Introduction to Modeling (Statistical Learning)

Visualizations and material from ISLR

Outline

The Language of Modeling

Motivating Examples

Mathematical Formulation

The Language of Modeling

Modeling - The process of developing a mathematical/ statistical tool to describe the relationship between a set of variables

Variable Types:

- Response - The outcome event or quantity that is being predicted
- Predictors - The variables that are used to model/predict the response

Modeling Outcomes - Interpretation or Prediction/Classification

- Interpretation - Understanding the relationship between the predictors and the response
- Prediction - Predicting the value of the response based on the predictors

The Language of Modeling

Examples of Modeling Questions:

Determining whether or not an email is spam.

Determining what type of media will contribute most to sales

Determining how much extra will a house be worth if it has a view of the river

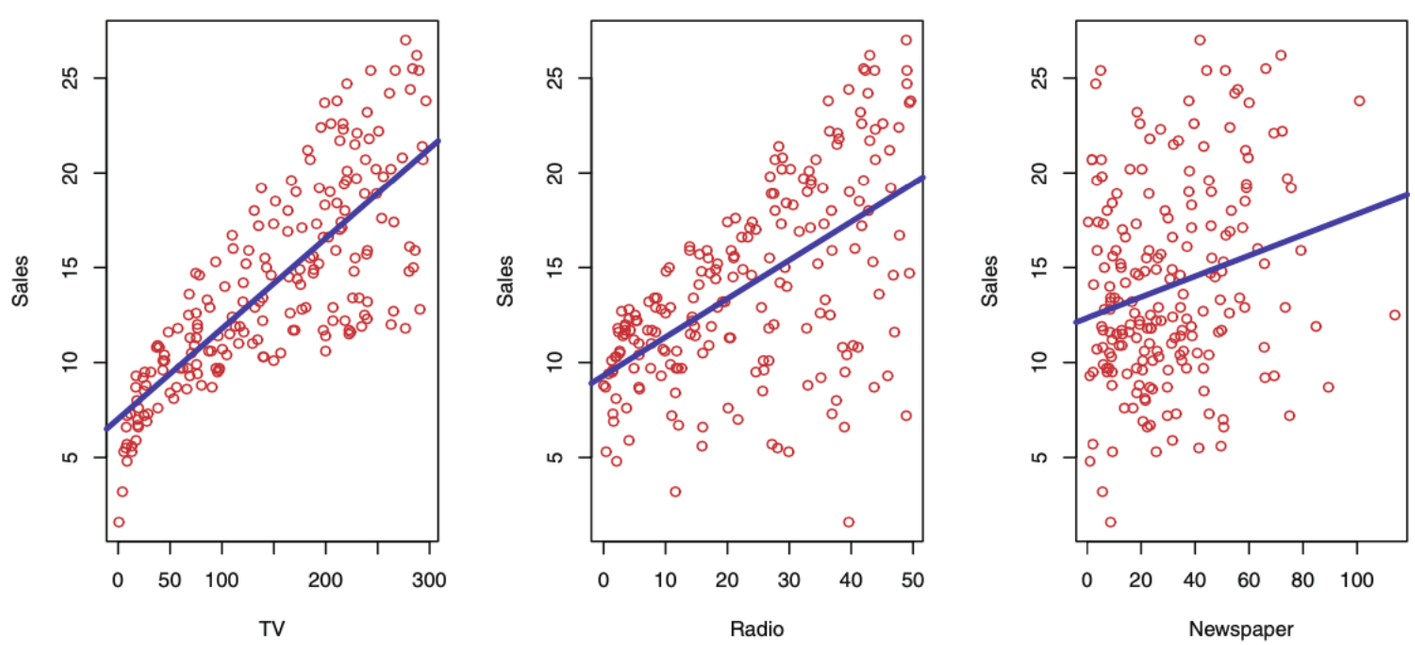
The Language of Modeling

Example: Advertising Data

Output: Sales $\longleftarrow Y$

Inputs: TV budget, radio budget, and newspaper budget

X_1 X_2 X_3



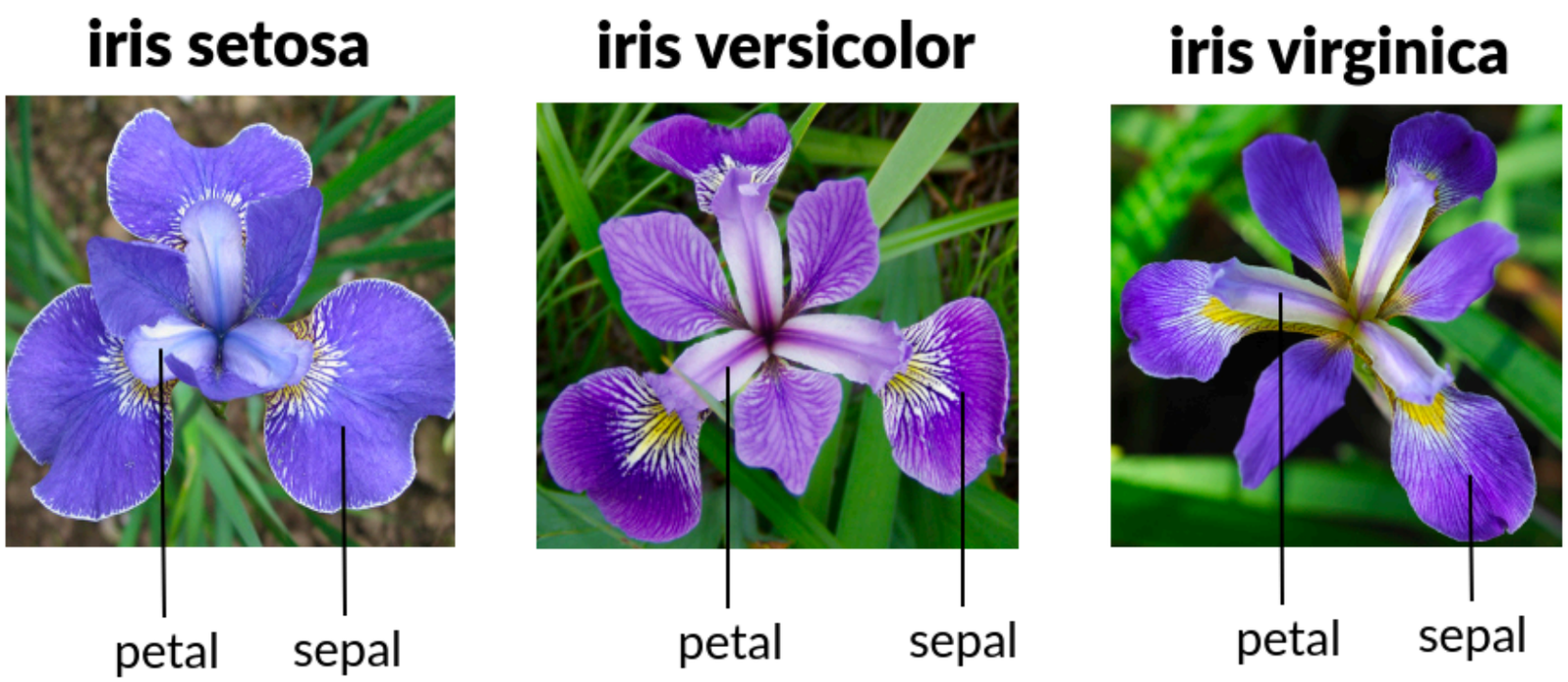
The Language of Modeling

Example: Fisher's Iris Data

Output: species

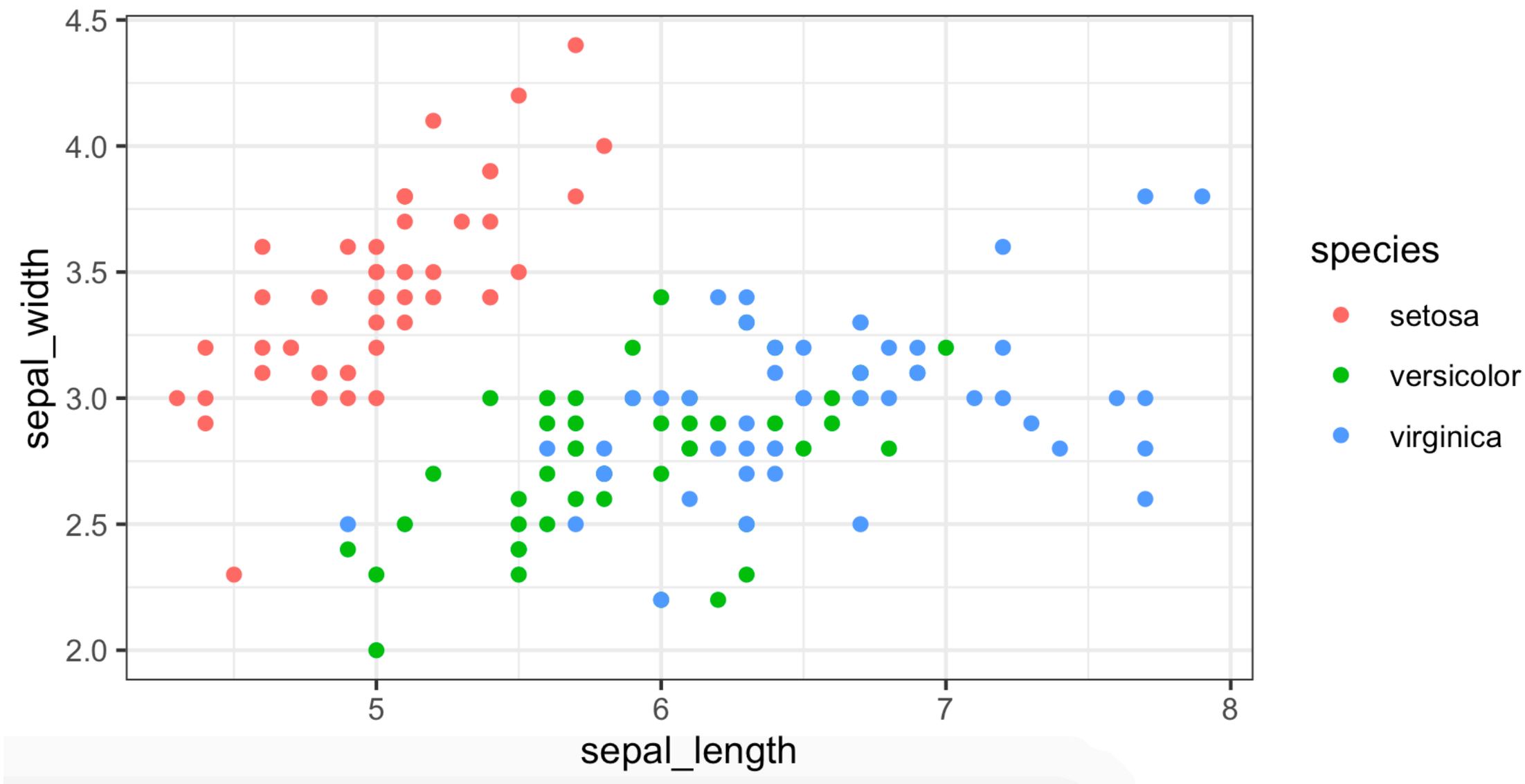
Inputs: petal and sepal lengths and widths

	sepal_length	sepal_width	petal_length	petal_width	species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa



Credit: https://thegoodpython.com/iris-dataset/

The Language of Modeling



The Language of Modeling

General Mathematical/Statistical Form

Quantitative Response: Y

Set of Predictors: $\boldsymbol{X} = (X_1, X_2, ..., X_p)$

$$Y = f(\boldsymbol{X}) + \epsilon$$

systematic terms

error term

Goal of statistical learning: estimate f!

The Language of Modeling

How do we estimate f?

Lots of approaches!!!

- linear vs. nonlinear
- parametric vs. nonparametric
- supervised vs. unsupervised
- statistics vs. machine learning
- regression vs. classification

General Approach

1. Pick a method to estimate f (and satisfy assumptions)
2. Let the method "learn" from a set of data (training)
3. Test the method out on another set of data (testing)
4. Asses current fit and iterate to try and find optimal fit.