Logistic Regression

Boston University CS 506 - Lance Galletti

Logistic Regression

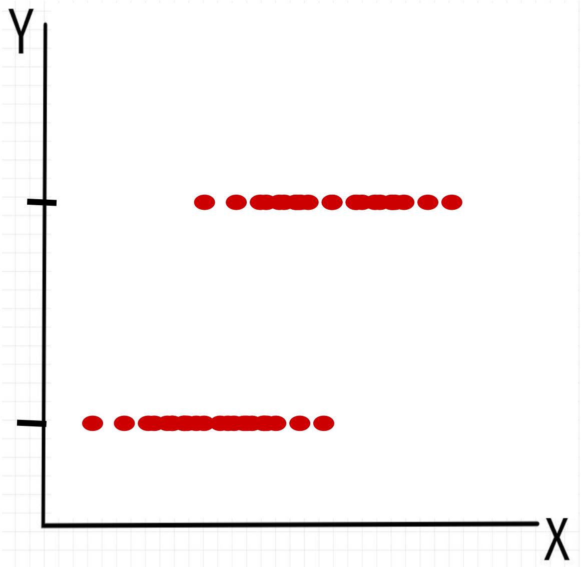
What if y_i is categorical? Can we use a linear function to predict y_i?

Assume we have **2 classes**.

Logistic Regression

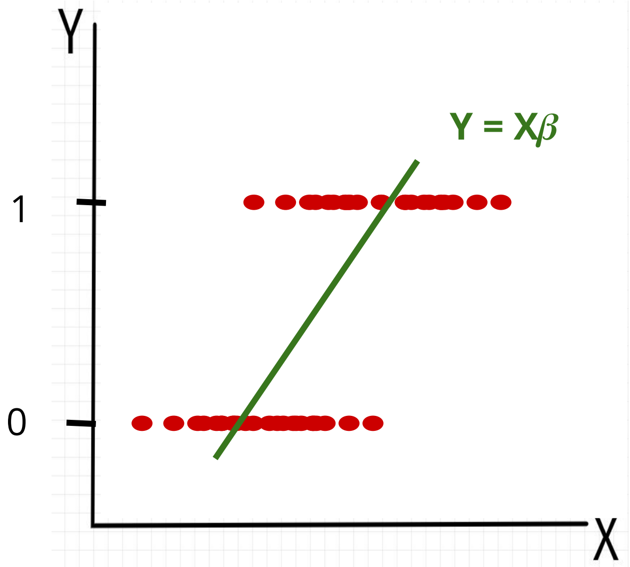
What if y_i is categorical? Can we use a linear function to predict y_i?

Assume we have **2 classes**.



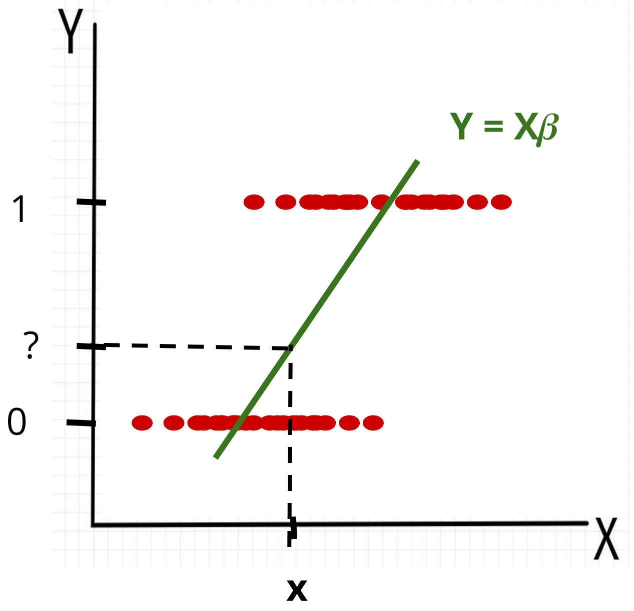
Logistic Regression

What will a linear model look like?



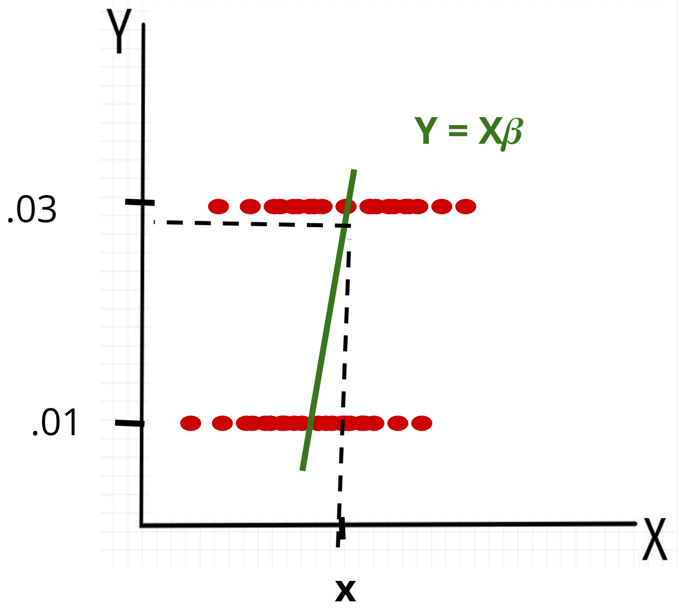
Logistic Regression

What will a linear model look like?



