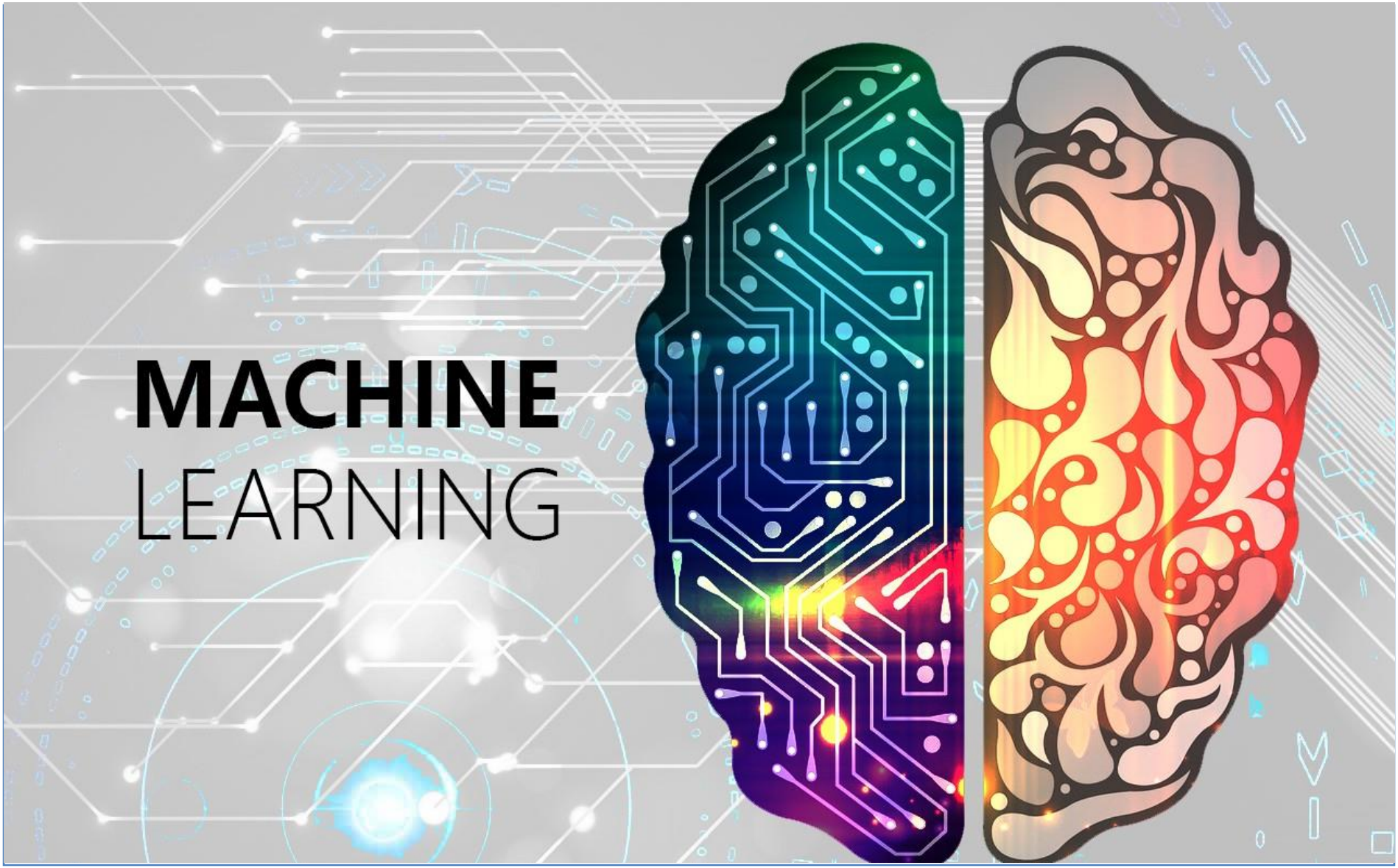
Machine Learning

Introduction To Machine Learning

GlobNet
HARNESSING SKILLS



Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Introduction

Machine learning (ML) is a sub-branch of AI that focuses on teaching computers how to learn without the need to be programmed for specific tasks. In fact, the key idea behind ML is that it is possible to create algorithms that learn from and make predictions on data.

