CDS

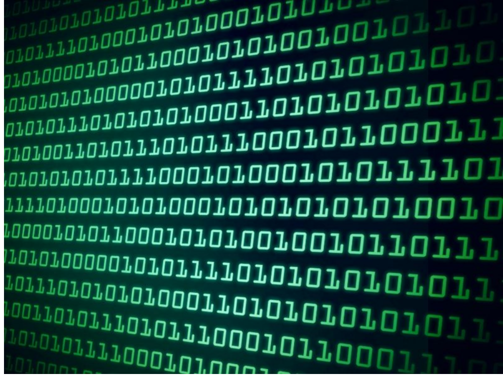
Cornell Data Science

Logistic Regression and Decision Trees

Recall: Regression is powerful.

http://cdn3-www.craveonline.com/assets/uploads/2015/07/Mission-Impossible-2.jpg

Regression for binary outcomes



Regression can also be used to:

- Detect whether someone is at risk for heart disease given health and family history
- Accept/reject applicants to Cornell Data Science based on GPA and performance in data science course

These are called **binary classification** problems.

http://www.johngarvens.com/wp-content/uploads/2013/03/binary-Language.jpg

Logistic Regression

Use a set of continuous variables ($x_{i,}$'s) to perform binary classification. Yields the **probability** that the outcome is 1.

Basic formula:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

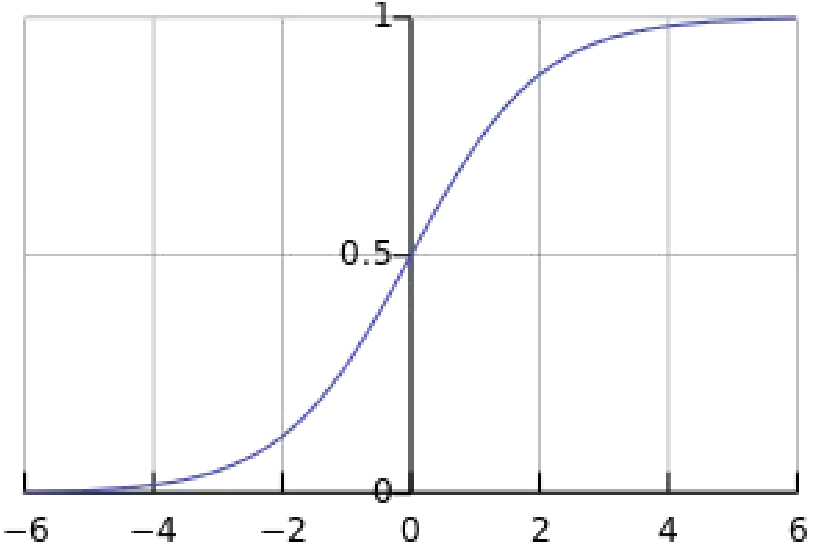
$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

(Recognize this?)

Logistic Function

Here's what $F(x)$ looks like.

Fancy name: **Sigmoid** function

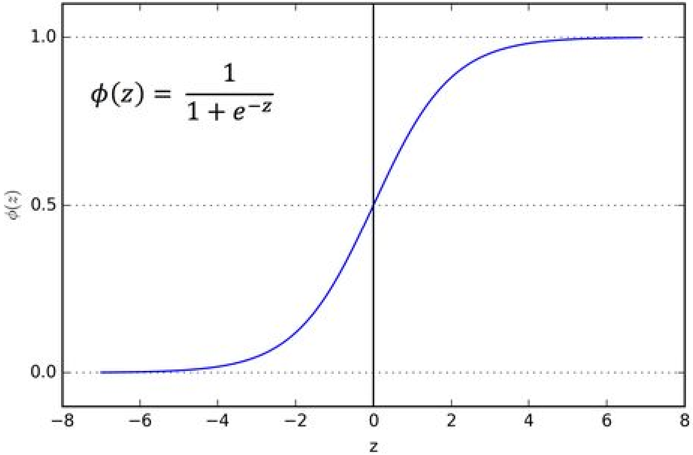


https://upload.wikimedia.org/wikipedia/commons/thumb/8/88/Logistic-curve.svg/320px-Logistic-curve.svg.png

Threshold

Where between 0 and 1 do we draw the line?

