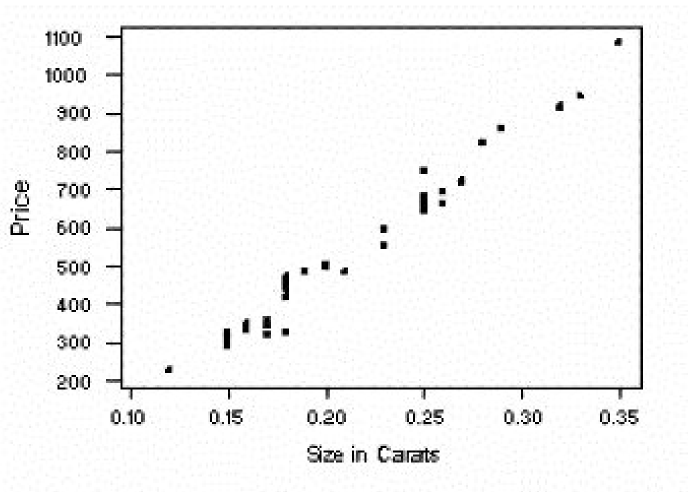
Logistic Regression and Decision Trees

Reminder: Regression

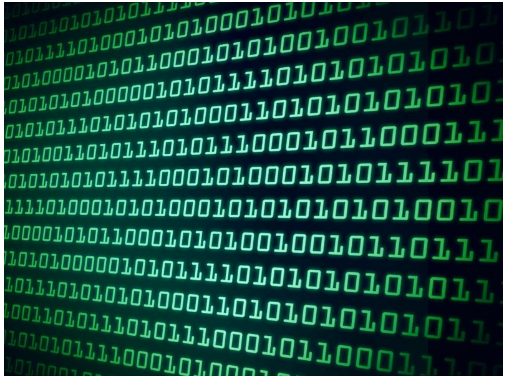
We want to find a **hypothesis** that explains the behavior of a **continuous** *y*



$$y = B_0 + B_1 x_1 + \dots + B_p x_p + \varepsilon$$

Source

Regression for binary outcomes



Regression can be used to **classify**:

- Likelihood of heart disease
- Accept/reject applicants to Cornell Data Science based on affinity to memes

Estimate **likelihood** using regression, convert to binary results

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

Logistic Regression

1) Fits a linear relationship between the variables
2) Transforms the linear relationship to an estimate function of the **probability** that the outcome is 1.

Basic formula:

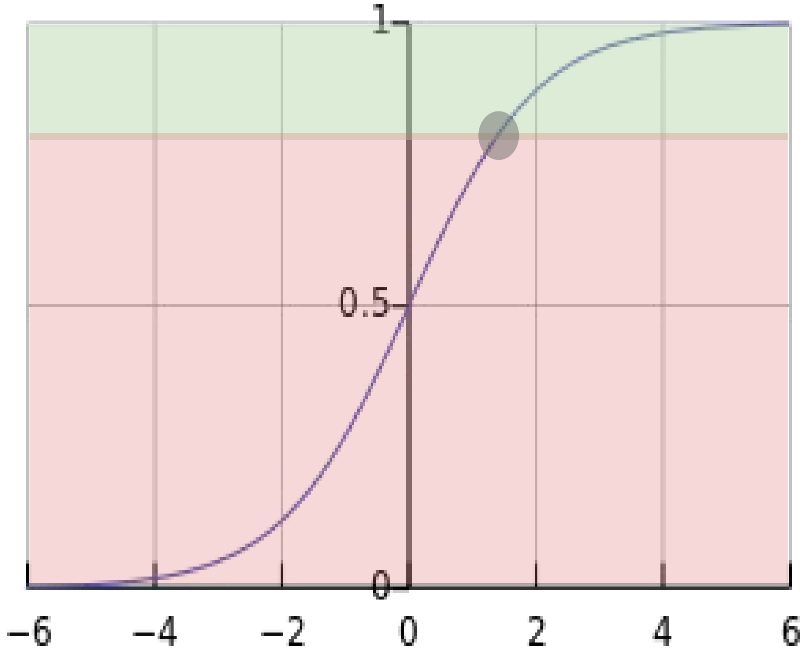
$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}$$

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

(Recognize this?)