This chapter will move into practical aspects of machine learning, primarily using Python's Scikit-Learn package.

The goals of this chapter are:

- To introduce the fundamental terminologies and concepts of ML.
- To introduce the Scikit-Learn API and show some examples of its use.
- To take a deeper dive into the details of several of the most important machine learning approaches.

Introduction To Machine Learning

What Is Machine Learning?

Machine Learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models proper parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Categories of Machine Learning

Machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

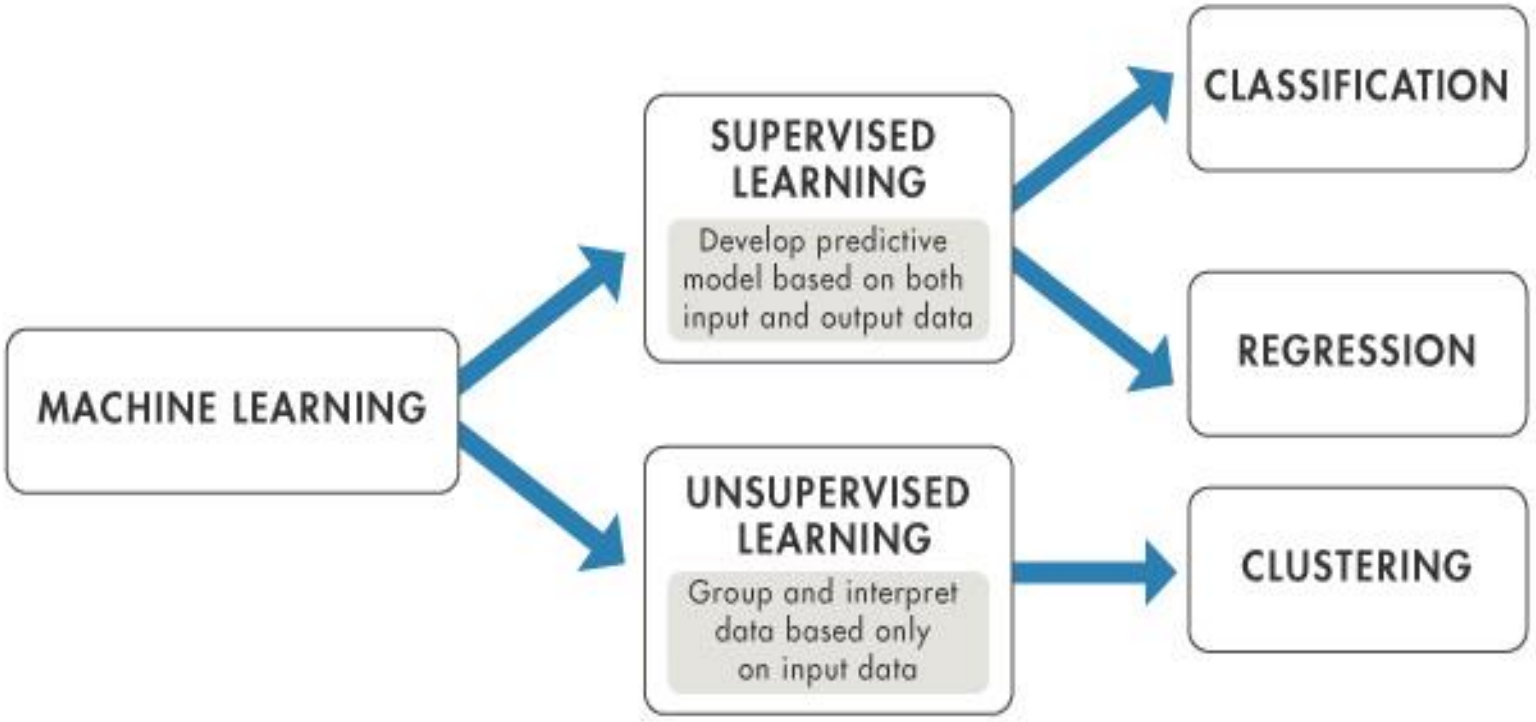
Categories of Machine Learning

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more proper representations of the data.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Categories of Machine Learning



Machine learning categorized into two main types: supervised learning and unsupervised learning.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Qualitative Examples of Machine Learning Applications

Let's take a look at a few very simple examples of a machine learning task. These examples are meant to give a simple, non-quantitative overview of the types of machine learning tasks we will be looking at in this chapter. Afterwards, we will go into more depth regarding the particular models and how they are used.

