DATA SCIENCE
LOGISTIC REGRESSION & CONFUSION MATRIX

I. WHAT IS LOGISTIC REGRESSION?

II. LOG, E, ODDS, AND LOG ODDS

III. REGRESSION: FROM LINEAR TO LOGISTIC

IV. INTERPRETING COEFFICIENTS

V. CONFUSION MATRIX

I. WHAT IS LOGISTIC REGRESSION?

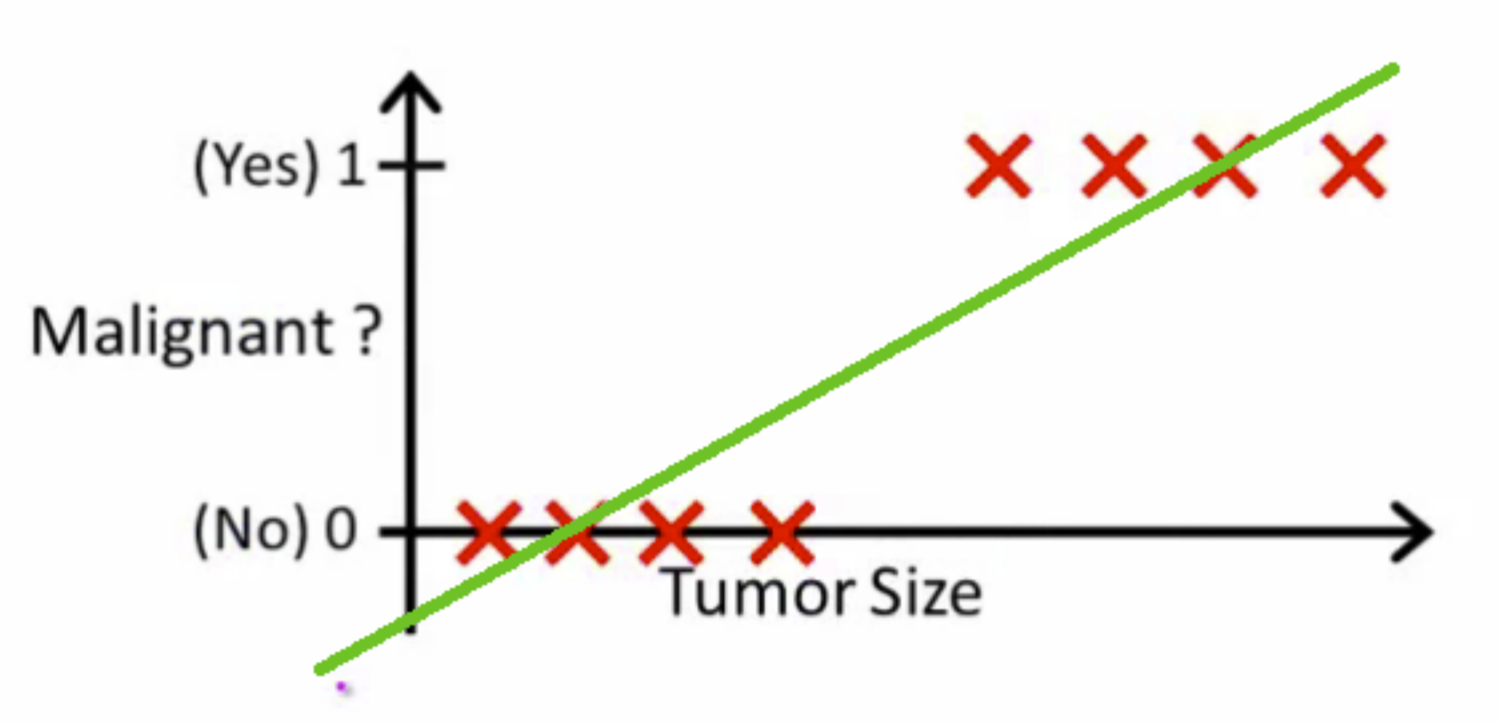
Q: What is logistic regression?

A: A generalization of the linear regression model used for *classification* problems.

