Week 6: An Overview of Machine Learning and Scikit-Learn

Kevin Lee

Department of Statistics
Western Michigan University

Outline

1. Overview of Machine Learning

2. Overview of Scikit-Learn

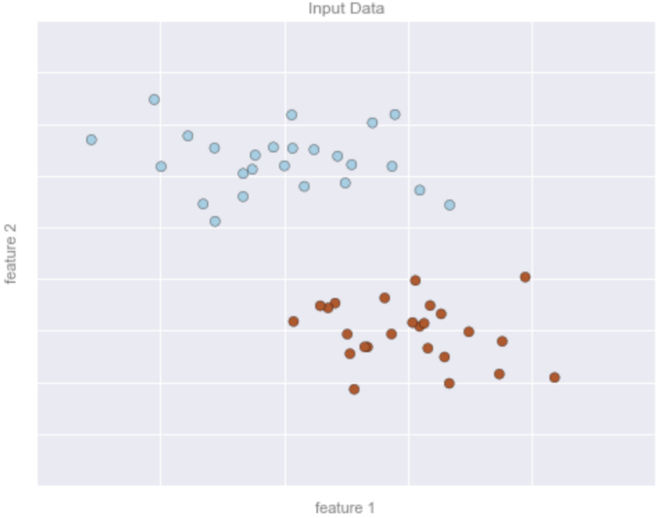
What is Machine Learning?

- Fundamentally, machine learning involves building mathematical models to help understand data.
- “Learning” enters the game when we give these models tunable parameters that can be adapted to observed data.
- Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data.

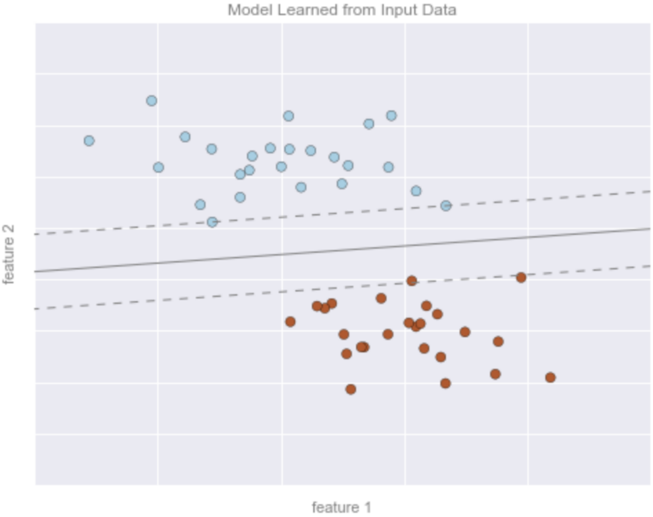
Categories of Machine Learning

- Supervised learning involves modeling the relationship between measured features of data and some label associated with the data; once the model is determined, it can be used to apply labels to new, unknown data.
  - Classification, the labels are discrete categories
  - Regression, the labels are continuous quantities.
- Unsupervised learning involves modeling the features of a data without reference to any label, and is often described as "letting the dataset speak for itself."
  - Clustering algorithms identify distinct groups of data.
  - Dimensionality reduction algorithms search for more succinct representations of the data.

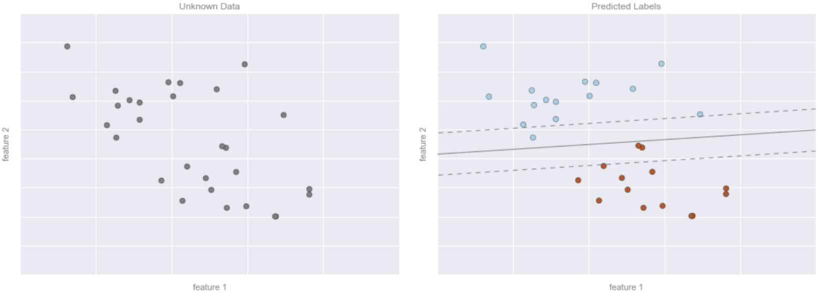
Classification: Predicting Discrete Labels



Classification: Predicting Discrete Labels



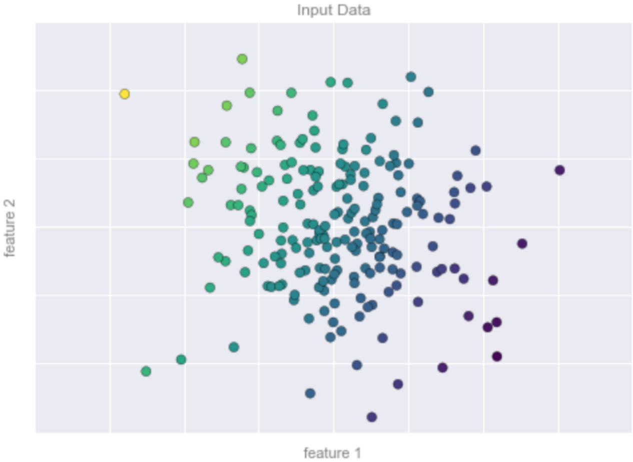
Classification: Predicting Discrete Labels