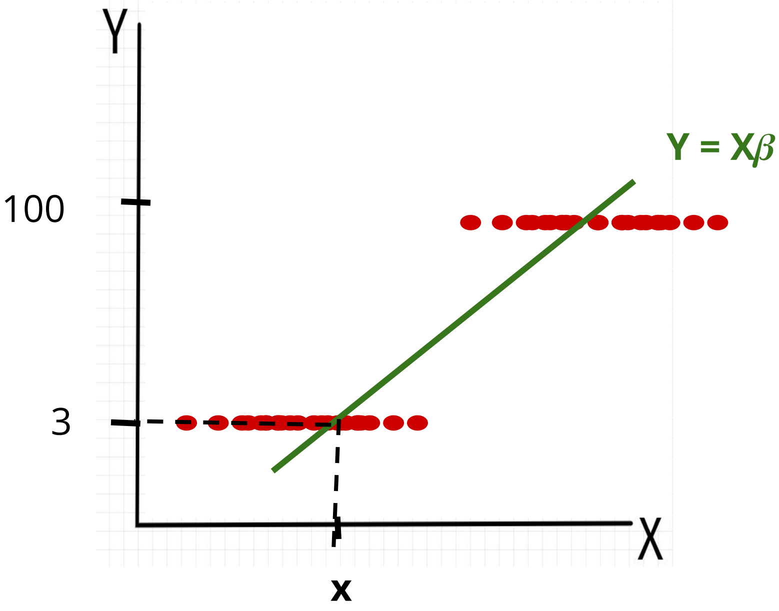
Logistic Regression

What if the numerical values of the classes change?



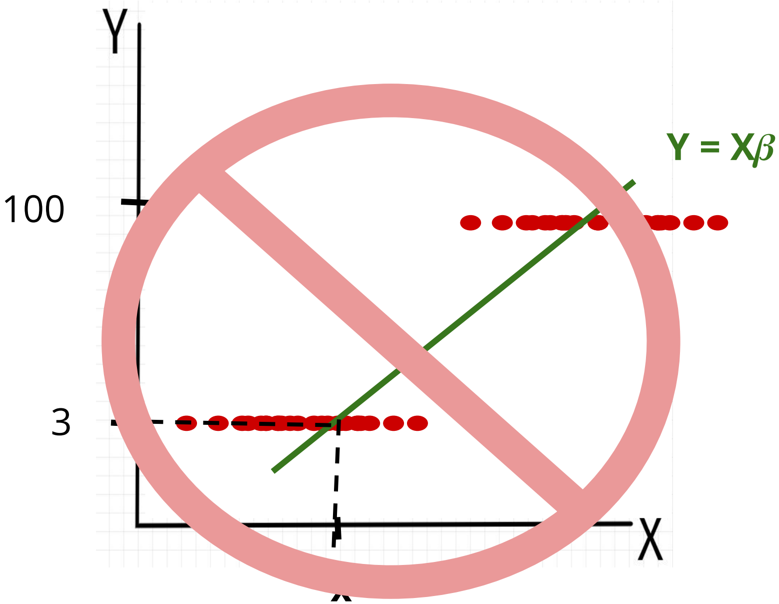
Logistic Regression

What if the numerical values of the classes change?



Logistic Regression

What if the numerical values of the classes change?



Logistic Regression

The numerical values associated with the class are **arbitrary numbers**. A model based on these numbers would be **meaningless…**

So we **should NOT model the class itself** with a linear model.

Logistic Regression

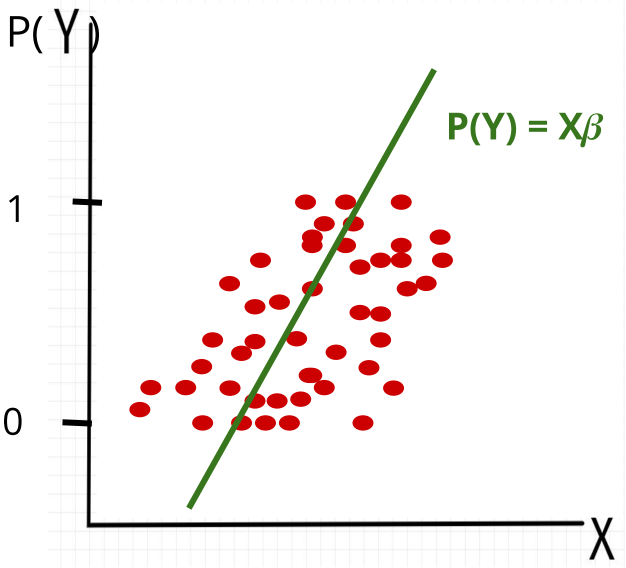
Notice that a linear function will predict a **continuum** of values. So we should find an interpretation / transformation of the class that is **continuous** for us to predict.

Logistic Regression

Can we use the probability of belonging to a given class as a proxy for how confidently we can classify a given point?