Classification: Predicting discrete labels

We will first take a look at a simple classification task, in which we are given a set of labeled points and want to use these to classify some unlabeled points.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Qualitative Examples of Machine Learning Applications

Classification: Predicting discrete labels

We will first take a look at a simple classification task, in which we are given a set of labeled points and want to use these to classify some unlabeled points.

Imagine that we have the data shown in Figure 6-1

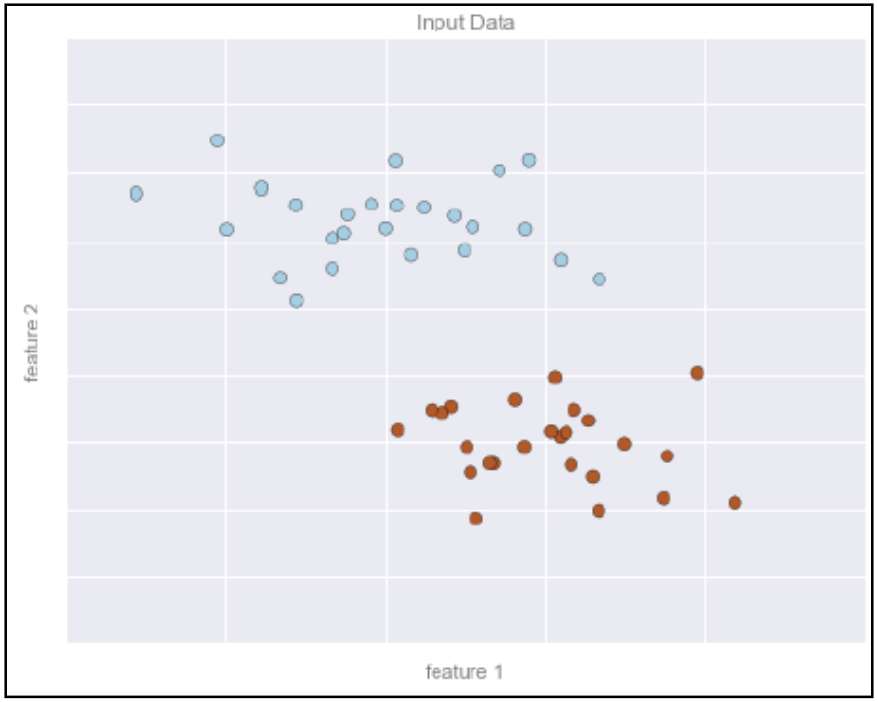


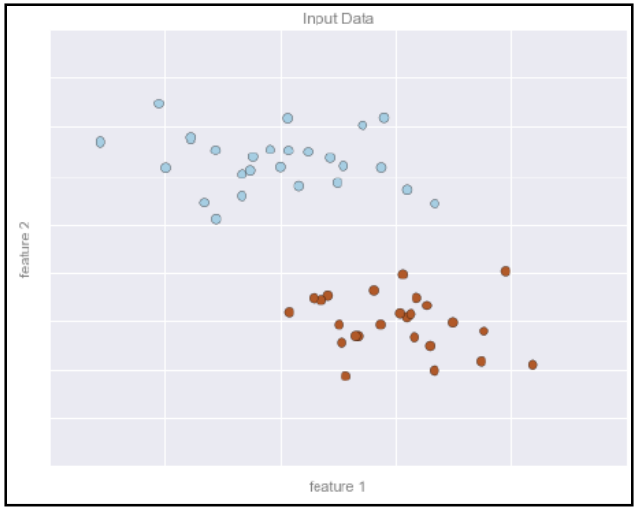
Figure 6-1. A simple data set for classification

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Qualitative Examples of Machine Learning Applications

Here we have two-dimensional data; that is, we have two features for each point, represented by the (x,y) positions of the points on the plane. In addition, we have one of two class labels for each point, here represented by the colors of the points. From these features and labels, we would like to create a model that will let us decide whether a new point should be labeled "blue" or "red."