- *F(x)* below threshold: predict 0
- *F(x)* above threshold: predict 1



https://sebastianraschka.com/images/faq/logisticregr-neuralnet/sigmoid.png

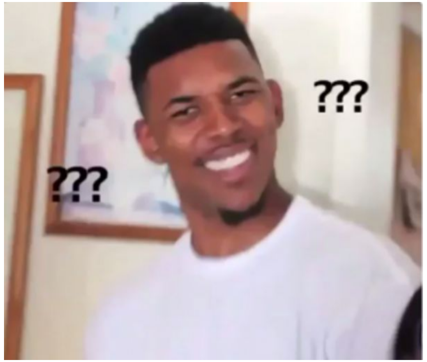
Making predictions

We now know enough to start making predictions.

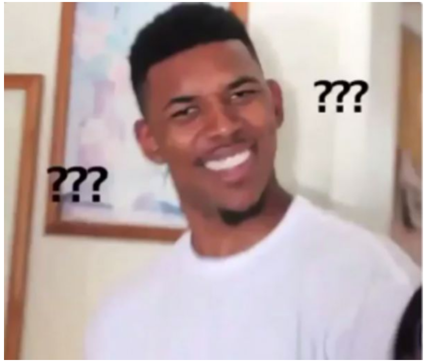
- Use the `glm` function with `family = “binomial”` to create our model.
- Use the `predict` function to predict probabilities.
- Use the `table` function to predict 0’s and 1’s based on a chosen threshold.

Demo time!

Confusion Matrix



	p’ (Predicted)	n’ (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative



Sensitivity

$$Sensitivity = True\ Positive / (True\ Positive + False\ Negative)$$

Also called **True Positive Rate**.

How many positives are correctly identified as positives?

Useful for:

- Airport security
- Initial diagnosis of fatal disease

https://cdn.theatlantic.com/assets/media/img/mt/2015/06/image42/lead_960.jpg?1433269612

Specificity

Specificity = True Negative/
(True Negative + False Positive)

Also called a **True Negative Rate**.

How many true negatives are classified as negative?

(The converse of specificity.)

<u>Sensitivity vs. specificity: Important trade-off!</u>

Question:

Name some examples of situations where you’d want to have a high specificity.

Overall Accuracy and Error Rate

Overall accuracy - proportion of all predictions that are true positives and true negatives

Accuracy =
(True Positive + True Negative)/Total

Overall error rate - proportion of all predictions that are false positives and false negatives

Error Rate =
(False Positive + False Negative)
/Total

Example

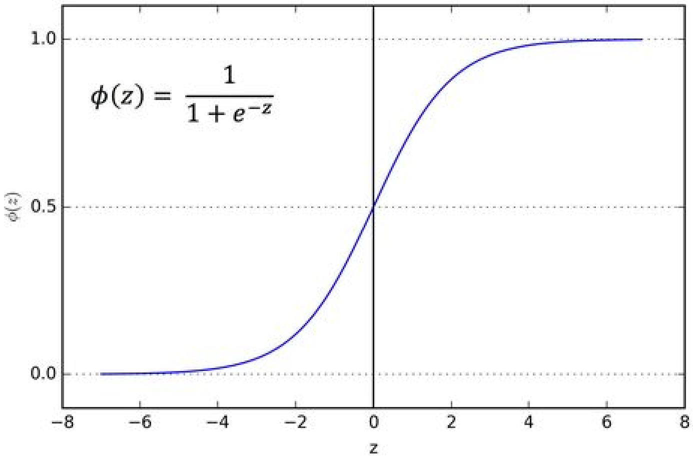
Given this confusion matrix, what is the:

- Specificity?
- Sensitivity?
- Overall error rate?
- Overall accuracy?

