Introduction to Data Science

Data Science Essentials

NASHVILLE SOFTWARE SCHOOL

Goals for today

- Review last session coding tasks
- Machine Learning, Part 2 – Tree-Based Models

Review last session coding tasks

week5_review notebook

Parametric Methods

Assume that your model takes a particular functional form. Fitting the model boils down to finding the best set of parameters.

Common examples: **linear regression**, **logistic regression, neural networks**

$$\text{mpg} = \beta_0 + \beta_1 \cdot (\text{hp}) + \beta_2 \cdot (\text{cyl})$$

Parameters

Nonparametric Methods

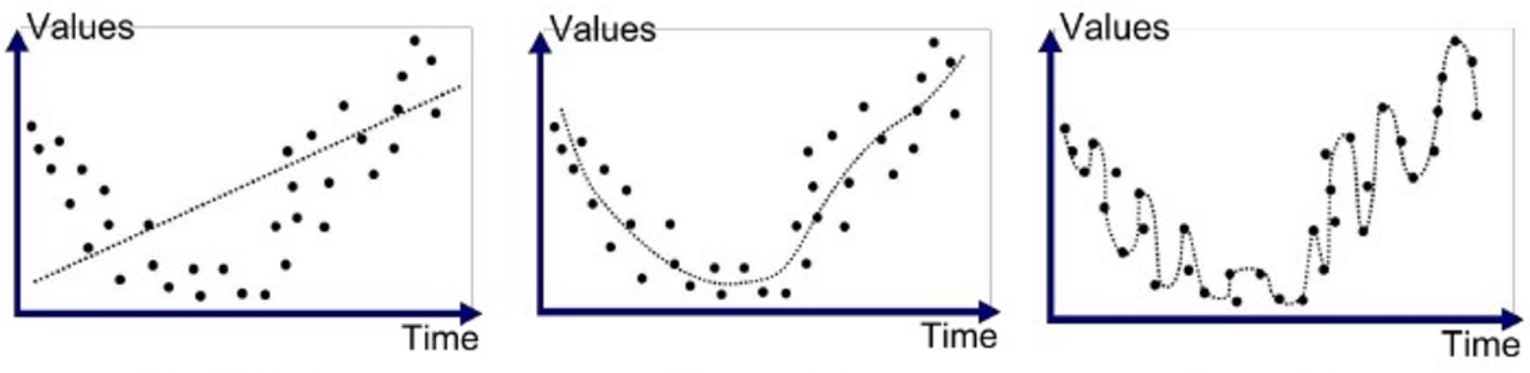
Do not assume a particular functional form for your model. Instead, use an estimate that gets as close as possible to the training data.

Common examples: **decision trees, random forests, gradient boosted trees, k-nearest neighbors**

Parametric vs. Nonparametric

Parametric methods will be low accuracy if the particular form that you chose is incorrect. Prone to **underfitting** - not fitting the training or the test data very well.

Nonparametric methods can fit the training data extremely well but not generalize well to new data. Prone to **overfitting** - fitting the noise. Larger amount of training data is needed to obtain an accurate model.