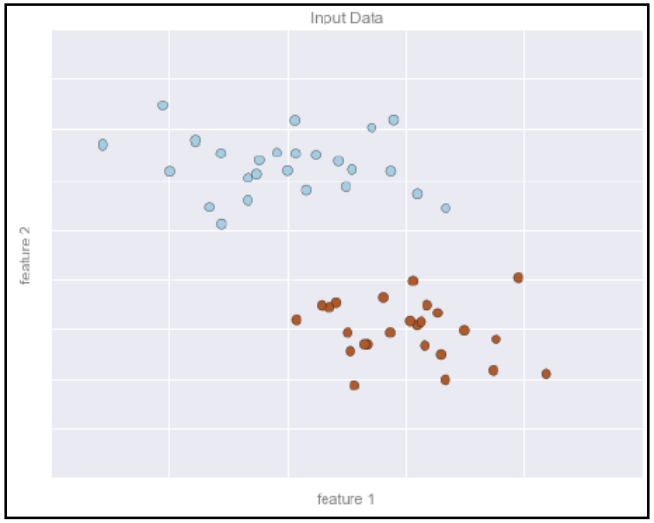
Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Qualitative Examples of Machine Learning Applications

There are a number of possible models for such a classification task, but here we will use an extremely simple one. We will make the assumption that the two groups can be separated by drawing a straight line through the plane between them, such that points on each side of the line fall in the same group. Here the model is a quantitative version of the statement "a straight line separates the classes," while the model parameters are the particular numbers describing the location and orientation of that line for our data.

The optimal values for these model parameters are learned from the data (this is the "learning" in machine learning), which is often called training the model.



Introduction To Machine Learning

Qualitative Examples of Machine Learning Applications

Figure 6-2 is a visual representation of what the trained model looks like for this data.

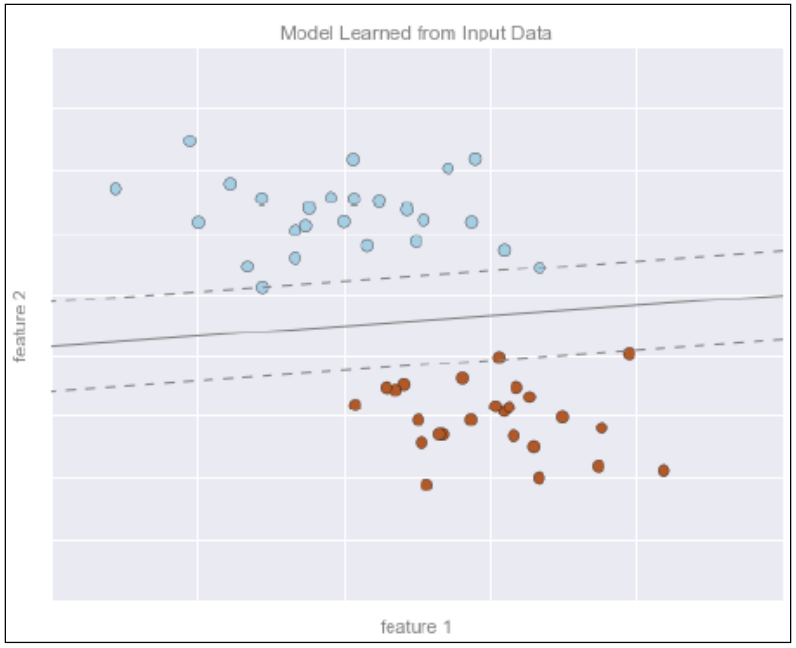


Figure 6-2. A simple classification model

Some of the

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Qualitative Examples of Machine Learning Applications

Now that this model has been trained, it can be generalized to new, unlabeled data. In other words, we can take a new set of data, draw this model line through it, and assign labels to the new points based on this model. This stage is usually called prediction. See the next figure.

