Lecture: Predictive Modeling

Predictive Modeling Framework

- In many situations, our goal is to describe a scientific process or phenemenon. To do so, we fit a model, which we hope is a reasonable approximation to reality that allows inferences. In this situation, simplicity or parsimony is important.

- In other situations, the goal is to predict (or forecast) future values. Generally, in these situations a prediction (point estimate or distribution) are generated for upcoming scenarios, but the underlying model parameters are not of interest.

- In other situations, the goal may seem to be prediction, but there is still a desire to describe the underlying scientific process. Resolving this requires clear communication.

Loss Functions

- With the prediction setting, often times the problem will come with a specific criteria that we seek to minimize.

- For instance when predicting continuous quantities, squared error loss $(actual - pred)^2$ or absolute error loss $|actual - pred|$ are common.

- With categorical, or binary data, a zero-one loss is common where there is zero loss for a correct prediction and a loss of one for an incorrect prediction.

- In both situations, you can also have non-symmetric loss functions where one kind of prediction error is more "costly".

- The loss functions that we previously discussed are focused on point predictions, but they can also evaluate predictive distributions or prediction intervals. One example is the Kaggle March Madness prediction competition, where rather than a prediction for each basketball game, contestants produce a winning probability.

  Then the log loss function is used: $y \times log(p) + (1 - y) \times log(1 - p)$, where y is the binary response.

- Conditional Bayes Factors are another way to evaluate predictive distributions.

Test / Training and Cross - Validation

- An important concept in predictive modeling is overfitting. Which enough parameters in a model, it is possible to get perfect predictions. However, that model will typically do a lousy job on another data set.

- Two common approaches for evaluating predictive models are cross-validation and the test/training set.

- Cross-validation partitions your data into k groups. Then all of the observations in the first $k - 1$ groups are used to fit a model and predictions are made for the k^{th} group. This process continues until all of the observations have been predicted once (and used in the model fitting $k - 1$ times).

- With the test / training set, typically around 70 percent of the data is used for training and predictions are made on the test set.

- Be careful of situations that require additional care: time series data and other forms of highly structured data.

