Logistic Regression

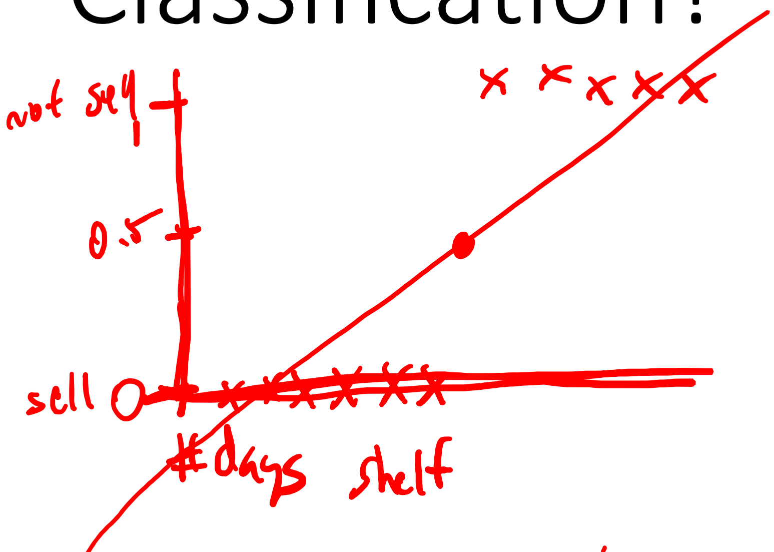
Logistic Regression

independent dependent

$y = mx + b$

- Logistic Regression is a terrible name!
  - It's not regression at all!
  - It's classification
- But as you'll see, how we do it is extremely similar to *linear* regression

Linear Regression for Classification?