Logistic Function

- Between 0 and 1
- X-axis spans (-inf, inf)

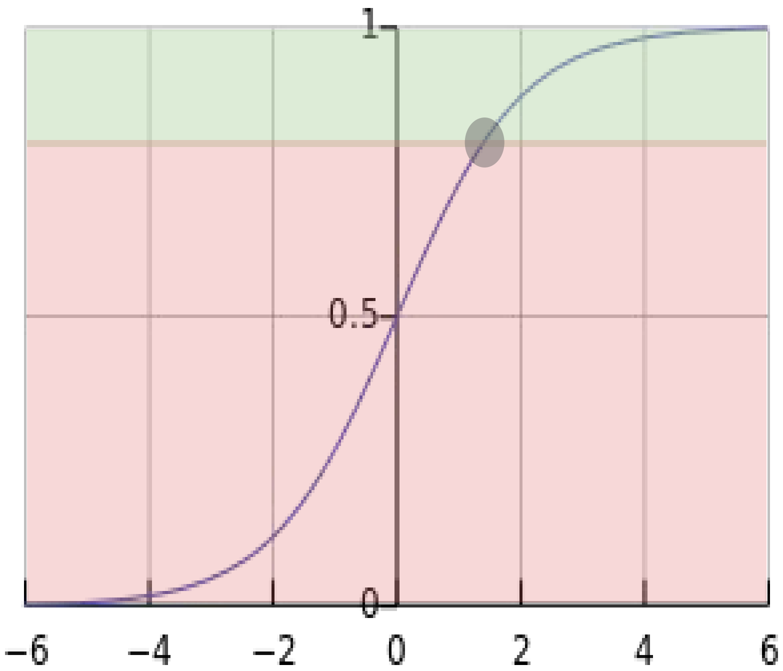


https://upload.wikimedia.org/wikipedia/commons/thumb/8/88/Logistic-curve.svg/320px-Logistic-curve.svg.png

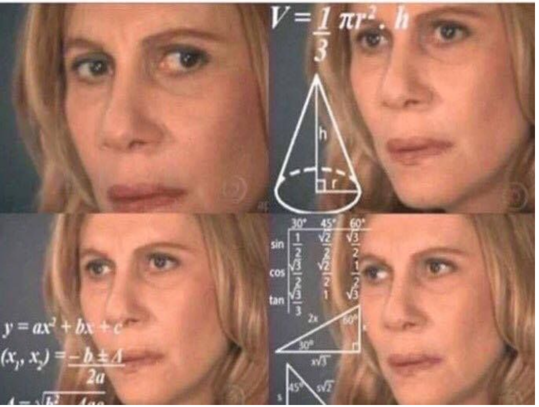
Threshold

Where between 0 and 1 do we draw the line?

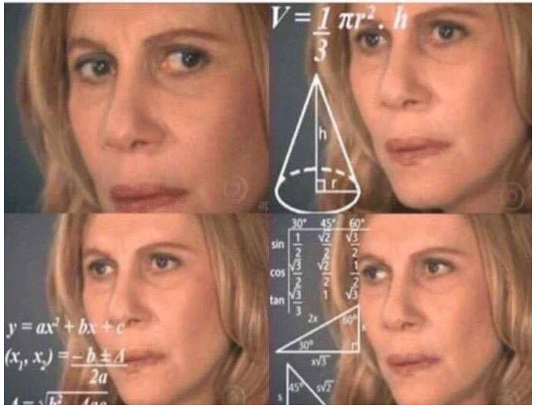
- $P(x)$ below threshold: predict 0
- $P(x)$ above threshold: predict 1



Confusion Matrix



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



Sensitivity

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

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 **True Negative Rate**.

How many negatives are correctly identified as negative?

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:

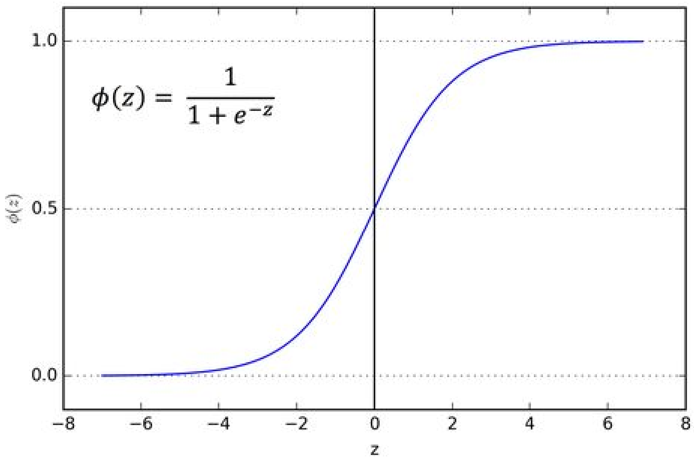
- Specificity?
- Sensitivity?
- Overall error rate?
- Overall accuracy?