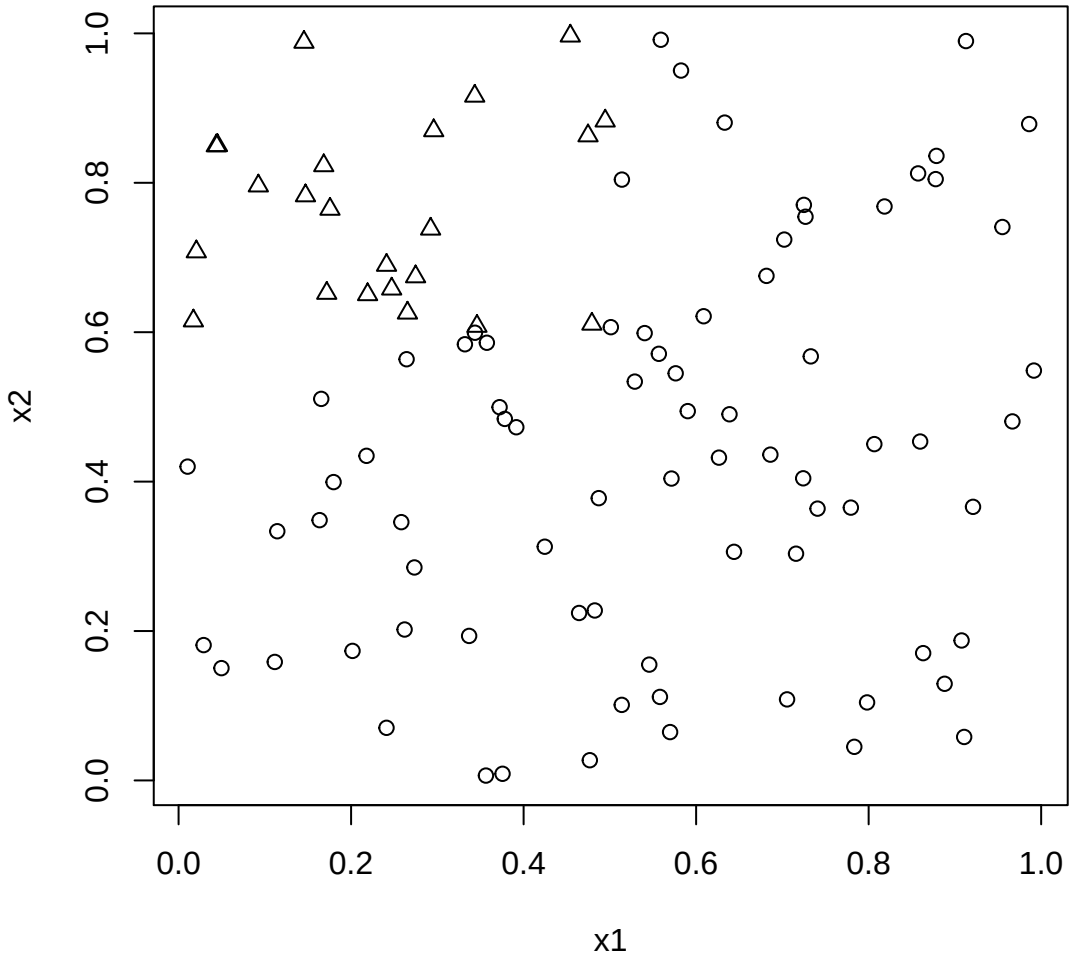
Tree Methods

- One of my favorite statistical papers is "Statistical Modeling: The Two Cultures" by Leo Breiman https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726, which focuses on "algorithmic