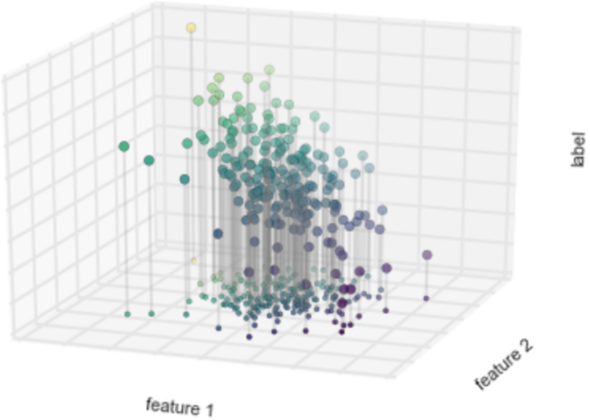
Unknown Data

Predicted Labels

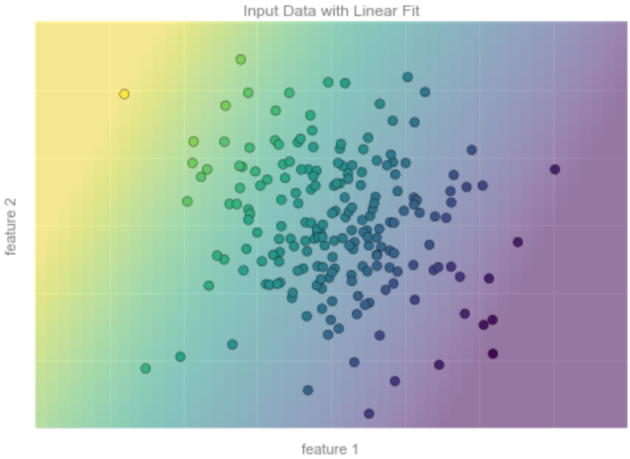
Regression: Predicting Continuous Labels



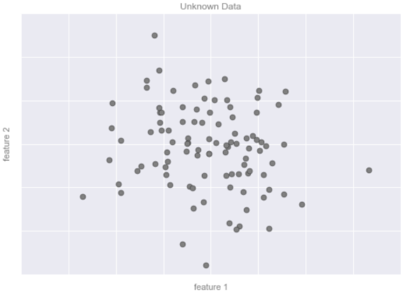
Regression: Predicting Continuous Labels