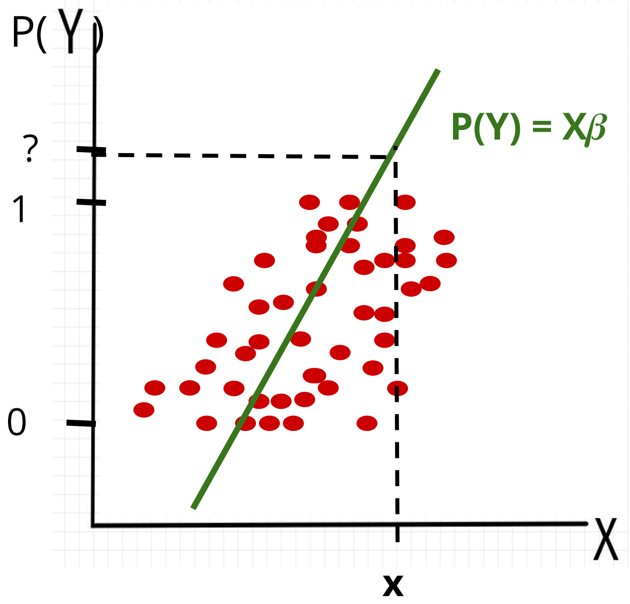
Logistic Regression

Can we use the probability of belonging to a given class as a proxy for how confidently we can classify a given point?



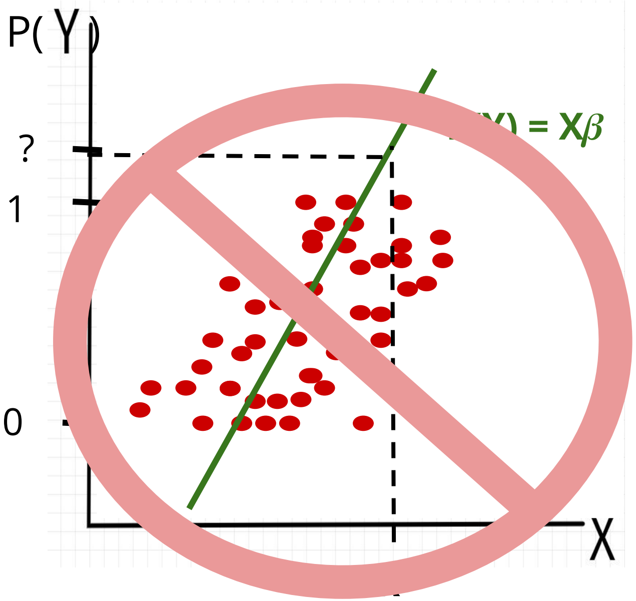
Logistic Regression

Can we use the probability of belonging to a given class as a proxy for how confidently we can classify a given point?



Logistic Regression

Can we use the probability of belonging to a given class as a proxy for how confidently we can classify a given point?



Logistic Regression

So it’s not just a continuum of values - the range of values needs to be (-∞, ∞)!

Define the odds = p / 1 - p where p = P(Y = class 1 | X)

Now the range of $X\boldsymbol{\beta}_{LS}$ is [0, ∞)

In order to get (-∞, ∞), let’s take the log of the odds! This is also convenient numerically because in the odds format, tiny variations in p have large effects on the odds!

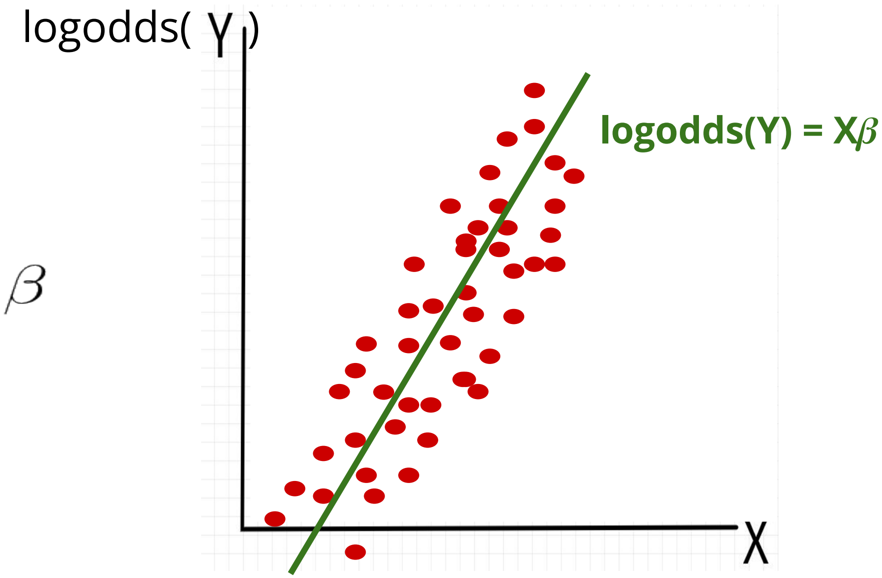
Logistic Regression

Our goal is to fit a linear model to **the log-odds of being in one of our classes** (in the 2-class case) i.e.

$$\log(\frac{P(Y=1|X)}{1-P(Y=1|X)}) = X\beta$$

Logistic Regression

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = X\beta$$



How do we make a prediction with this model?