	p’ (Predicted)	n’ (Predicted)
p (Actual)	146	32
n (Actual)	21	590

Thresholds matter

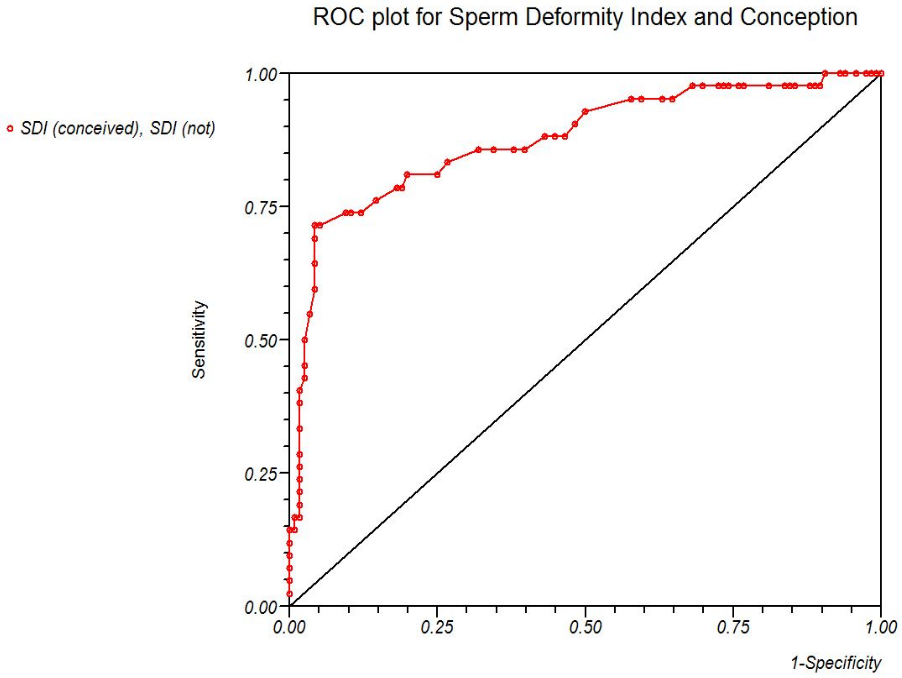
- Low threshold
  - **Lower** specificity
  - **Higher** sensitivity
- High threshold
  - **Higher** specificity
  - **Lower** sensitivity



https://sebastianraschka.com/images/faq/logisticregr-neuralnet/sigmoid.png

ROC Curve

- Visual representation of specificity vs sensitivity tradeoff.
- Allows us to choose a threshold according to our priorities



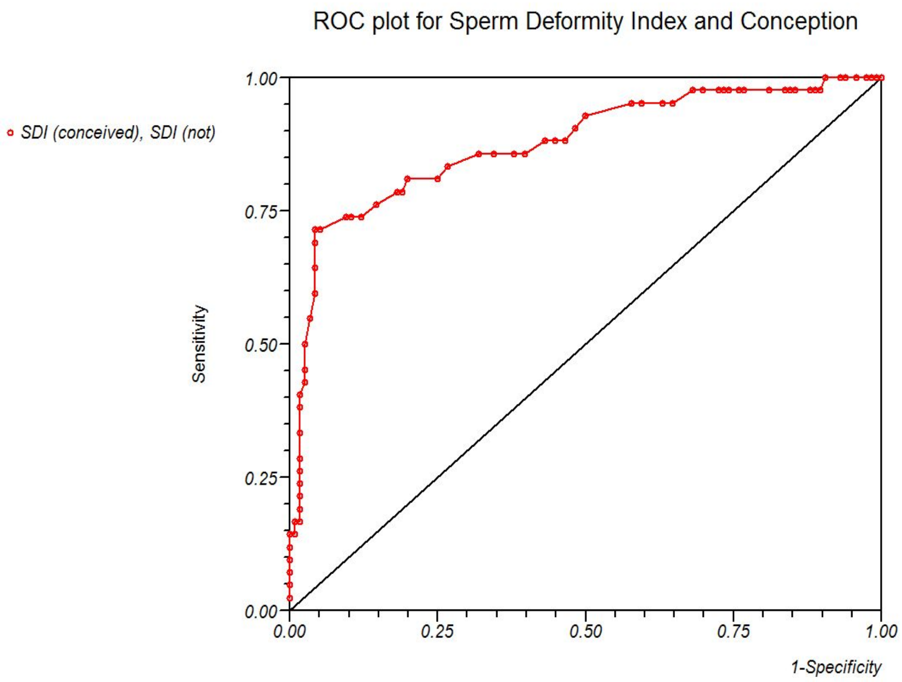
Area Under Curve

$$\text{AUC} = \int ROC\text{-}curve$$

Always between 0.5 and 1.

