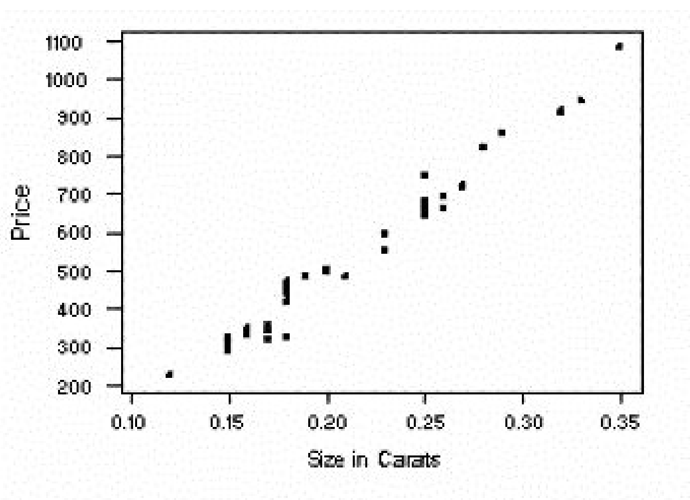
CDS

Cornell Data Science

Logistic Regression and Decision Trees

Reminder: Regression

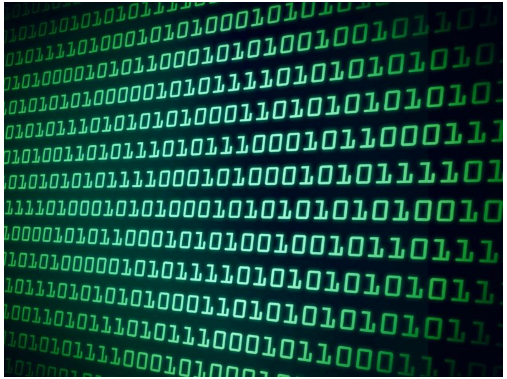
We want to find a **hypothesis** that explains the behavior of a **continuous** *y*.



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

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

Source

Conditional Probability

The probability that an event (A) will occur given that some condition (B) is true

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

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:

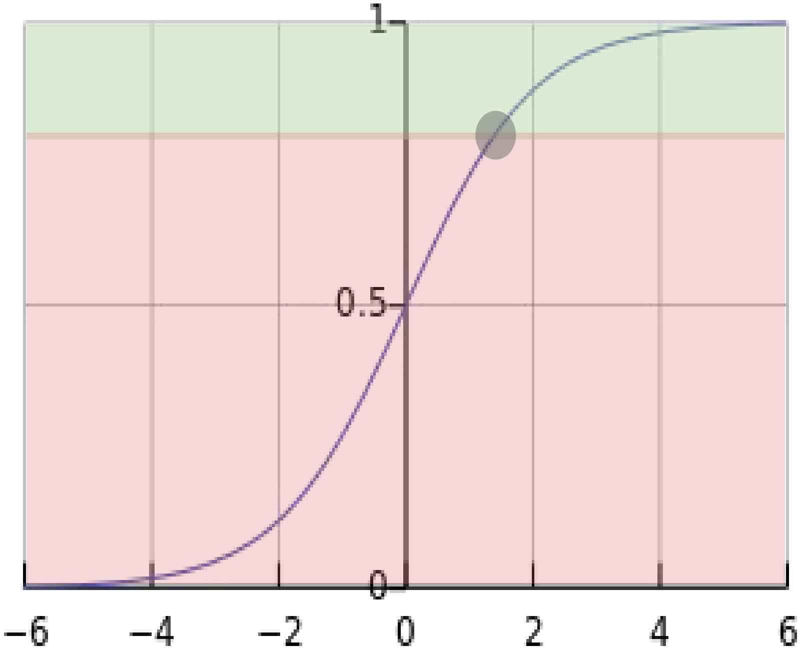
$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}$$

(Recognize this?)