DECISION RULE:

IF $P(Y=1|X) > ½$ THEN 1 ELSE 0

Logistic Regression

Suppose we have such a model. How do we recover the P(Y=1|X)?

$$\log(\frac{P(Y=1|X)}{1-P(Y=1|X)}) = \alpha + \beta X$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\alpha+\beta X}$$

Logistic Regression

Suppose we have such a model. How do we recover the P(Y=1|X)?

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \alpha + \beta X$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\alpha+\beta X}$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} + 1 = e^{\alpha+\beta X} + 1$$

Logistic Regression

Suppose we have such a model. How do we recover the P(Y=1|X)?

$$\log(\frac{P(Y=1|X)}{1-P(Y=1|X)}) = \alpha + \beta X$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\alpha+\beta X}$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} + 1 = e^{\alpha+\beta X} + 1$$

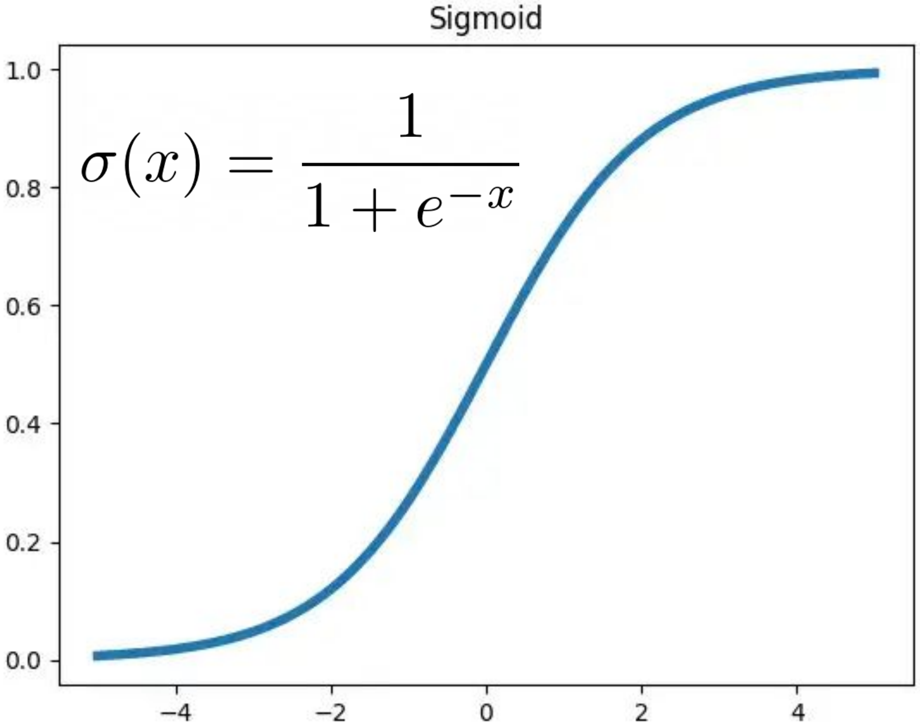
$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\alpha+\beta X} + 1$$

$$P(Y=1|X) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

The function we apply to our probability to obtain the log odds is called the **logit** function. The function used to retrieve our probability from the log odds is called **logit^{-1}** or **sigmoid**