models" that do not have an underlying probabilistic or data generating mechanism.

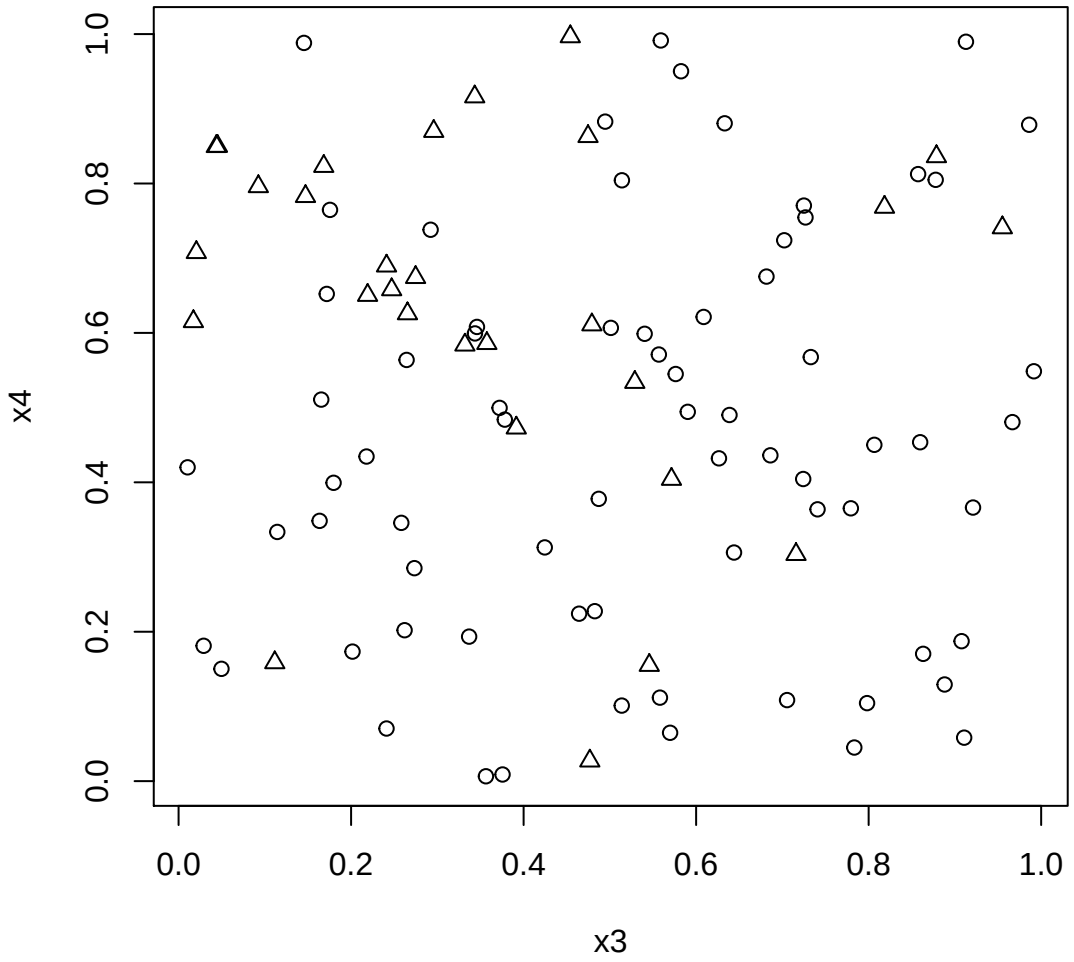
- One such example of an algorithmic model is a decision tree - also known as Classification And Regression Tree (CART).

