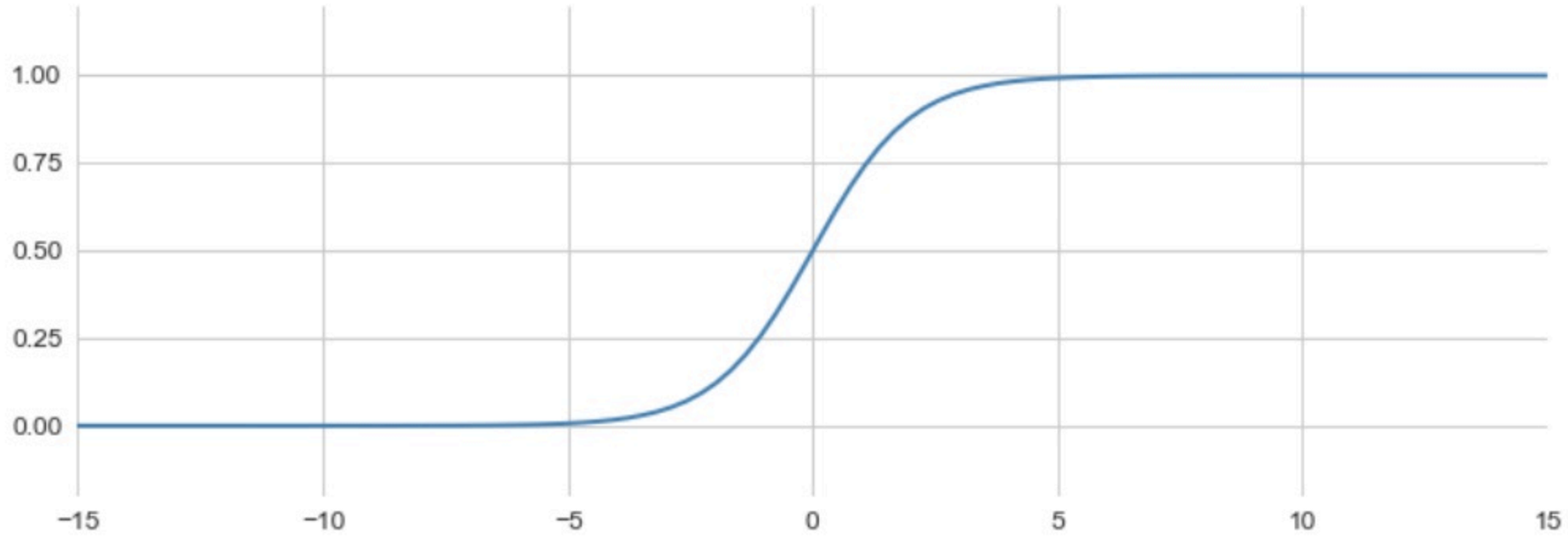


Enter the sigmoid Function

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

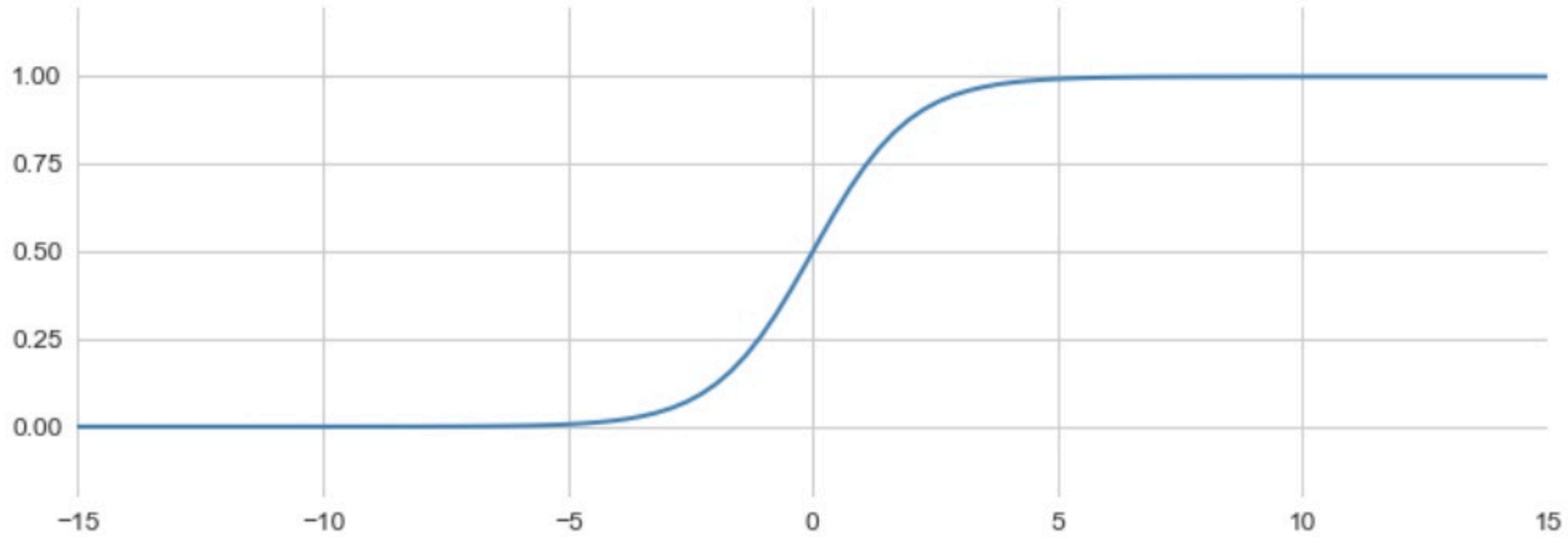
$z \to -\infty \to 0$