$$\log(\frac{P(Y=1|X)}{1-P(Y=1|X)}) = \alpha + \beta X$$

$$P(Y=1|X) = \sigma(\alpha + \beta x)$$

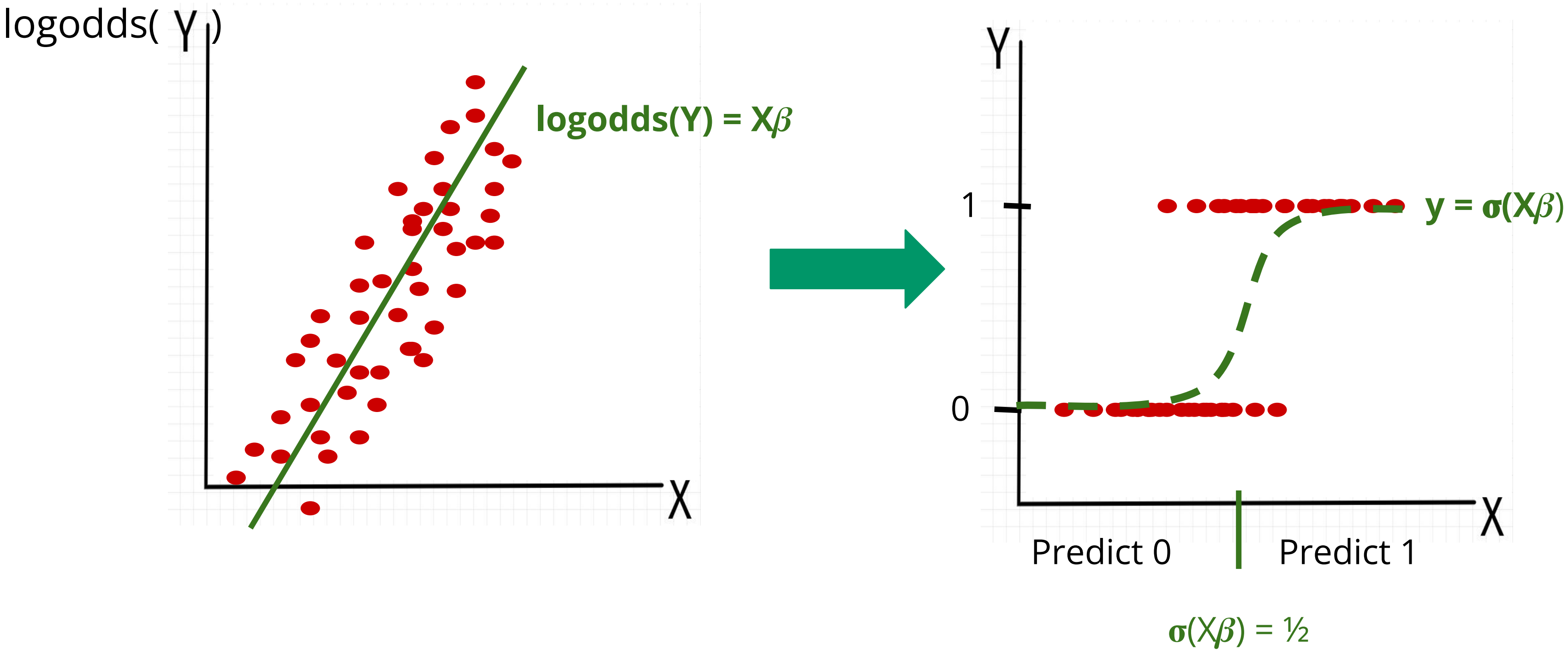


Sigmoid

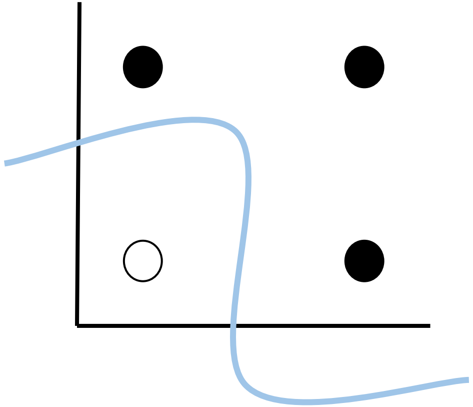
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

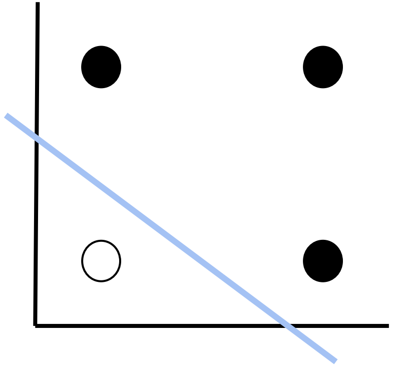
DECISION RULE:

IF P(Y=1|X) > ½ THEN 1 ELSE 0



What does the decision boundary look like?

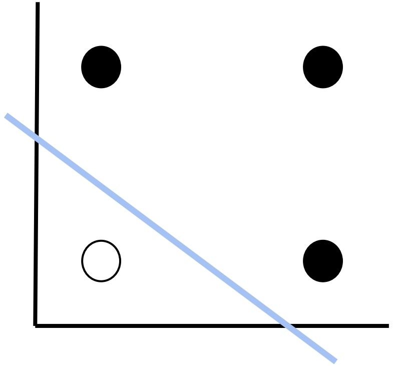




What does the decision boundary look like?

Decision Boundary is where **P(Y = 1 | X) = ½**

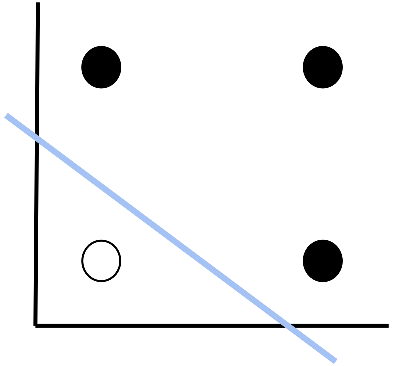
$$P(Y = 1|X) = \frac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$$



What does the decision boundary look like?

Decision Boundary is where **e^{wx+b} = 1**

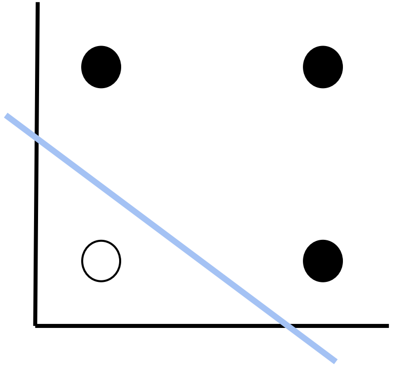
$$P(Y=1|X)=\frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$



What does the decision boundary look like?

Decision Boundary is where **wx+b = 0**

$$P(Y = 1|X) = \frac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$$



Worksheet a) -> c)

Maximum Likelihood Estimator

How do we learn our model? I.e. the α and β parameters.

Maximum Likelihood Estimator

How do we learn our model? I.e. the α and β parameters.

We know:

$$P(y_i|x_i) = \begin{cases} \sigma(\alpha + \beta x_i) & \text{if } y_i = 1 \\ 1 - \sigma(\alpha + \beta x_i) & \text{if } y_i = 0 \end{cases}$$

Maximum Likelihood Estimator

How do we learn our model? I.e. the α and β parameters.