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

Thresholds matter (a lot!)

What happens to the specificity when you have a

- Low threshold?
- High threshold?

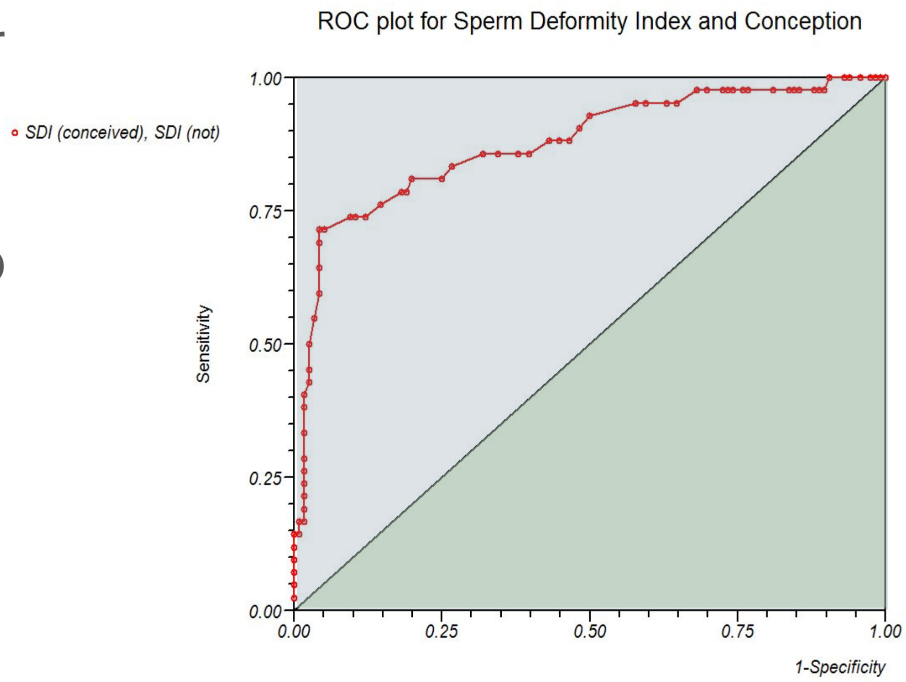


https://sebastianraschka.com/images/faq/logisticregr-neuralnet/sigmoid.png

ROC Curve

Receiver **O**perating **C**haracter

- Visualization of trade-off
- Each point corresponds to specific threshold value



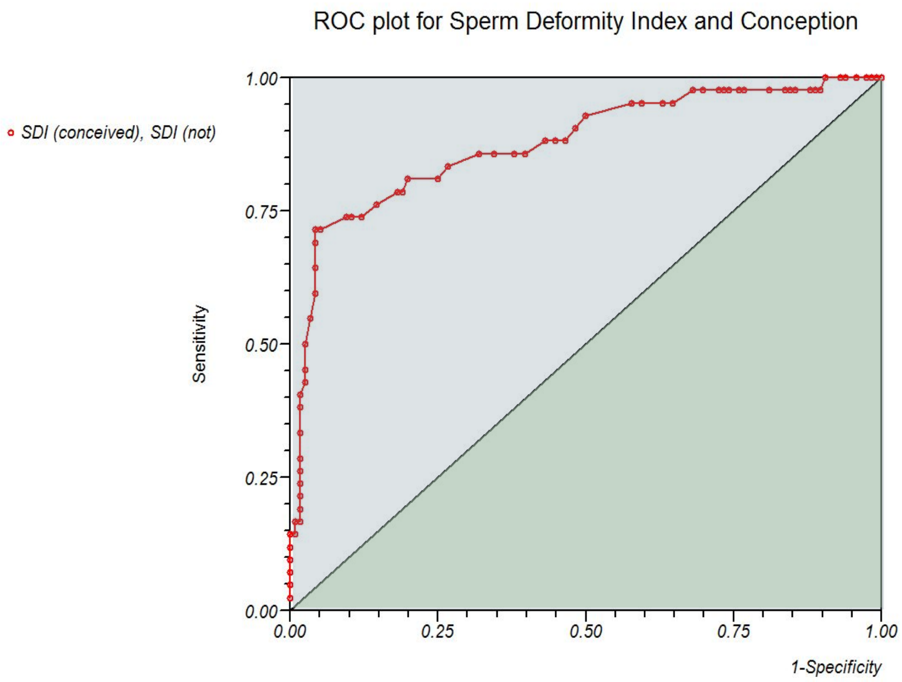
Area Under Curve

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

Always between 0.5 and 1.