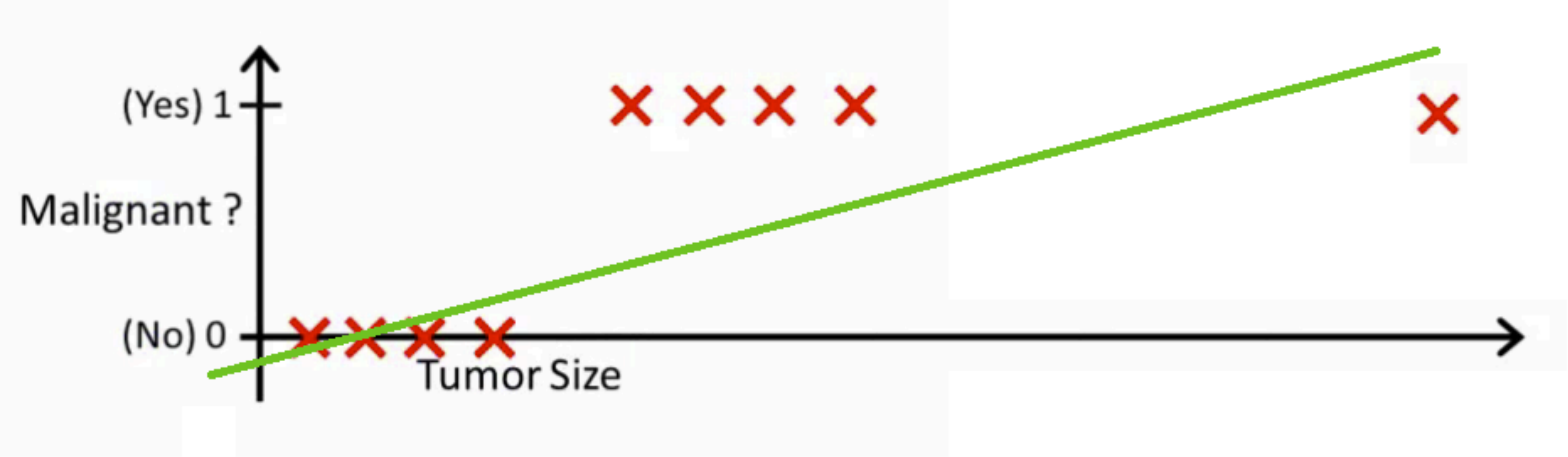
The output of logistic regression is a probability of being in a specific class, i.e. it falls between 0 and 1.

- Why not just use Linear Regression with a threshold?

```
lm = LinearRegression()
lm.fit(X_train, y_train)
y_pred_prob = lm.predict(X_test)
y_pred = np.where(y_pred_prob < 0.5, 0, 1)
```



Source: Andrew Ng, "Introduction to Machine Learning"



Source: Andrew Ng, “Introduction to Machine Learning”

II. LOG, E, ODDS, AND LOG ODDS

- e is the base rate of growth for continually growing processes
- You may remember this from compound interest formulas.
- When you see “log”, it usually means “ln”.
- e and ln are inverses of each other.
- $e^{\ln(x)} = x$
- $\ln(e^x) = x$

- Probability is a measure of likelihood that an event will occur. π
- 1 – Probability is the likelihood of an event not occurring. $1-\pi$
- The odds are the probability that an event will occur divided by the probability that it won't occur.
- The log odds are the natural logs of the odds.