We know:

$$P(y_i|x_i) = \begin{cases} \sigma(\alpha + \beta x_i) & \text{if } y_i = 1 \\ 1 - \sigma(\alpha + \beta x_i) & \text{if } y_i = 0 \end{cases}$$

$$P(y_i|x_i) = \sigma(\alpha + \beta x_i)^{y_i}(1 - \sigma(\alpha + \beta x_i))^{1-y_i}$$

Maximum Likelihood Estimator

So we can define the probability of having seen the data we saw:

$$L(\alpha, \beta) = \prod_{i=1}^{n} P(y_i | x_i)$$

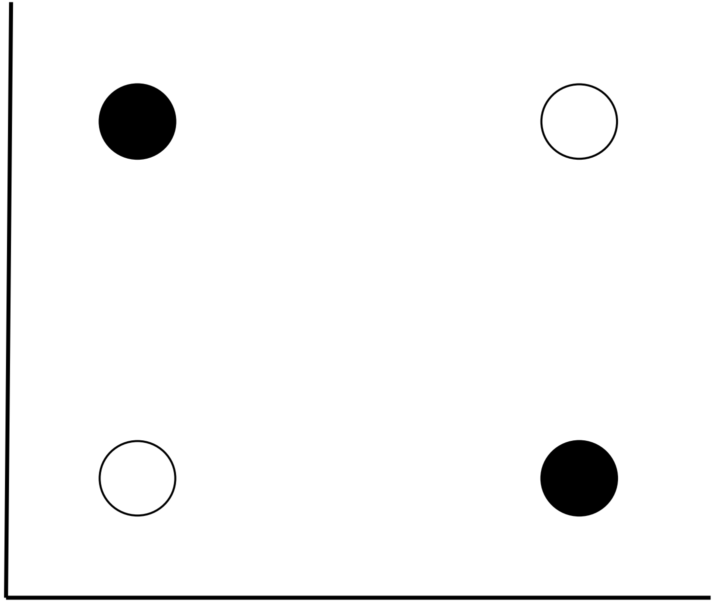
$$= \prod_{i=1}^{n} \sigma(\alpha + \beta x_i)^{y_i} (1 - \sigma(\alpha + \beta x_i))^{1-y_i}$$

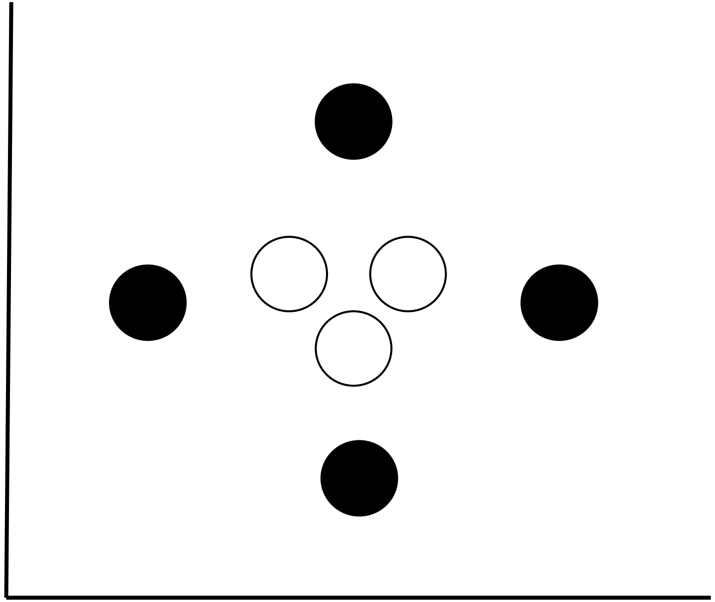
And try to maximize this quantity!

Maximum Likelihood Estimator

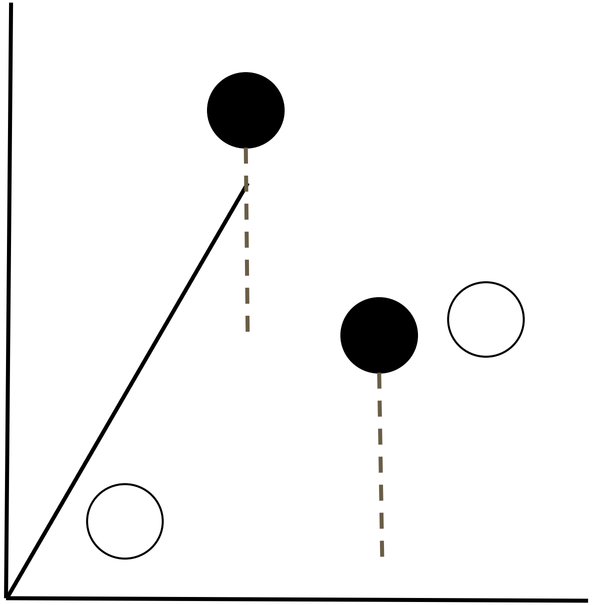
We will learn how to solve this next week - it’s not as simple as linear regression