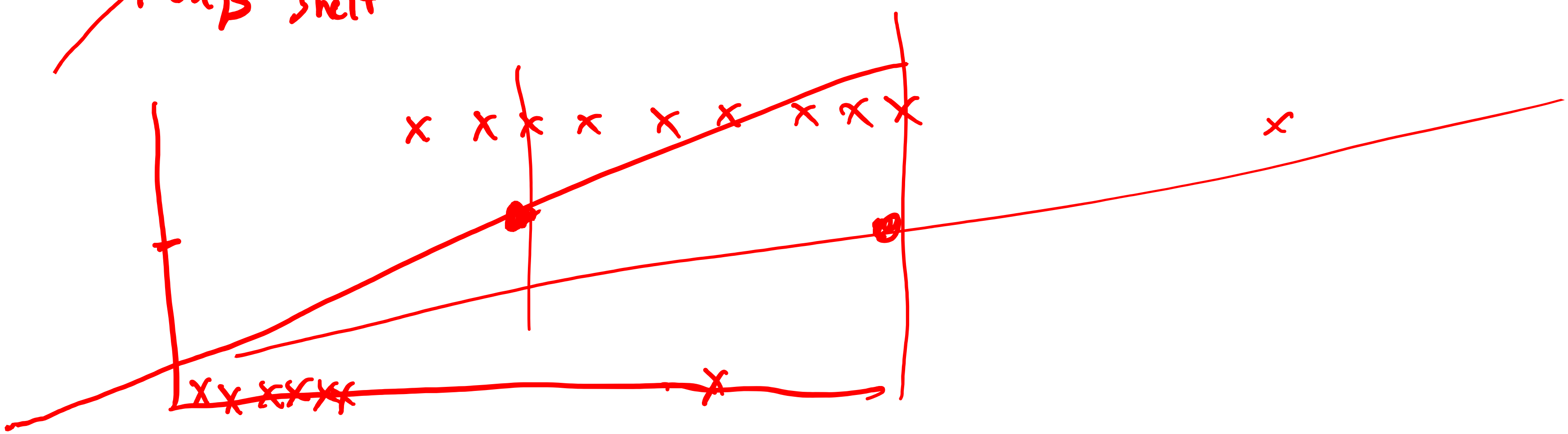


$y \in \{0, 1\}$

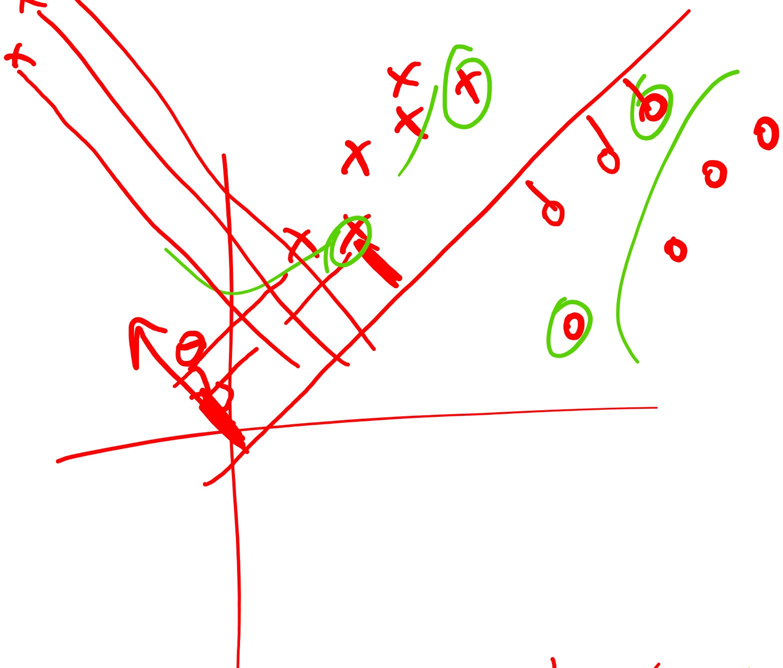
sell no

$\hat{y} \geq 0.5 \rightarrow 1$

$\hat{y} < 0.5 \rightarrow 0$



Logistic Regression – Line as decision boundary



$y \in \{0, 1\}$

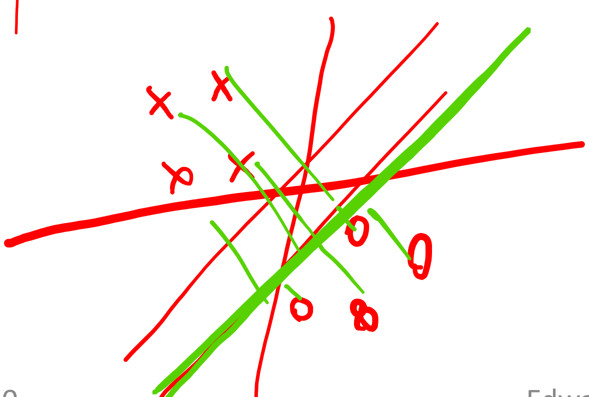
$ax + by - c = 0$

$\theta^T x = 0$

$\underline{\theta \cdot x_i}$ = distance to boundary

$\theta \cdot p_i$ = 1,000 units?

$\underline{\theta x_i} \Rightarrow$ Prob cat $[0 - 1]$



Logistic Regression

- With logistic regression we assume binary classification and want to provide a probability for the positive class:

$$0 \leq P(y = 1) \leq 1$$

- Recall from *linear* regression we computed

$$g(x, \theta) = x\theta$$

- We can alter this for use in computing $P(y = 1)$ as:

$$P(y = 1) = g(x, \theta) = \frac{1}{1 + e^{-x\theta}}$$

sigmoid
logistic
squashing

Logistic Regression

$\theta(x) \to 1000$

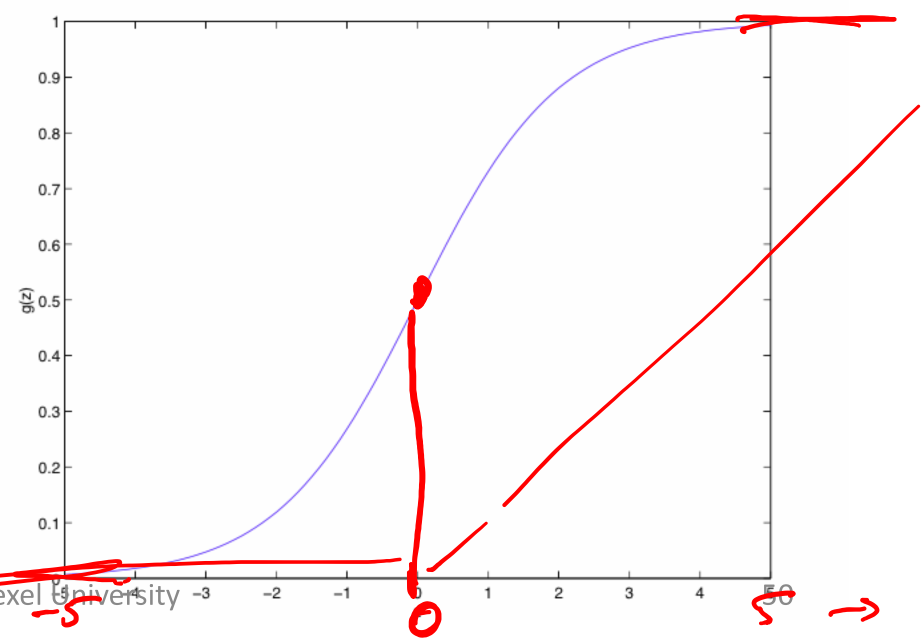
$\to .9999$

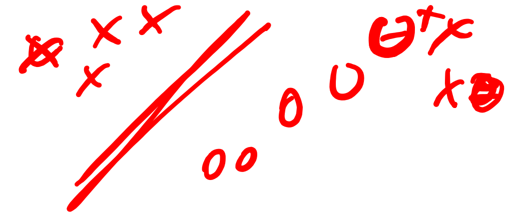
$\theta x = -1000$

$\to 0$

$$P(y=1|x) = g(x,\theta) = \frac{1}{1+e^{-x\theta}}$$

- This function, Let $g(z) = \frac{1}{1+e^{-z}}$ is called the *sigmoid* or *logistic* function
  - Tends to 0 as z decreases
  - Tends to 1 as z increases
- This has the nice characteristic in that it's differentiable
  - Why might that be important?!





Logistic Regression

- If we consider
  - $P(y=1|x,\theta) = \frac{1}{1+e^{-x\theta}} = g(x,\theta)$
- Then we can compute the probability of being from the negative class as:
  - $P(y=0|x,\theta) = 1 - g(x,\theta)$
- Ultimately we want to find the parameters θ to minimize the classification error
  - Or conversely, to find the parameters θ to maximize the correct class likelihood

$P(y|x,\theta) = g(x,\theta)^{y}(1-g(x,\theta))^{(1-y)}$

$y=1 = g(x,\theta)$

$y=0 \; (1-g(x,\theta))$

Fit Parameters Based on Maximum Likelihood

MLE

Maximum Likelihood estimation

- Given a supervised observation (x, y), we can compute the **likelihood** that we are correct as

$$\ell(y|x,\theta) = \big(g(x,\theta)\big)^{y}\big(1 - g(x,\theta)\big)^{(1-y)}$$

- Doing this for the entire dataset, since the observations are *conditionally independent* of one another we have:

iid → independently identically distributed

$$\ell(Y|X,\theta) = \prod_{t=1}^{N} \ell(Y_t, |X_t, \theta) = \prod_{t=1}^{N} (g(X_t,\theta)^{Y_t}\big(1 - g(X_t,\theta)\big)^{(1-Y_t)}$$

Log Likelihood

$\log(ab) = \log a + \log b$

$\log(a^b) = b \log a$

- So what do we do with this likelihood $\ell(Y|X,\theta)$?
  - We want to maximize it!
  - Or minimize $-\ell(Y|X,\theta)$
- So we're going to want to take the derivative
- But taking the derivative of a product of a lot of things involves a very long expansion
- Let's instead first take the log of this
  - Doing so will result in a sum which is easier to take the derivative of.
  - So now we want to *maximize the log likelihood*

Log Likelihood Rules

$\log(a^b) = b \log a$

$\log(ab) = \log a + \log b$

$$\prod_{i=1}^{N} g(x_i,\theta)^{y_i} \cdot (1 - g(x_i,\theta))^{(1-y_i)}$$

$$\sum_{i=1}^{N} y_i \ln g(x_i;\theta) + (1-y_i)\ln(1-g(x_i,\theta))$$

Log Likelihood

- From the properties of logarithms
  - $\log_b(mn) = \log_b(m) + \log_b(n)$
  - $\log_b(m^n) = n \cdot \log_b(m)$
- Returning to our likelihood for a single observation, $\ell(y, |x, \theta)$, we get
  - $\ell(y|x, \theta) = ln\left(g(x, \theta)^y \left(1 - g(x, \theta)\right)^{(1-y)}\right)$
  - $= \ln(g(x, \theta)^y) + \ln\left(\left(1 - g(x, \theta)\right)^{1-y}\right)$
  - $= y\, ln\big(g(x, \theta)\big) + (1 - y)\ln\big(1 - g(x, \theta)\big)$