- Classification and Regression Trees partition the predictor space into distinct sets that have different models or predictions.

- Consider the figure below and write out a statistical model for predicting triangles or circles.

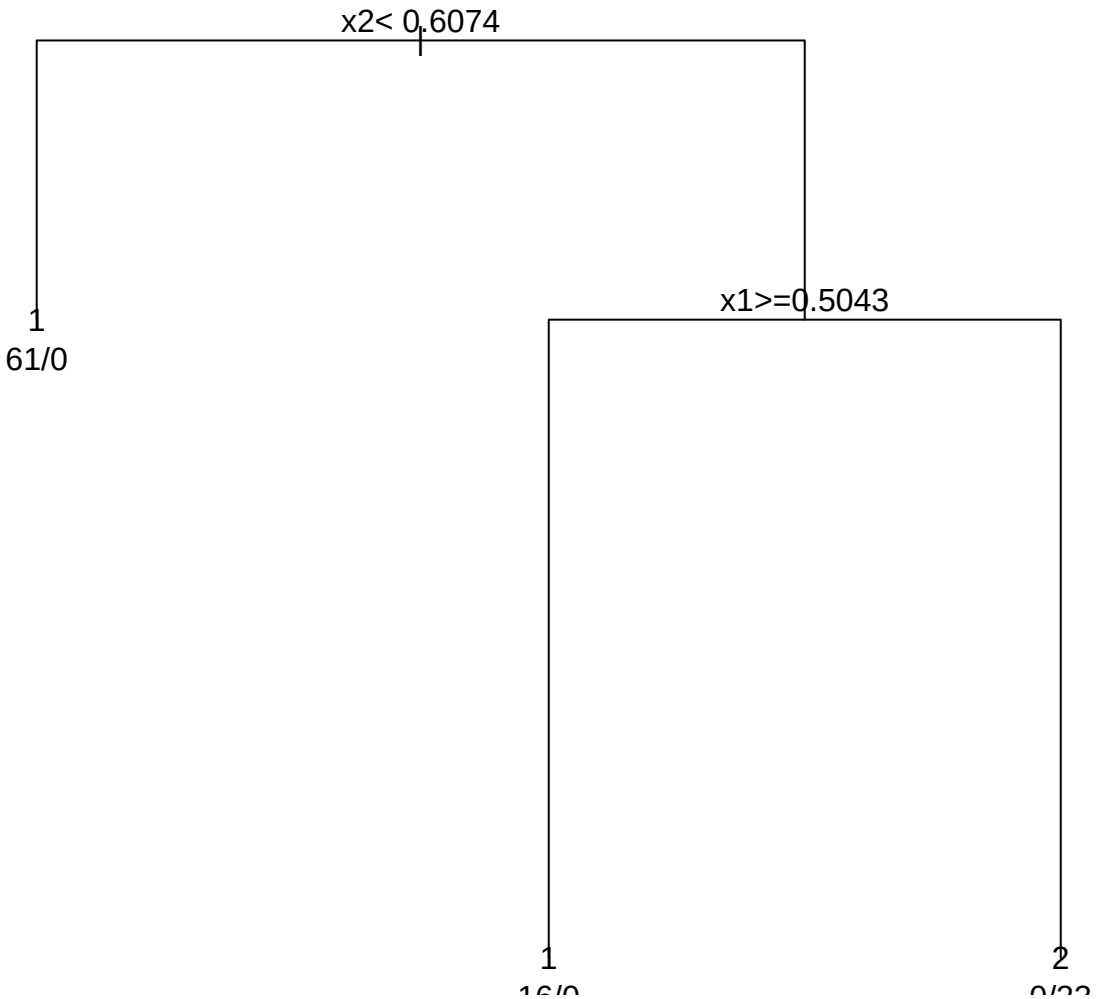


- Now, how about the figure below?