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

Logistic Function

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

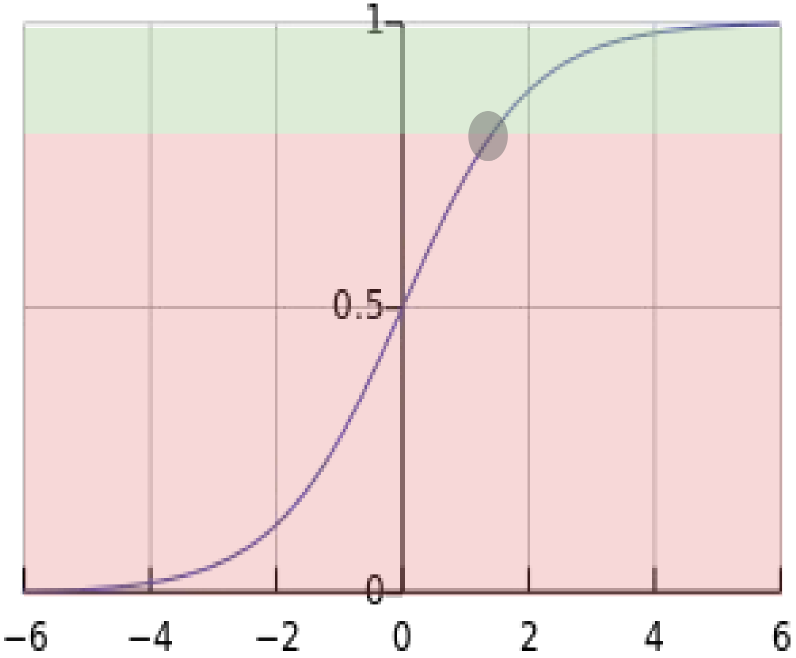


Source

Threshold

Where between 0 and 1 do we draw the line?

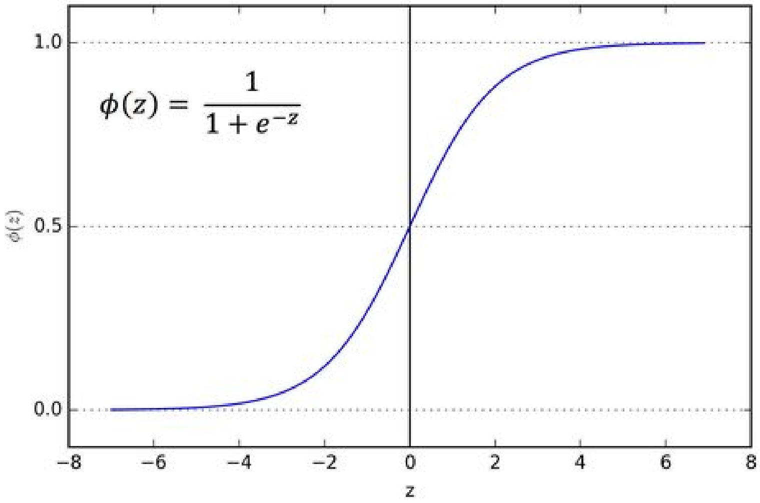
- *P(x)* below threshold: predict 0
- *P(x)* above threshold: predict 1



Thresholds matter (a lot!)

What happens to the specificity when you have a

- Low threshold?
- High threshold?