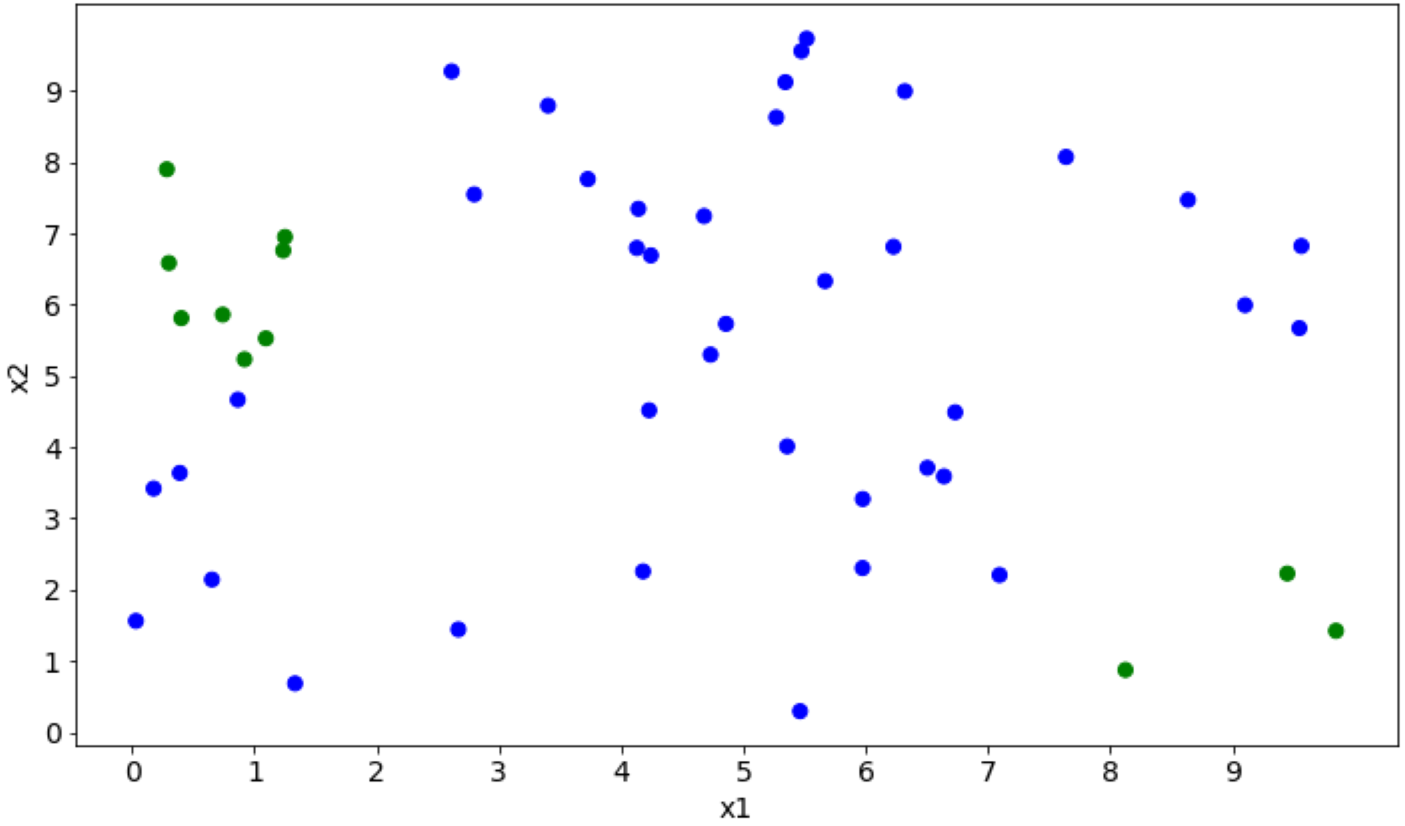


Underfitted   Good Fit/Robust   Overfitted

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Decision Trees

Makes predictions by using a set of sequential, hierarchical decisions leading to a final outcome. Can work on datasets that would be difficult to predict using parametric methods.



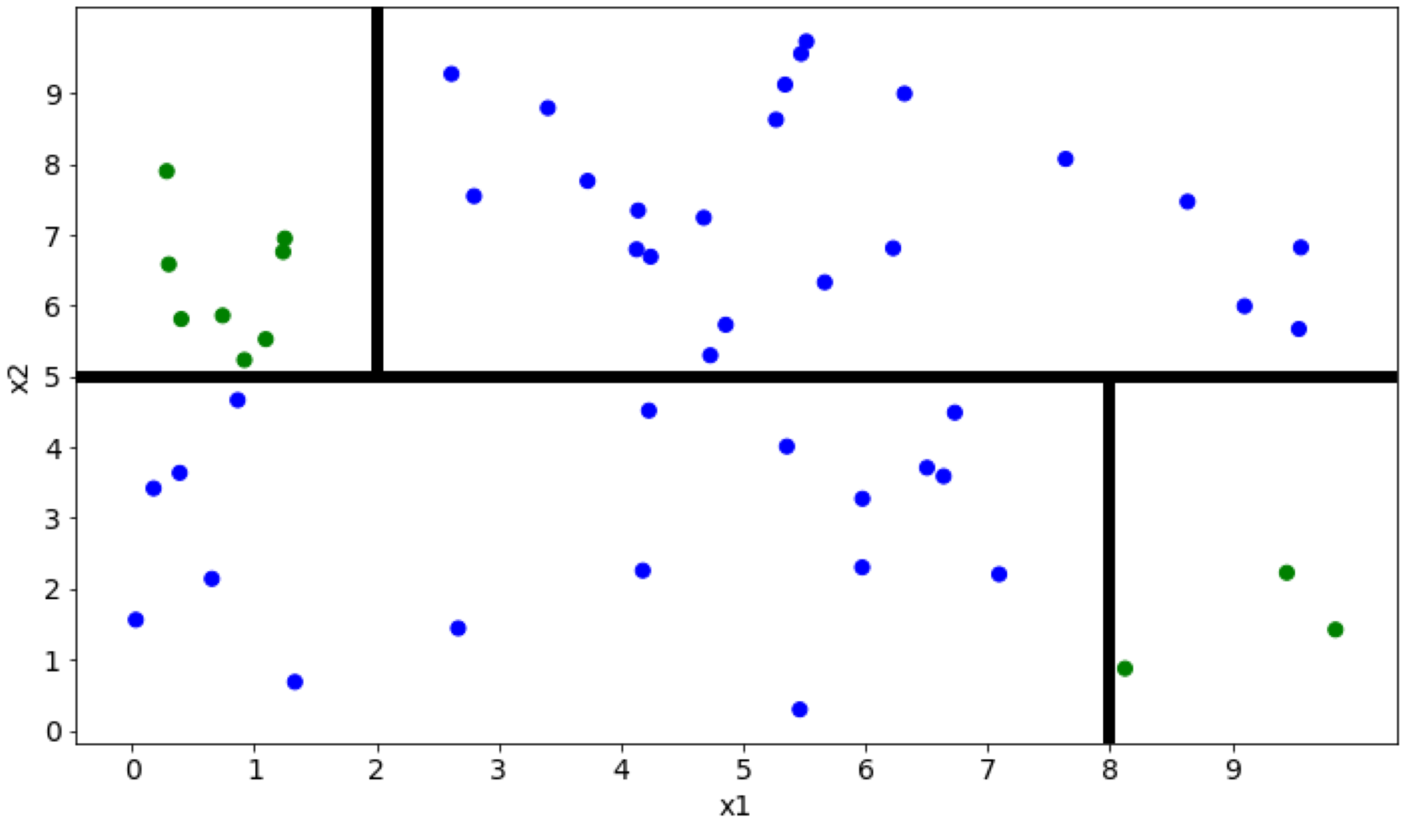
Can we build a model that distinguishes blue points from green points?