$z \to \infty \to 1$



It Has Everything!

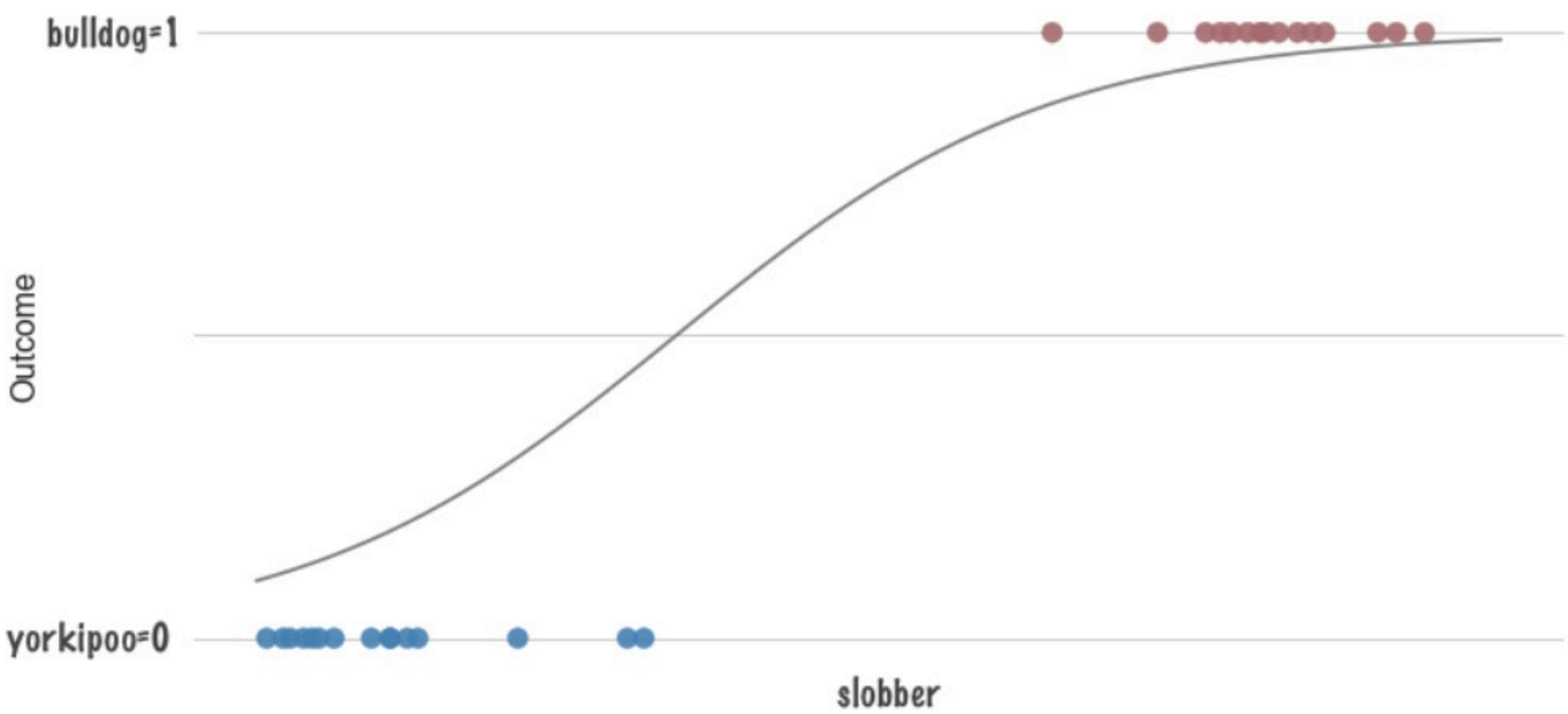
- Behaves like a probability ($0 < \text{sigm}(z) < 1$)
- Distinguishes between points
- It's really smooth (important later)