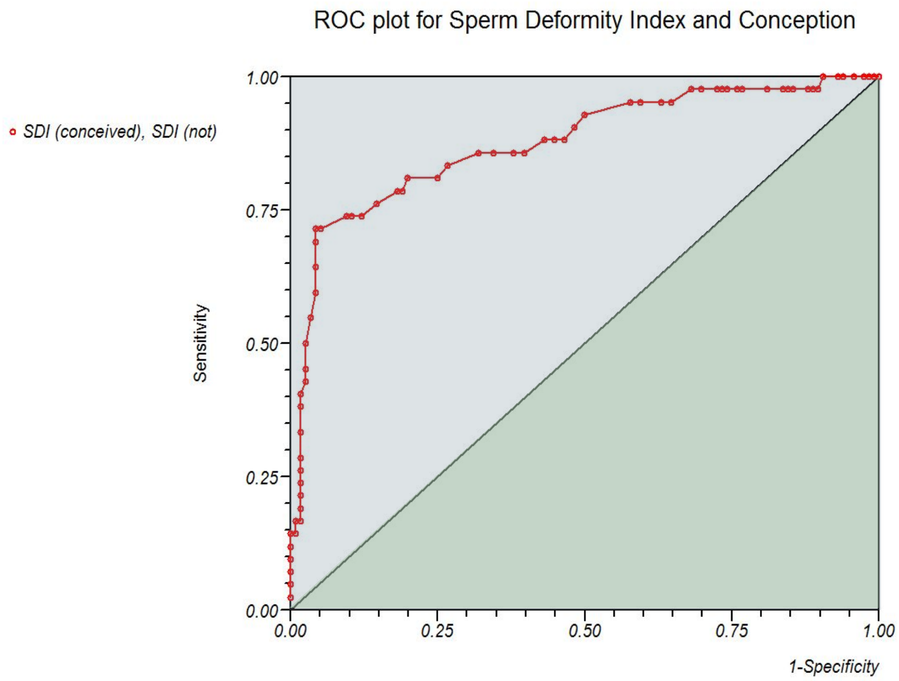


Source

ROC Curve

Receiver **O**perating **C**haracteristic

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



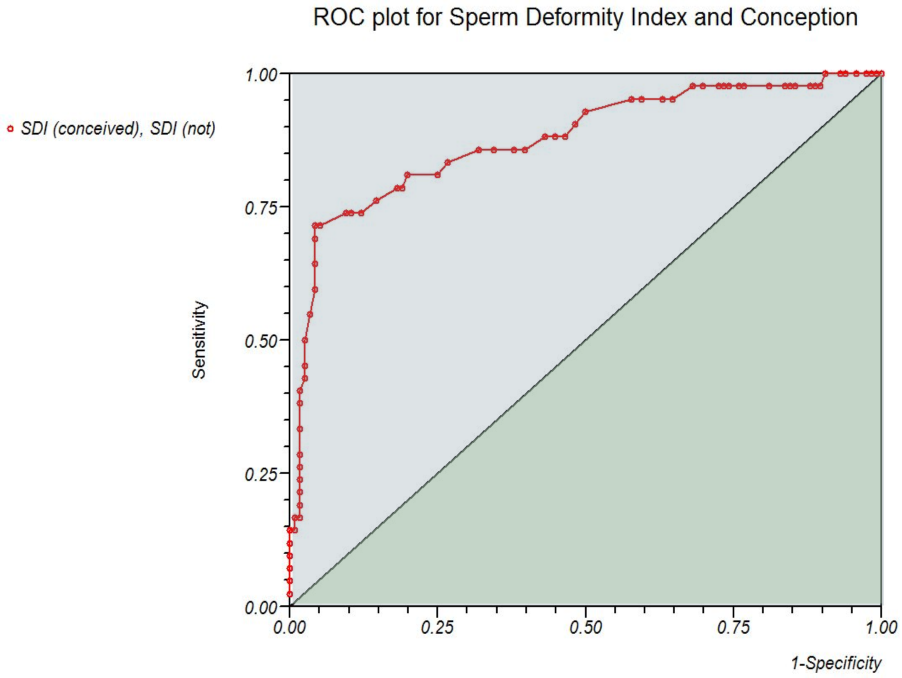
Area Under Curve

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

Always between 0.5 and 1.

Interpretation:

- 0.5: Worst possible model
- 1: Perfect model



When to Use Regression

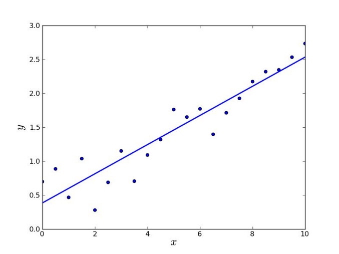
- Works well on (roughly) linearly separable problems
- Outputs probabilities for outcomes
- Can lack **interpretability**, which is an important part of any useful model

Demo!

Review: Supervised Learning

Regression

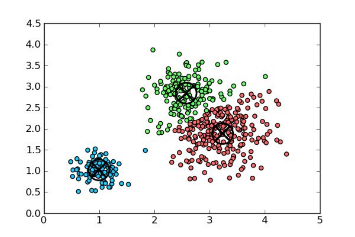
“How much?”
Used for *continuous* predictions