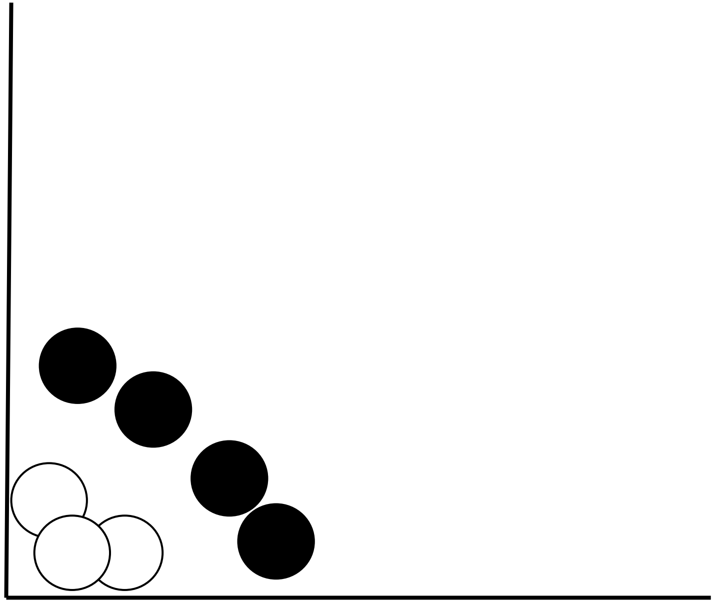
What if the data is not linearly separable