Log Likelihood

- Since we're taking the log of product of this for each instance we get a sum!

$$\ell(Y|X,\theta) = \log P(Y|X,\theta) = \sum_{t=1}^{N} Y_t\, ln\big(g(X_t,\theta)\big) + (1 - Y_t)\ln\big(1 - g(X_t,\theta)\big)$$

- Ideally we'd like to take the derivative of this with respect to θ, set it equal to zero, and solve for θ to find the maxima
  - The closed form approach
  - But this doesn't exist ☹
- So what's our other approach?
  - Do partial derivatives on the parameters and use gradient descent! (actually in this case gradient ascent, since we're trying to maximize)
  - First do this for a single observation and single parameter
    - Then vectorize!

Log Likelihood Derivation

$$\left(\frac{y}{g(x,\theta)} - \frac{1-y}{1-g(x,\theta)}\right) \cdot x_j \, \cancel{g(x,\theta)\,(1-g(x,\theta))}$$

$$\theta = \theta - \frac{\eta}{N} X^T \left(g(x,\theta) - y\right)$$

x_θ

$$\frac{(1-g(x,\theta))\,y}{(1-g(x,\theta))(g(x,\theta))} - \frac{g(x,\theta)\,(1-y)}{g(x,\theta)\,(1-g(x,\theta))}$$

$$\frac{\partial J}{\partial \theta_j} = \left(y - g(x,\theta)\right) \cdot x_j$$

$$X = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_N \end{bmatrix} \cdot \begin{bmatrix} (y - g(x,\theta)) \\ (y - g(x,\theta)) \\ \vdots \\ (y - g(x,\theta)) \end{bmatrix}$$

$$\frac{y - \cancel{y\, g(x,\theta)} - g(x,\theta) + \cancel{y\, g(x,\theta)}}{\cancel{(1-g(x,\theta))\, g(x,\theta)}}$$

$$\Theta = \Theta + \frac{\eta}{N} X^T \left(y - g(x,\theta)\right)$$

Log Likelihood Derivation

$$x_i \cdot g(x, \theta) \cdot \frac{e^{-x\theta}}{1+e^{-x\theta}}$$

$$\frac{1+\cancel{e^{-x\theta}}}{1+e^{-x\theta}} - \frac{\cancel{e^{-x\theta}}}{1+e^{-x\theta}}$$

thing

$$x_i \cdot g(x,\theta) \cdot (1 - g(x,\theta))$$

$$\frac{1}{1+e^{-x\theta}}$$

$$1 - \text{thing} = g(x,\theta)$$

$$1 - g(x,\theta) = \text{thing}$$

To Maximum Likelihood

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \frac{\partial}{\partial \theta_j} \left(y \ln\big(g(x,\theta)\big) + (1-y) \ln\big(1-g(x,\theta)\big) \right)$$

- First off, a reminder…

$$\frac{\partial}{\partial x}(\ln x) = \frac{1}{x} \cdot \frac{\partial}{\partial x}(x)$$

- Therefore

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \frac{y}{g(x,\theta)} \frac{\partial}{\partial \theta_j}\big(g(x,\theta)\big) + \frac{(1-y)}{1-g(x,\theta)} \frac{\partial}{\partial \theta_j}\big(1-g(x,\theta)\big)$$

- But what is $\frac{\partial}{\partial \theta_j}(g(x,\theta)$?

To Maximum Likelihood

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \frac{y}{g(x,\theta)} \frac{\partial}{\partial \theta_j}\left(g(x,\theta)\right) + \frac{(1-y)}{1-g(x,\theta)} \frac{\partial}{\partial \theta_j}\left(1-g(x,\theta)\right)$$

- $\frac{\partial}{\partial \theta_j} g(x,\theta) = \frac{\partial}{\partial \theta_j}\left(\frac{1}{1+e^{-x\theta}}\right) = \frac{\partial}{\partial \theta_j}\left(1+e^{-x\theta}\right)^{-1}$
- $= -1\left(0 - x_j e^{-x\theta}\right)\left(1+e^{-x\theta}\right)^{-2} = \frac{x_j e^{-x\theta}}{\left(1+e^{-x\theta}\right)^2}$
- $= x_j \frac{1}{1+e^{-x\theta}} \frac{e^{-x\theta}}{1+e^{-x\theta}}$
- $= x_j g(x,\theta)\left(1-g(x,\theta)\right)$