$$Odds = \frac{\pi}{1-\pi} \qquad Log-Odds = \ln\left(\frac{\pi}{1-\pi}\right)$$

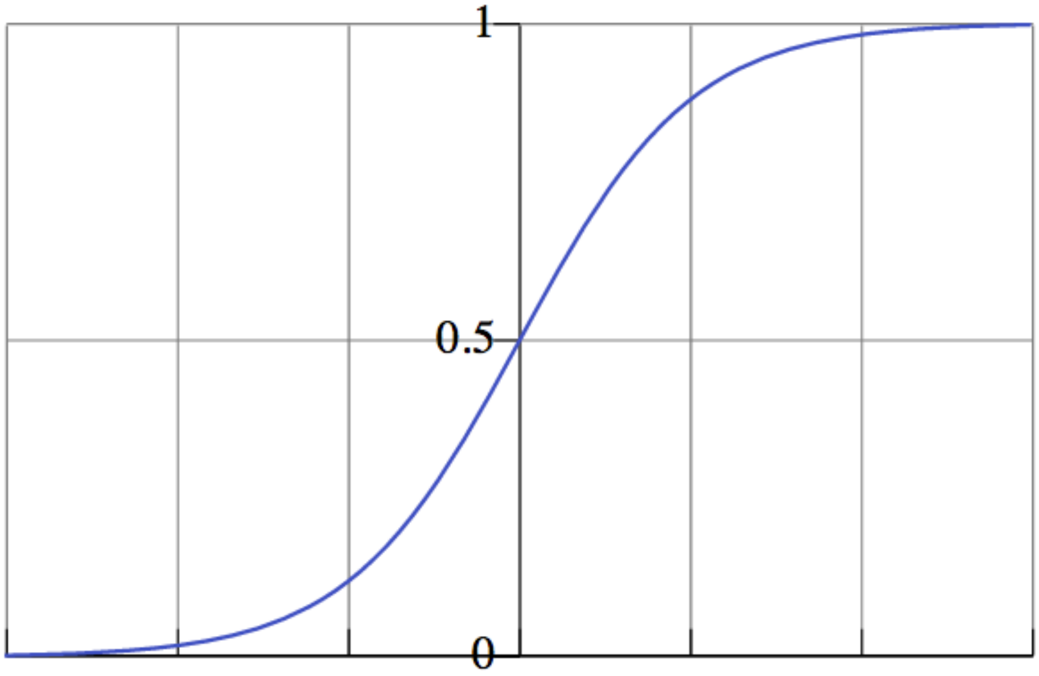
- As the probability increases, the odds increase.
- As the odds increase, the log odds increase.
- Take three minutes to confirm that you get the numbers below for odds and log odds.

Probability	Odds	Log odds
0.01	0.0101	-4.5951
0.25	0.3333	-1.0986
0.50	1.0	0
0.75	3.0	1.0986
0.99	99	4.5951

III. REGRESSION: FROM LINEAR TO LOGISTIC

- In linear regression, we used a set of predictors to predict the value of a continuous response.
- In logistic regression, we use a set of predictors to predict the *probability* of (binary) class membership.
- These probabilities are mapped to *class labels*, allowing us to use this for classification.