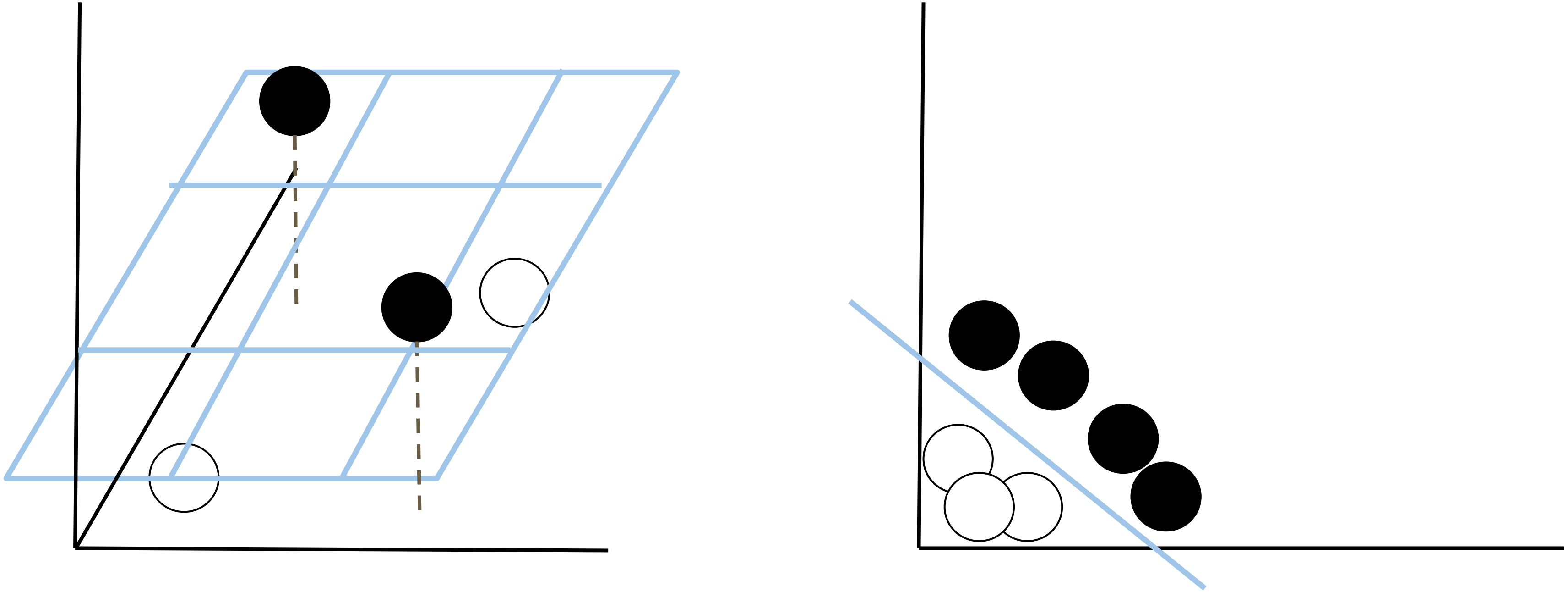


What if the data is not linearly separable



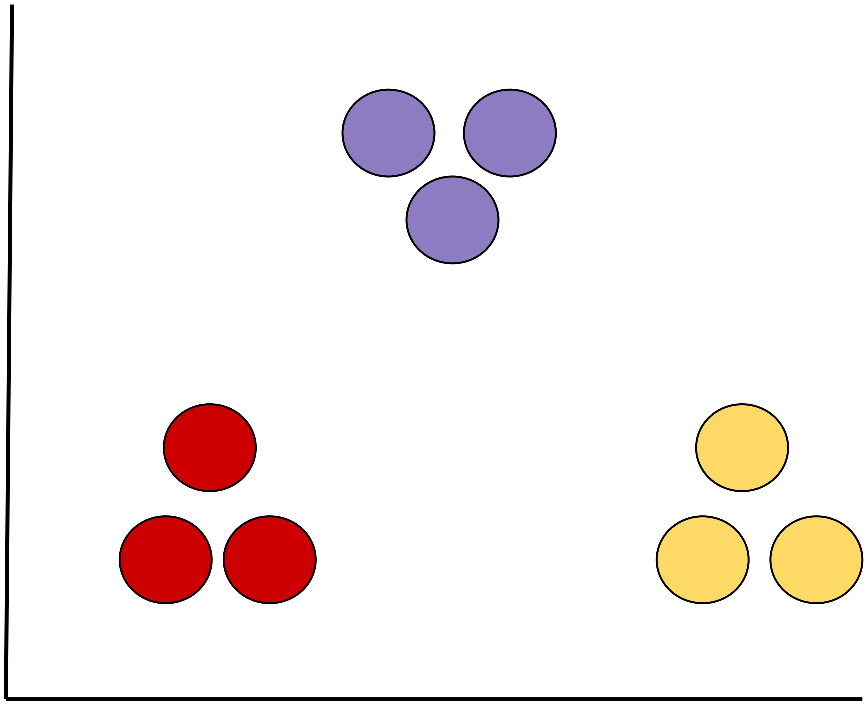


What if the data is not linearly separable



Worksheet d) -> h)

What if there are 3 classes



What if there are 3 classes

Setup:

$$\log\left(\frac{P(Y=0|X)}{P(Y=2|X)}\right)=\beta_0 X$$

$$\log\left(\frac{P(Y=1|X)}{P(Y=2|X)}\right)=\beta_1 X$$

$$P(Y=2|X)=1-(P(Y=1|X)+P(Y=0|X))$$

What if there are 3 classes

$$P(Y = 0|X) = P(Y = 2|X)e^{\beta_0 X}$$

$$P(Y = 1|X) = P(Y = 2|X)e^{\beta_1 X}$$

$$P(Y = 2|X) = 1 - (P(Y = 1|X) + P(Y = 0|X))$$

What if there are 3 classes

$$P(Y = 0|X) = P(Y = 2|X)e^{\beta_0 X}$$

$$P(Y = 1|X) = P(Y = 2|X)e^{\beta_1 X}$$

$$P(Y = 2|X) = 1 - (P(Y = 2|X)e^{\beta_0 X} + P(Y = 2|X)e^{\beta_1 X})$$

What if there are 3 classes

$$P(Y=0|X) = P(Y=2|X)e^{\beta_0 X}$$

$$P(Y=1|X) = P(Y=2|X)e^{\beta_1 X}$$

$$P(Y=2|X) = \frac{1}{1+e^{\beta_0 X}+e^{\beta_1 X}}$$

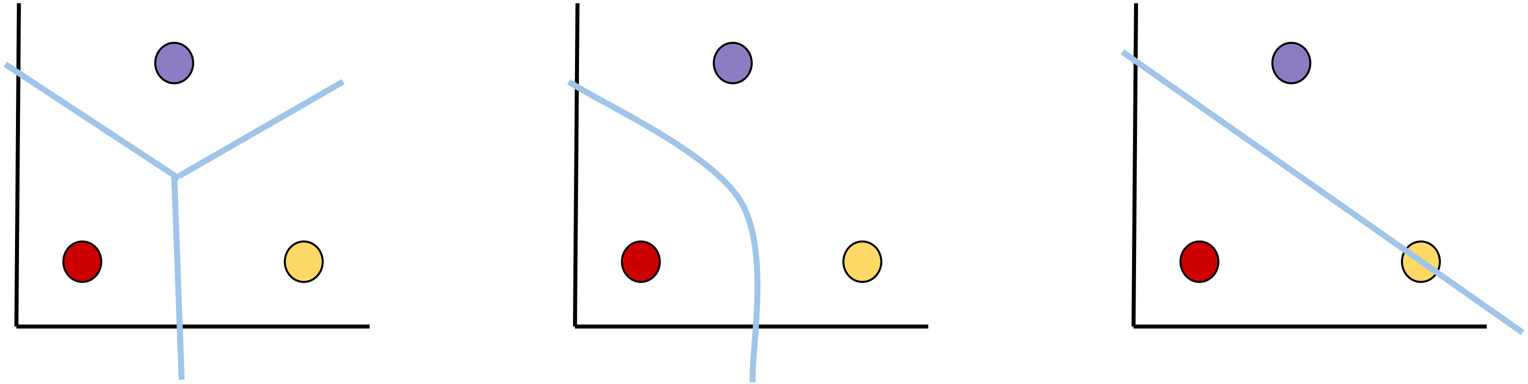
What if there are 3 classes

$$P(Y = 0|X) = \frac{e^{\beta_0 X}}{1 + e^{\beta_0 X} + e^{\beta_1 X}}$$

$$P(Y = 1|X) = \frac{e^{\beta_1 X}}{1 + e^{\beta_0 X} + e^{\beta_1 X}}$$

$$P(Y = 2|X) = \frac{1}{1 + e^{\beta_0 X} + e^{\beta_1 X}}$$

What does the decision boundary look like?



Worksheet i) ->