Interpretation:

- 0.5: Worst possible model
- 1: Perfect model



Implementing in R

1) Construct a model using the glm() function
   a) Specify family “binomial”
2) Use pROC package to:
   a) Create an ROC object using roc()
   b) Plot using plot() and pick a certain threshold value
   c) Obtain AUC using auc()
3) Predict using predict() function and use

Case Study:

Predicting Supreme Court Decisions

And Justice For All

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

- **Dependent variable**: Did the supreme court overturn the lower court’s decision?
- **Independent variables:** various properties of the case

Article on FiveThirtyEight: http://53eig.ht/2mnLUI3

Source

Let's Try Regression

Logistic Regression used on the courts using different predictor variables!

- If case is from 2nd circuit court: +1.66
- If case is from 4th circuit court: +2.82
- If lower court decisions were liberal: -1.22
- ...

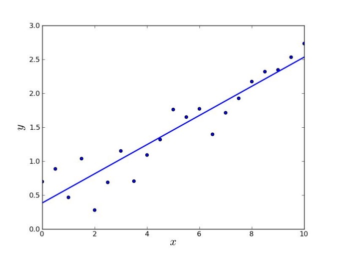
Is this intuitive?

Source

Review: Supervised Learning

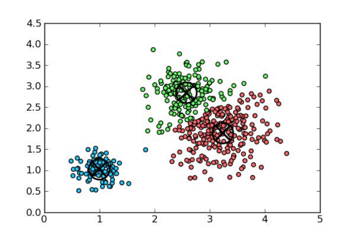
Regression

"How much?"
Used for *continuous* predictions



Classification

"What kind?"
Used for *discrete* predictions



https://lh3.ggpht.com/-d16xbB_2m94/Uml0Gi76XtI/AAAAAAAAI94/5ezC-ZqZWz8/s1600/linearRegression.png

http://blog.mpacula.com/wp-content/uploads/2011/04/kmeans1.png

Pitfall of Regression

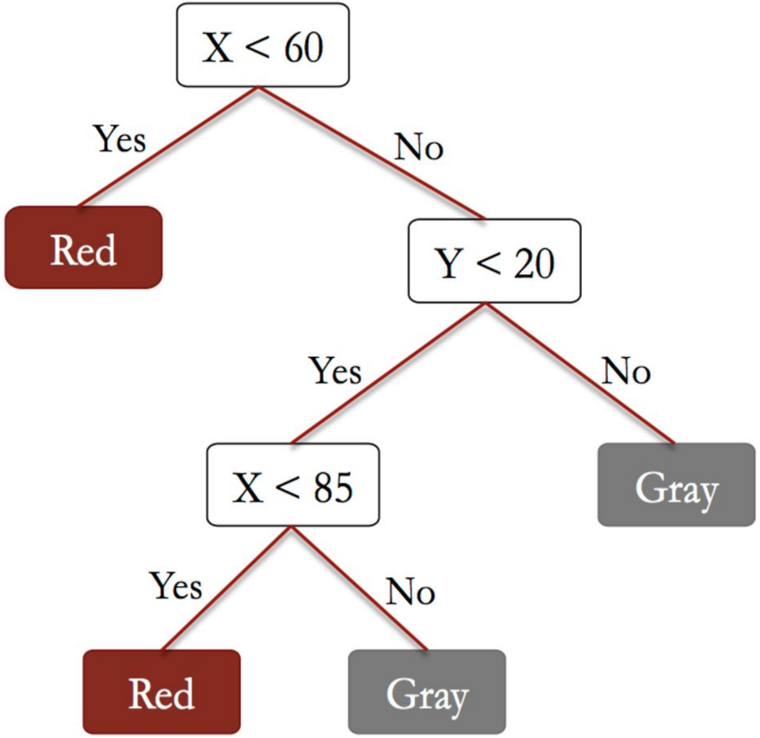
Problem: Regression models are **additive** and assume linearity, which won't help us much here.

This approach can lack **interpretability**, which is an important part of any useful model.