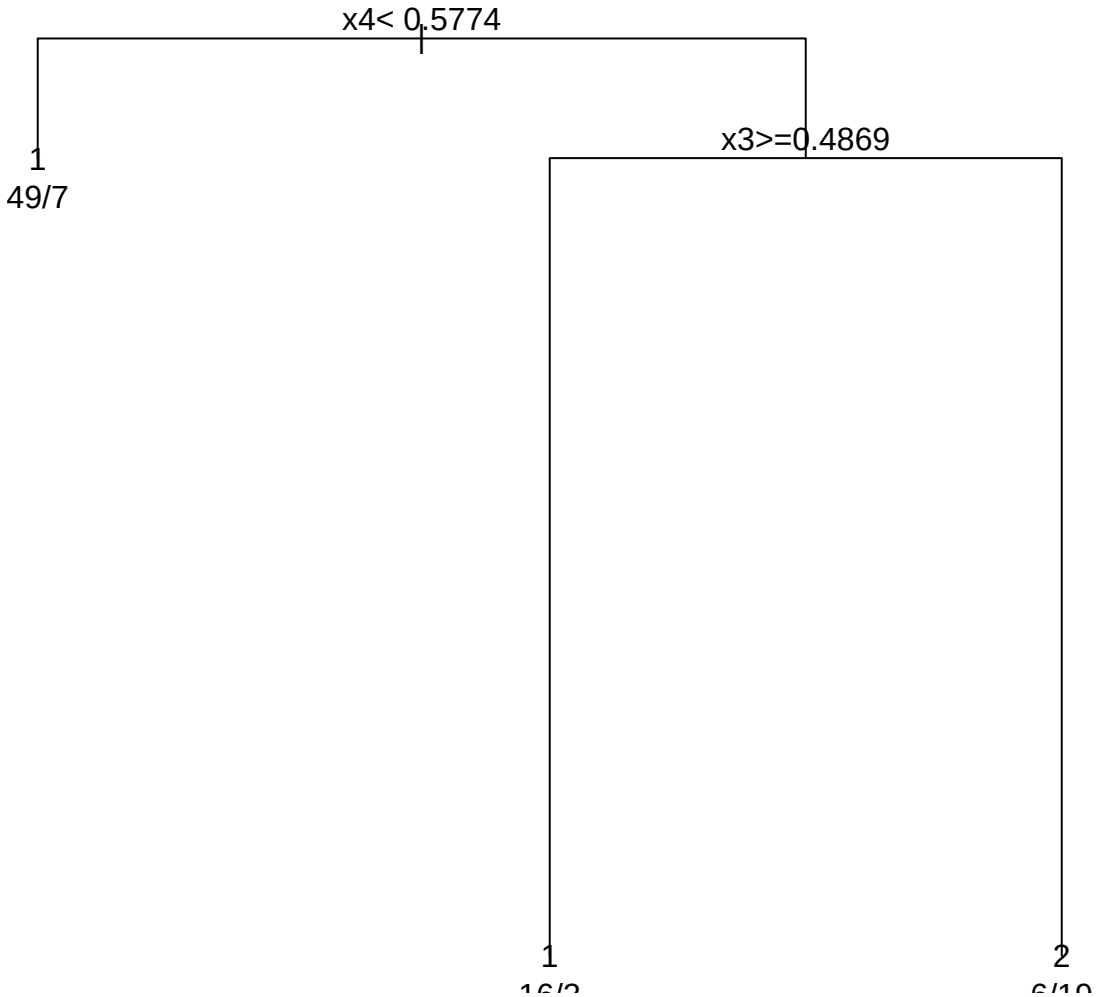


- Using R, we can fit a decision tree with `rpart`.

```
library(rpart)
tree.df <- data.frame(y1 = point.type1, x1 = x1, x2 = x2,
                      y2 = point.type2, x3 = x3, x4 = x4)
tree1 <- rpart(y1 ~ x1 + x2, data = tree.df, method = 'class')
plot(tree1, compress = T)
text(tree1, use.n = TRUE)
```




```
tree2 <- rpart(y2 ~ x3 + x4, data = tree.df, method = 'class')
plot(tree2, compress = T)
text(tree2, use.n = TRUE)
```



- The `rpart` objects have the usual generic r functions: predict and residuals.

- These examples were both classification trees, where a value of one or zero is predicted for each partition.

- Regression trees are also possible, where each terminal node would have an indpendent tree.

Ensemble Methods

- Ensemble methods combine several *weak* learners to reach a consensus decision.

- Consider a set of independent models that each correctly predict a binary outcome 60 percent of the time.

  - Then $Pr[Y = 1|M_1 = 1] = .6$

  - $Pr[Y = 1|M_1 = 1, M_2 = 1] = \frac{.6^2}{.6^2+.4^2} = 0.6923077$

  - $Pr[Y = 1|M_1 = 1, M_2 = 1, \ldots, M_5 = 1] = 0.8836364$

- A common example of an ensemble method is a random forest. A random forest is composed of a set of decision trees; hence, the forest moniker.

- The random forest attempts to create "pseudo-indepent" trees by randomly selecting a subset of the covariates to include and a random samples, with replacement, or the observations.

- In the classification setting, a majority rule algorithm is implemented for a final, consensus choice.

- The performance of a random forest is based on two criteria: the strength of the predictions for each individual tree and the correlation between those trees.

Bayesian Trees

- The seminal paper for Bayesian Trees is Bayesian CART Model Search by Chipman, George, and McCulloch.

- Finding and evaluating trees is difficult. Most procedures, including random forest, include a greedy search algorithm.

- In a Bayesian setting, the tree search is controlled by the prior distribution and a stochastic search algorithm.

- Priors are placed on a stochastic generating mechanism for the tree. In particular, there are two fundamental concepts that control tree generation: split probability and rule application.

- The splitting probability determines whether or not to split a given node.

- The rule application determines which features (covariates) to split on and how to implement that split.

- Given these prior distributions, the prior probability of a tree can be algorithmically computed.

- A stochastic search algorithm using Metropolis-Hastings is implemented with the following proposal functions: GROW, PRUNE, CHANGE, and SWAP.

- Posterior predictive distributions use the ensemble of trees for predictions.