To Maximum Likelihood

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = \frac{y}{g(x,\theta)} \frac{\partial}{\partial \theta_j} \big(g(x,\theta)\big) + \frac{(1-y)}{1-g(x,\theta)} \frac{\partial}{\partial \theta_j} \big(1-g(x,\theta)\big)$$

- From the previous slide we have

$$\frac{\partial}{\partial \theta_j} g(x,\theta) = x_j g(x,\theta)\big(1-g(x,\theta)\big)$$

- Putting it all together (and simplifying) we get:

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = x_j\big(y-g(x,\theta)\big)$$

To Maximum Likelihood

$$\frac{\partial}{\partial \theta_j} \ell(y|x,\theta) = x_j\big(y - g(x,\theta)\big)$$

- Vectorizing this for all parameters we have

$$\frac{\partial \ell}{\partial \theta} = x^T\big(y - g(x,\theta)\big)$$

- Vectorizing this to be the mean gradient over all observations we have:

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{N} X^T\big(Y - g(X,\theta)\big)$$

Gradient Ascent Rule

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{N} X^T \big(Y - g(X, \theta)\big)$$

- We want this to go towards a maxima
- So let's update θ as

$$\theta := \theta + \eta \left(\frac{\partial}{\partial \theta} \ell(Y|X, \theta) \right)$$

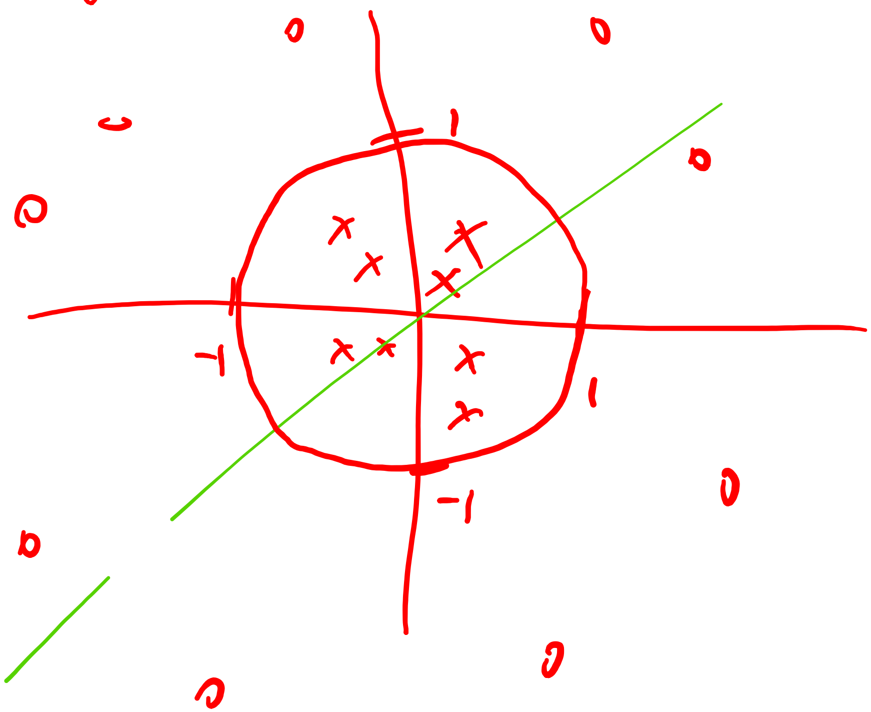
$$\theta = \theta + \frac{\eta}{N} X^T \big(Y - g(X, \theta)\big)$$

- This is (almost) the same form as the least squared error for linear regression!!!!

Non-linear decision boundaries

$\theta_0 + \theta_1 x_1$

$\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 \ldots$



$$\begin{array}{ccccc} -1 & 0 & 0 & 1 & 1 \\ \theta_0 & +\theta_1 x_1 & +\theta_2 x_2 & +\theta_3 x_1^2 & +\theta_4 x_2^2 \end{array}$$

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \Rightarrow -1 + x_1^2 + x_2^2 = 0$$

$x_1^2 + x_2^2 = 1$

$x^2 + y^2 = 1$