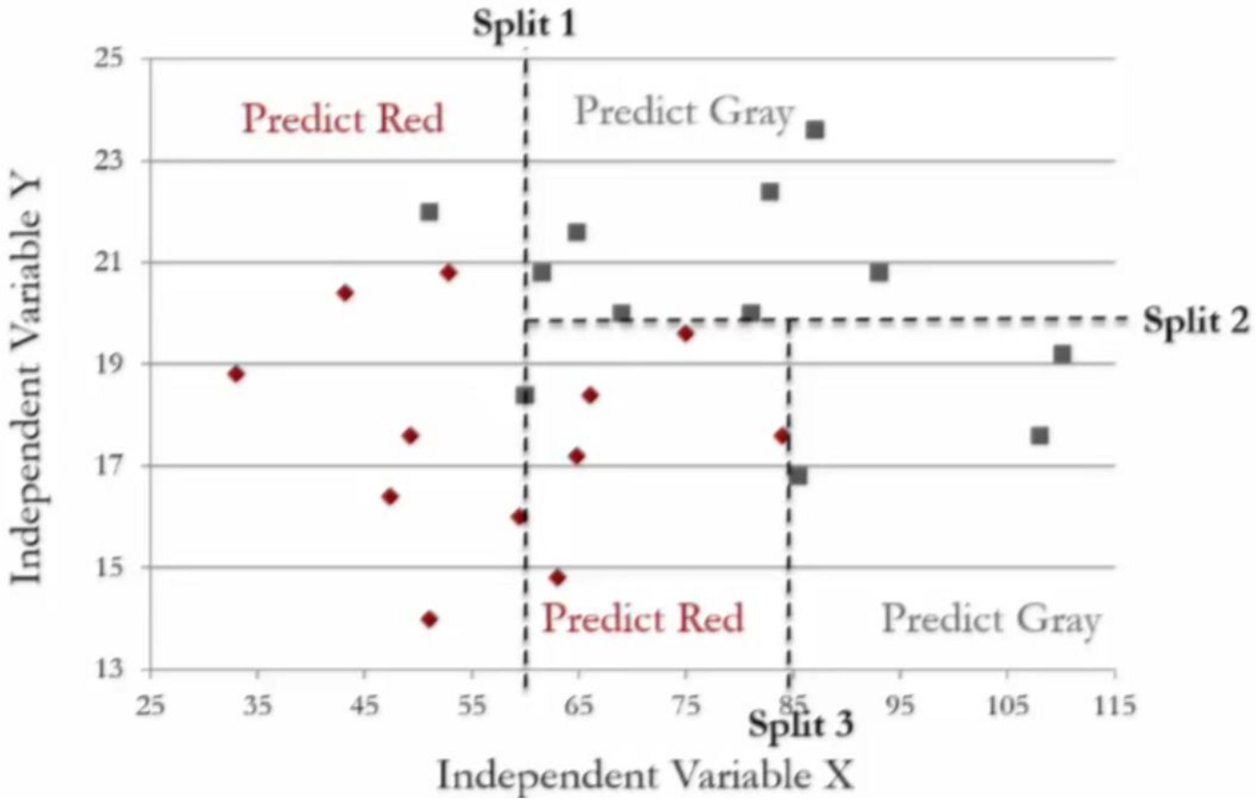
Source

Instead: CART (Classification and Regression Trees)

- At each node, split on variables
- Each split maximizes reduction of sum of squares for regression trees
- Very interpretable
- Models a non-linear relationship!



Splitting the data



◆ = red

■ = gray

How to grow trees(class. tres)

Gini Impurity

- 1 minus probability that random guess *i* (probability p_i) is correct
- Lower is better

$$1 - \sum p_i^2$$

Entropy

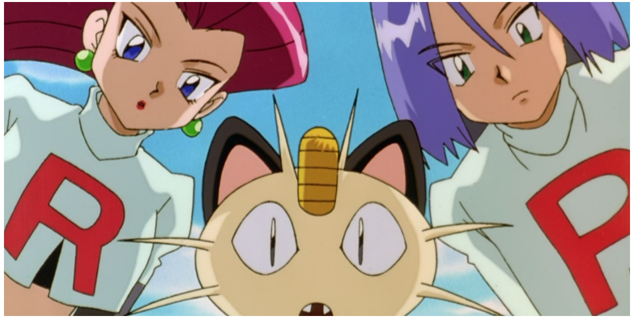
- Information theory concept
- Measures mixed-ness, unpredictability of population
- Lower is better

$$-\sum p_i \log p_i$$

https://image.slidesharecdn.com/untitled-150504104846-conversion-gate01/95/a-taste-of-random-decision-forests-on-apache-spark-21-638.jpg?cb=1430736669

Implementing in R

1) Construct a model using the rpart() function
    a) Specify classification (“class”) or regression (“anova”)
    b) Specify tree depth, splits, complexity using control
2) Plot or print the complexity of model using
3) Prune the tree
4) Predict using predict() function
5) Pick a method/threshold to convert predictions to a final classification



Demo Time!

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

Coming Up

Your problem set: Project 1 released!

Next week: More advanced classification models

See you then!