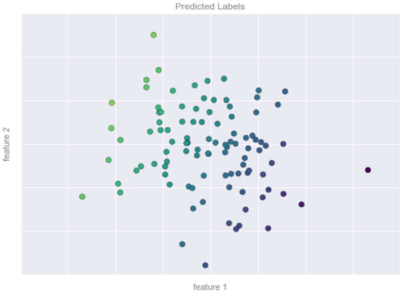


Regression: Predicting Continuous Labels

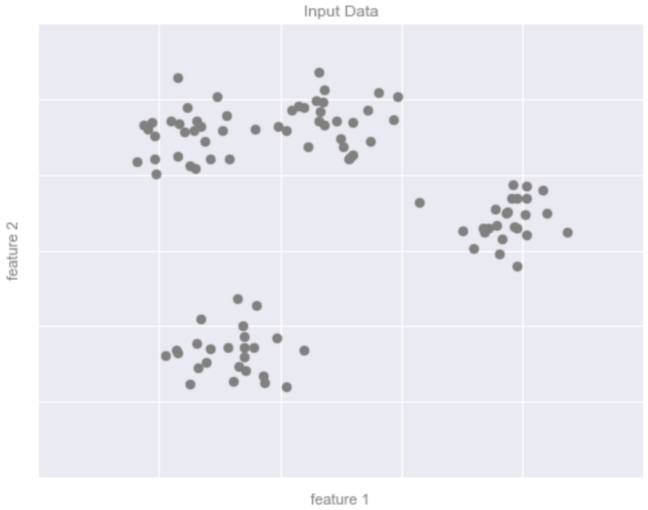


Regression: Predicting Continuous Labels

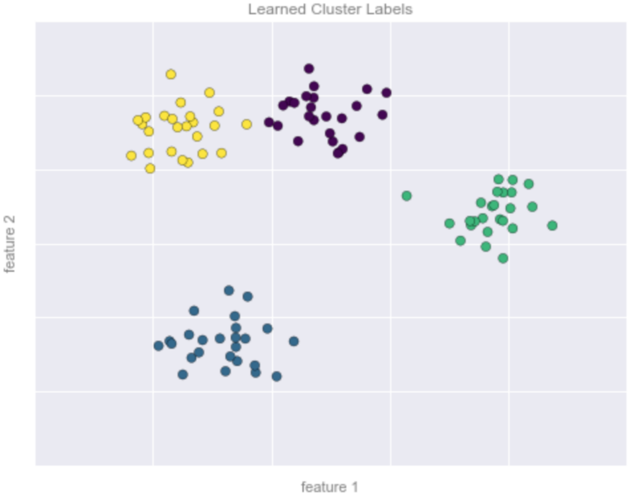




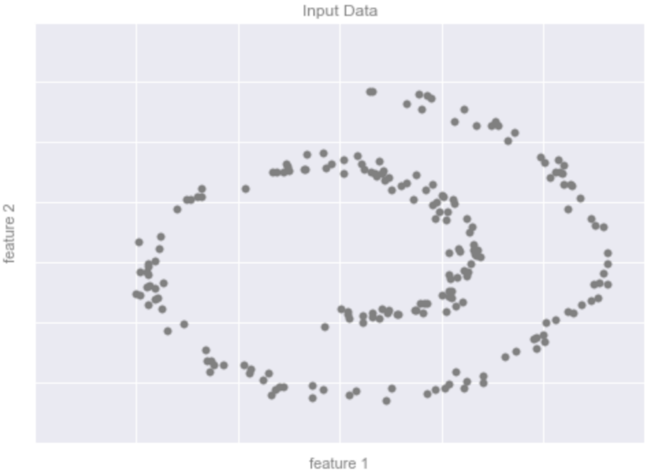
Clustering: Inferring Labels on Unlabeled Data



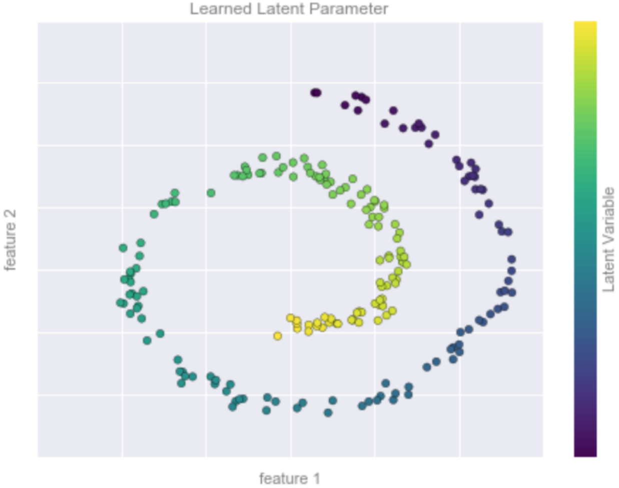
Clustering: Inferring Labels on Unlabeled Data



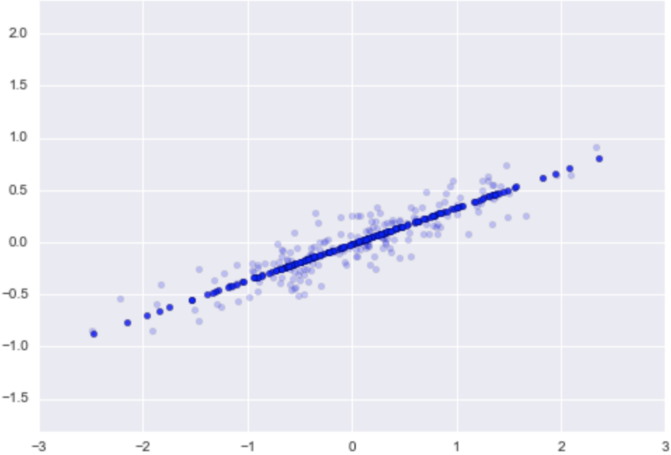
Dimensionality Reduction: Finding Low-Dimensional Representation of Data



Dimensionality Reduction: Finding Low-Dimensional Representation of Data



Dimensionality Reduction: Finding Low-Dimensional Representation of Data



Outline

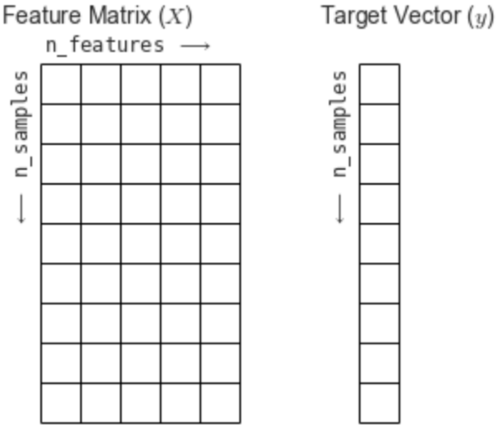
1. Overview of Machine Learning

2. Overview of Scikit-Learn

Introducing Scikit-Learn

- This project was started in 2007 as a Google Summer of Code project by David Cournapeau.
- In 2010 Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort and Vincent Michel of INRIA took leadership of the project and made the first public release in 2010.
- Scikit-Learn provides a large number of common machine learning algorithms.
- Scikit-Learn is characterized by a clean, uniform, and streamlined API, as well as by very useful online documentation.

Data Representation in Scikit-Leran


Feature Matrix (X)
n_features ⟶
n_samples ⟶
Target Vector (y)
n_samples ⟶