Interpretation:

- 0.5: Worst possible model
- 1: Perfect model



Pitfalls of Regression

Let’s build a model for predicting supreme court decisions.

- **Dependent variable**: Did the supreme court overturn the lower court’s decision?

- **Independent variables:** properties of the case
  - Lower court, issue, type of people involved, ideological direction of lower court...

http://static.makers.com/Ruth-Bader-Ginsburg-Women-On-The-Bench.jpg

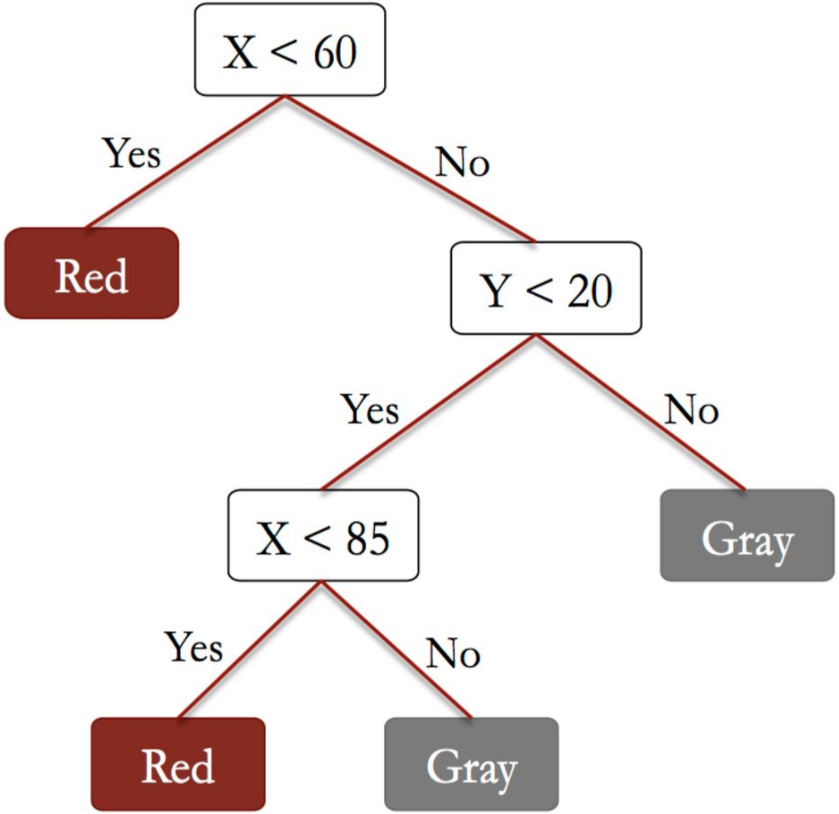
Pitfalls of Regression

- Linear and Logistic Regression assumes linearity
- Gets significant variables and their weights
  - If case is from 2nd circuit court: +1.66
  - If case is from 4th circuit court: +2.82
  - If lower court decisions was liberal: -1.22
- But what does that mean???
  - Difficult to tell which properties are more important
  - Hard to use this data to make prediction