The Plan

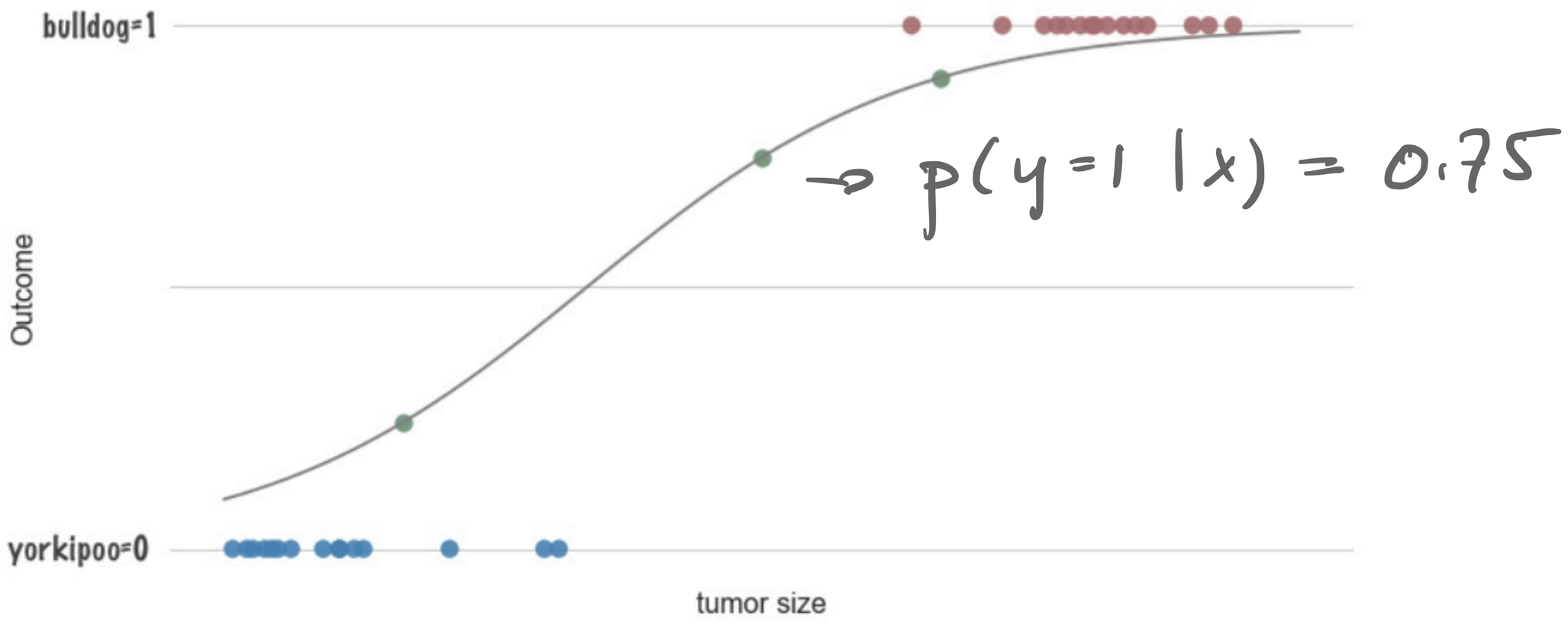
$$p(y = 1 \mid x_1; \hat{\mathbf{w}}) = \text{sigm}(\hat{w}_0 + \hat{w}_1 x_1)$$

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



The Plan

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



The Plan

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

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

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

New dot-product notation:

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

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

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

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

$x = (1, 0, 1, 0, 1)$

$p(y=1 \mid \{mom, nigeria\}) =$

$= sign(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4)$

$= sign(0.1 + 2.0 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1)$

$= sign(2.1) = 0.89$

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

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$
$$0.1 + 2 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1 = 2.1$$