Source

Classification

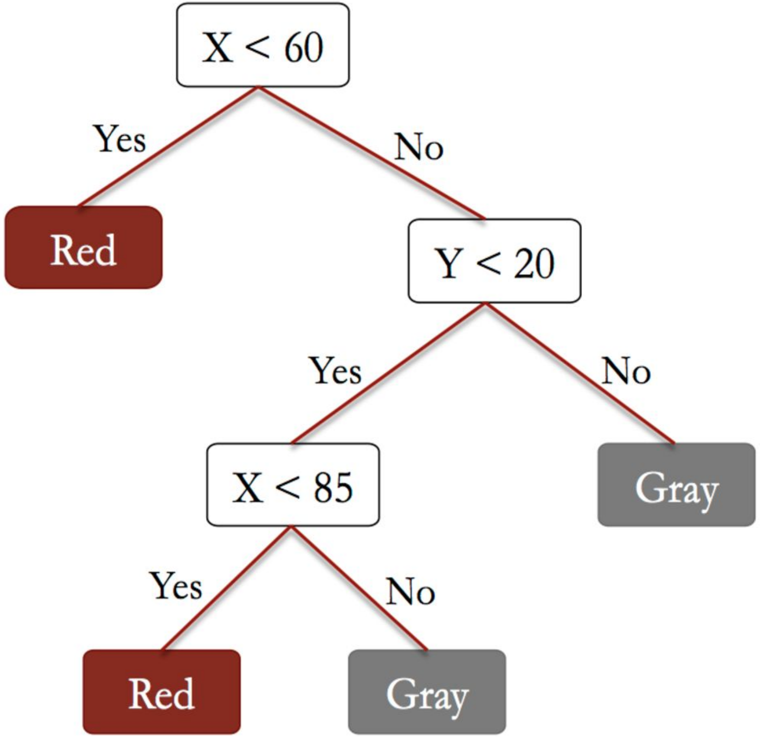
“What kind?”
Used for *discrete* predictions



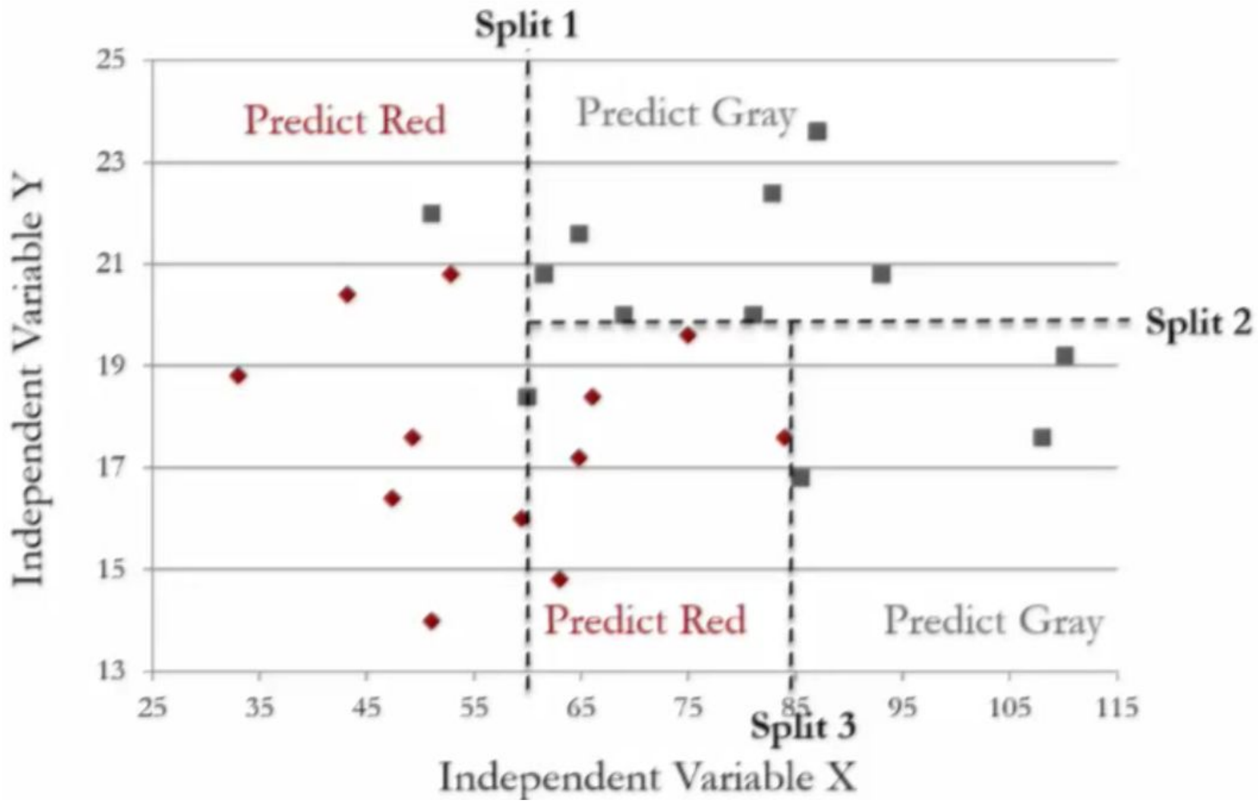
Source

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!