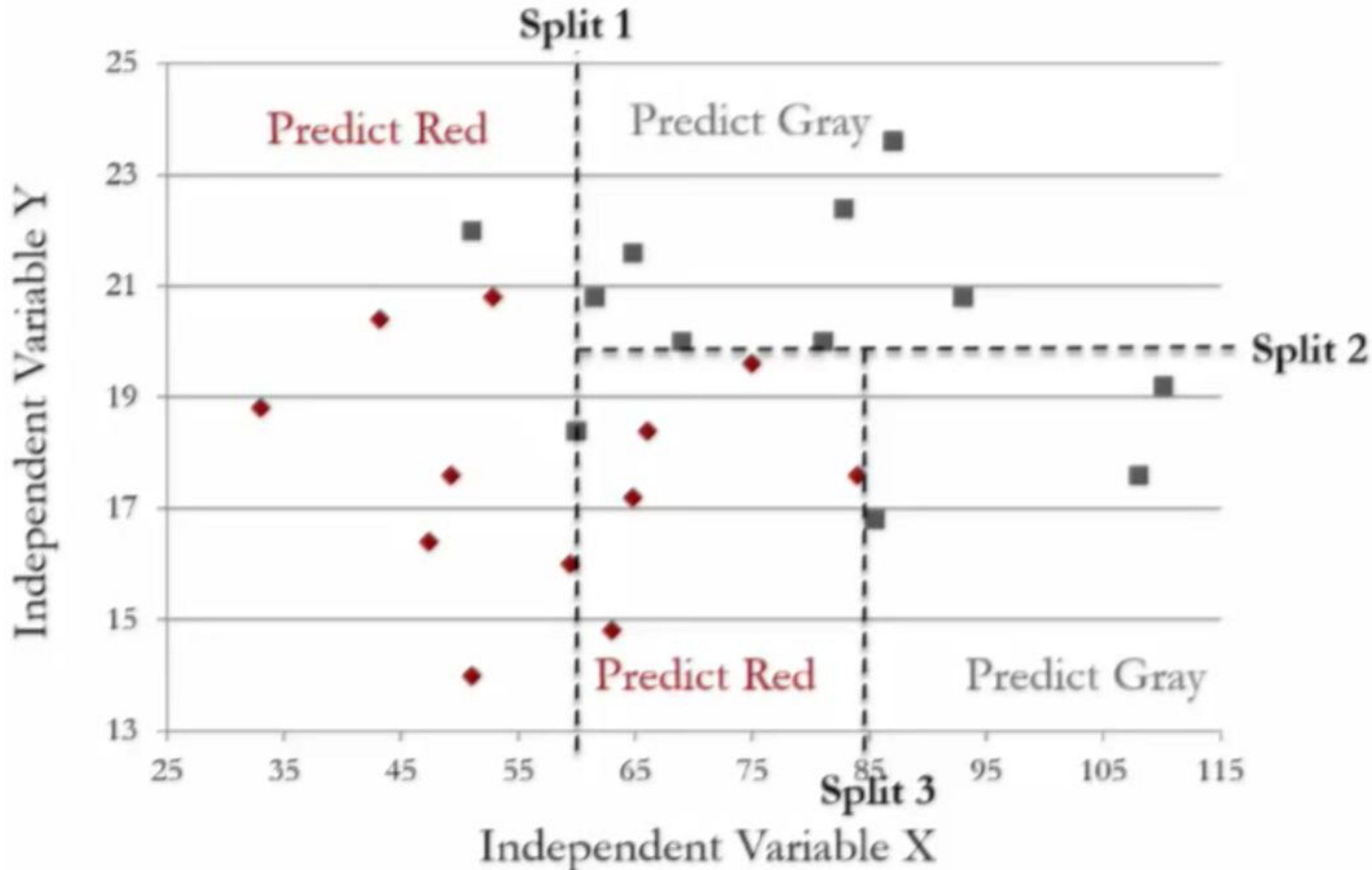
https://openclipart.org/image/2400px/svg_to_png/26849/johnny-automatic-scales-of-justice.png

Instead.... CART (Classification and Regression Trees)

- Build a tree by splitting variables
- To predict the outcome for an observation, follow splits and at the end pick the **most frequent** outcome
- *Does not assume linear model!*



Splitting the data



◆ = red

■ = gray

How CART works

- In each branch (boxes in graph), we have a number of outcomes
    - affirm or reverse in our court data

- Compute percentage of data in each branch
    - Example: 10 affirm, 2 reverse -> 10/(10+2)= 87% affirm

- Like logistic regression, we use **threshold value** to make prediction
    - This example 0.5 threshold would pick most frequent outcome
    - Vary our threshold value to compute ROC curve

Demo Time!

http://www.rako.com/Diamond/images/Classic_shopping_cart.jpg

Coming Up

Your problem set: Kaggle Titanic Dataset

Next week: More advanced classification models

See you then!

CDS

Cornell Data Science