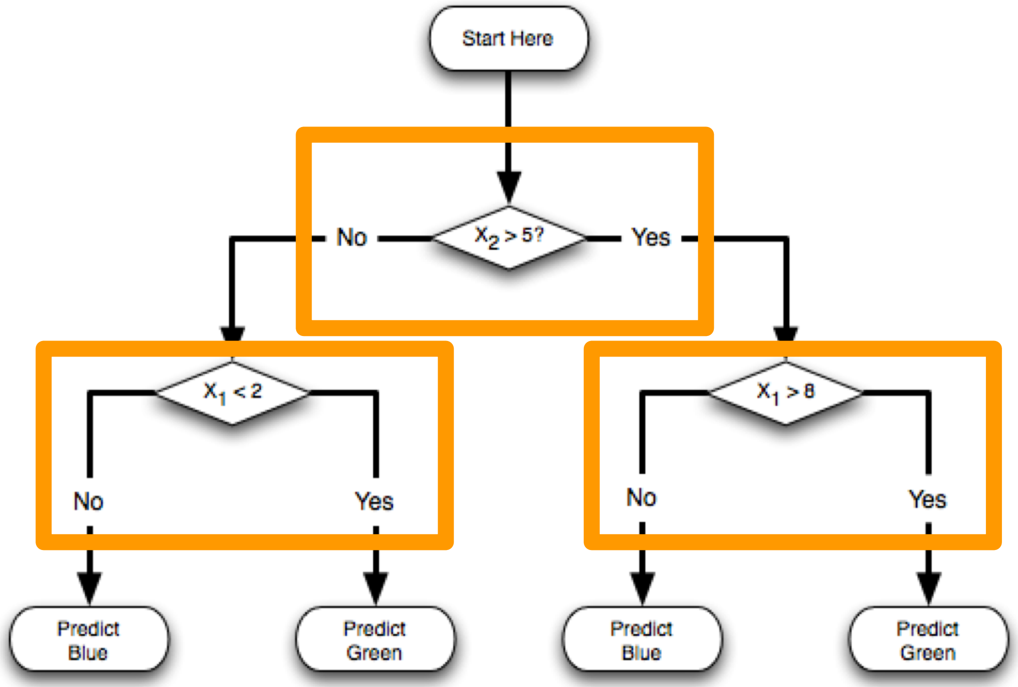
It is impossible to separate these points by a straight line, so we could not use a traditional logistic regression model here (unless we applied a complicated transformation to the dataset).

Instead, we can opt for a decision tree model.

Decision Trees

Makes predictions by using a set of sequential, hierarchical decisions leading to a final outcome. Can work on datasets that would be difficult to predict using parametric methods.





Random Forests

Single decision trees will often not generalize well to new data.

One common way to improve this is by building multiple trees and aggregating the predictions (aka creating an **ensemble**).

Random Forest models create a large number of decision trees, each trained on a subset of the training data and a subset of the features in order to decorrelate and reduce the variance of predictions.

To make the final prediction, the predictions from each tree are averaged.

NASHVILLE
SOFTWARE
SCHOOL

Model building and evaluation

Model_evaluation_Part2 notebook

Next Steps:

Your turn! Find more data to try and improve your model. Try using a random forest model to predict whether a county's cost-income ratio is above or below the mean for TN (hint: first create a label for the data that answers that).

- **October 16 - data storytelling and presentation; work in teams to create a 7-10 minute presentation of your findings**
- **October 23 – Team presentations + panel with data scientists who have walked in your shoes**

NASHVILLE SOFTWARE SCHOOL

Questions?