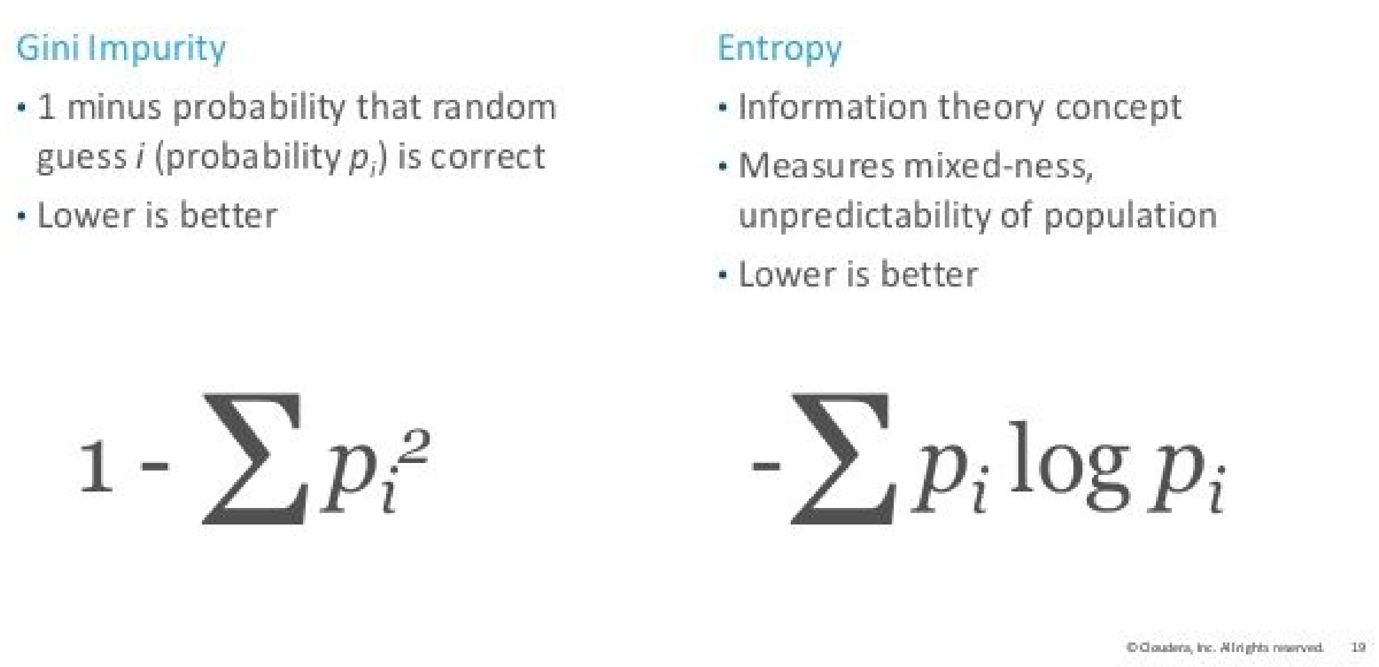
Splitting the data



◆ = red

■ = gray

How to grow trees (class.trees)



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

$$-\sum p_i \log p_i$$

Source

How to grow?

- Start at the top of the tree
- Split attributes one by one
  - Split based on impurity or entropy
- Assign the values to the leaf nodes
- Repeat
- Prune for overfitting

Decision Tree Parameters

- Used to control specificity of the tree
  - max_depth
  - max_leaf_nodes
  - min_samples_split
    - minimum number of cases needed for a branch

When to Use Decision Trees

- Easy to interpret
- Prone to overfitting

Demo!

Coming Up

Your problem set: Continue working on Project Part B

Next week: Unsupervised Learning

See you then!