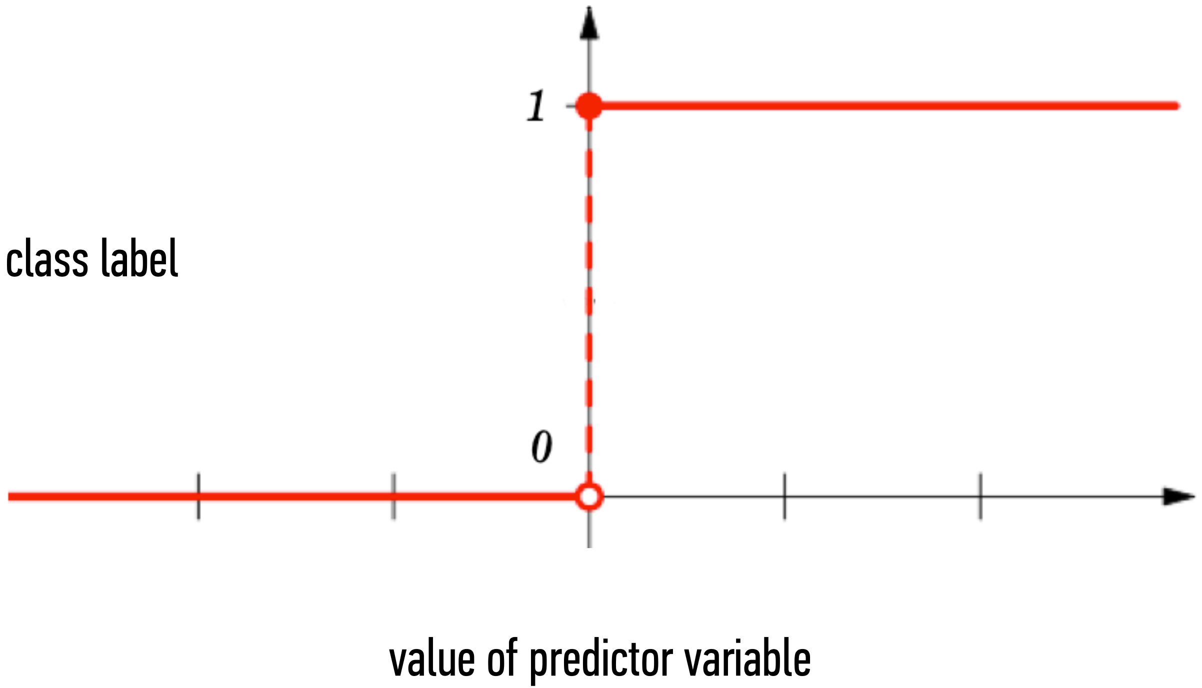
probability of belonging to class



value of predictor variable

NOTE

Probability predictions look like this.



NOTE

Probabilities are "snapped" to class labels (e.g., by threshholding at 50%).

- Exercise: In the following examples, should linear or logistic regression be used?
  - Predict how much money you'll spend on a rental from Air Bnb.
  - Predict whether an email is marked as spam or not.
  - Predict the salary of a new college grad.
  - Predict whether the salary of a new college grad is greater than or less than the median American income.

- One of the key differences between linear and logistic regression is the response variable (what you're predicting).
- In linear regression, the response is modeled by a linear combination of the predictors.

$$Y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \ldots + \varepsilon = X\beta + \varepsilon$$

- In logistic regression, the *log odds* of the outcome is modeled by a linear combination of the predictors.

$$Y = P(event)$$

$$\ln\left(\frac{Y}{1-Y}\right) = X\beta + \varepsilon$$

- Solving the previous equation, we get...

$$\ln\left(\frac{Y}{1-Y}\right) = X\beta + \varepsilon \quad \rightarrow \quad Y = \frac{e^{X\beta}}{1+e^{X\beta}}$$

- We now have an equation to predict the probability of class membership.

IV. INTERPRETING COEFFICIENTS

- Now that we have an equation, what do the coefficients mean?
- For every unit increase in X, there is a β increase in the log odds of class membership and vice versa.
- For every unit increase in X, there is a e^{β} increase in the odds of class membership and vice versa.
- This doesn't mean for every increase in X, you add e^{β} to the previous odds. It means for every increase in X, you multiple e^{β} to the previous odds.

- Let's say I am trying to predict whether your heart is unhealthy or not, yes or no.
- I have one predictor, number of cheeseburgers eaten.
- If I fit a logistic regression model to my data, I get a coefficient of 0.16288268. This means for every cheeseburger I eat, I increase the odds of having an unhealthy heart by exp(0.16288268) = 1.18.
- You could also interpret this as increase the odds by 18%.

V. CONFUSION MATRIX

CONFUSION MATRIX

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Basic Terminology:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)

Accuracy:

- Overall, how often is it **correct**?
- (TP + TN) / total = 150/165 = 0.91

Misclassification Rate (Error Rate):

- Overall, how often is it **wrong**?
- (FP + FN) / total = 15/165 = 0.09

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Sensitivity:

- When actual value is **positive**, how often is prediction **correct**?
- TP / actual yes = 100/105 = 0.95
- "True Positive Rate" or "Recall"

False Positive Rate:

- When actual value is **negative**, how often is prediction **wrong**?
- FP / actual no = 10/60 = 0.17

Specificity:

- When actual value is **negative**, how often is prediction **correct**?
- TN / actual no = 50/60 = 0.83