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

$0.89 > 0.5 \Rightarrow$ predict $\hat{y} = 1$

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: $\mathbf{x} = \{mom, nigeria\}$ = SPAM is .89

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$
$$0.1 + 2 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1 = 2.1$$

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

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

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$
$$0.1 + 2 \cdot 0 - 1 \cdot 1 - 0.5 \cdot 0 + 3 \cdot 1 = 2.1$$

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

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

Question: 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: $\mathbf{x} = \{mom, work, viagra, mom\}$ ⇒ (1, 1, 2, 1, 5)

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

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

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$
$$0.1 + 2 \cdot 1 - 1 \cdot 2 - 0.5 \cdot 1 + 3 \cdot 0 = -0.4$$

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

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: $\mathbf{x} = \{\}$ = (1,0,0,0,0)

Yes, this is an empty email $w^T x = 0.1$

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

$$z = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$
$$0.1 + 2 \cdot 0 - 1 \cdot 0 - 0.5 \cdot 0 + 3 \cdot 0 = 0.1$$

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

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.

An Odd(s) View of Logistic Regression

Our inevitable path to Logistic Regression and the **sigmoid** function began with our insistance on modeling $p(y = 1 \mid \mathbf{x}; \mathbf{w})$ as a bonafide probability.

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

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

An Odd(s) View of Logistic Regression

In statistics, the odds of an event occuring 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

ODDS = $\frac{.1}{1-.1} = \frac{1}{9}$ $\quad \text{odds} = \frac{p}{1-p}$ $\quad$ ODDS = $\frac{3/4}{1-3/4} = 3$

Example: If $p = .75$, then odds = 3 and we would say the odds are 3 to 1 in favor of the event occurring.

Example: If $p = 0.1$, then odds = $1/9$ and we would say the odds are 9 to 1 against the event occurring.

Note: While p is constrained to the interval $[0, 1]$, *odds* can range from zero to ∞.

An Odd(s) View of Logistic Regression

In Logistic Regression the event we're modeling is $p(y = 1 \mid \mathbf{x}; \mathbf{w})$ which we model as

$$p = \text{sigm}(\mathbf{w}^T \mathbf{x})$$

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

$$\text{odds} = \frac{\text{sigm}(\mathbf{w}^T \mathbf{x})}{1 - \text{sigm}(\mathbf{w}^T \mathbf{x})} = \exp(\mathbf{w}^T \mathbf{x})$$

EFY: Do this algebra yourself!

That sure looks nice, but it'll look even better if we take the log of both sides

An Odd(s) View of Logistic Regression

Taking the log of both sides, we have

$$\ln(\text{odds}) = \mathbf{w}^T\mathbf{x} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_Dx_D$$

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

$$\text{odds} = \exp(\mathbf{w}^T\mathbf{x})$$

This gives us a new interpretation of the logistic regression weight w_i

An Odd(s) View of Logistic Regression

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

$$\text{odds} = \exp(w_0 + w_1x_1 + \cdots + w_4x_4) \Rightarrow$$

$$\text{odds} = \exp(w_0 + w_1(x_1 + 1) + \cdots + w_4x_4)$$

$w_1x_1 + w_1$

An Odd(s) View of Logistic Regression

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

$$\text{odds} = \exp(w_0 + w_1 x_1 + \cdots + w_4 x_4) \Rightarrow$$

$$\text{odds} = \exp(w_1)\exp(w_0 + w_1 x_1 + \cdots + w_4 x_4)$$

An Odd(s) View of Logistic Regression

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

$$\text{odds} = \exp(w_0 + w_1 x_1 + \cdots + w_4 x_4) \Rightarrow$$

$$\text{odds} = \exp(w_1)\exp(w_0 + w_1 x_1 + \cdots + w_4 x_4)$$

Interpretation: If the number of instances of "viagra" go up in an email by 1, then the odds of the message being SPAM go up by a factor of $\exp(w_1) = \exp(2) \approx 7$

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

In Class

- **Get in groups and get out your computers!**

$$p(y=1 \mid x) = \text{sigm}(w_0 + w_1 x_1 + w_2 x_2)$$

$$\text{sigm}(w^T x) = 0.5$$

$$\Leftrightarrow \quad w^T x = 0$$

$$w_0 + w_1 x + w_2 x_2 = 0$$

$$x_2 = \frac{1}{w_2}(-w_0 - w_1 x_1)$$

In Class

In Class

In Class

In Class

In Class

In Class