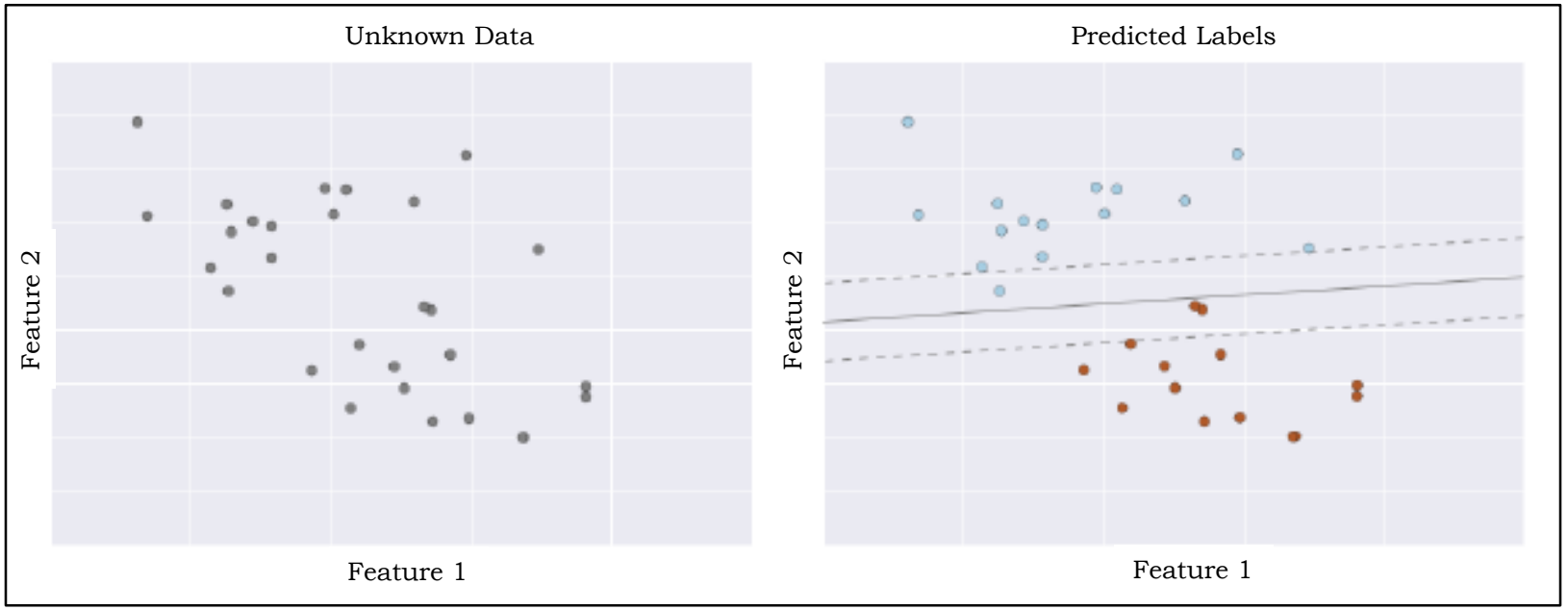


Figure 6-3. Applying a classification model to new data

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Qualitative Examples of Machine Learning Applications

This is the basic idea of a classification task in machine learning, where "classification" indicates that the data has discrete class labels. At first look this may look fairly trivial: it would be relatively easy to simply look at this data and draw such a discriminatory line to accomplish this classification. A benefit of the machine learning approach, is that it can generalize to much larger datasets in many more dimensions.

For example, this is similar to the task of automated spam detection for email; in this case, we might use the following features and labels:

- feature 1, feature 2, etc. normalized counts of important words or phrases ("Insurance," "Life-Security," etc.)
- label "spam" or "not spam"

Introduction To Machine Learning

Qualitative Examples of Machine Learning Applications

For the training set, these labels might be determined by individual inspection of a small representative sample of emails; for the remaining emails, the label would be determined using the model. For a suitably trained classification algorithm with enough well-constructed features (typically thousands or millions of words or phrases), this type of approach can be very effective.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Regression: Predicting continuous labels

In comparison with the discrete labels of a classification algorithm, we will next look at a simple regression task in which the labels are continuous quantities.

Consider the data shown in Figure 6-4, which consists of a set of points, each with a continuous label.

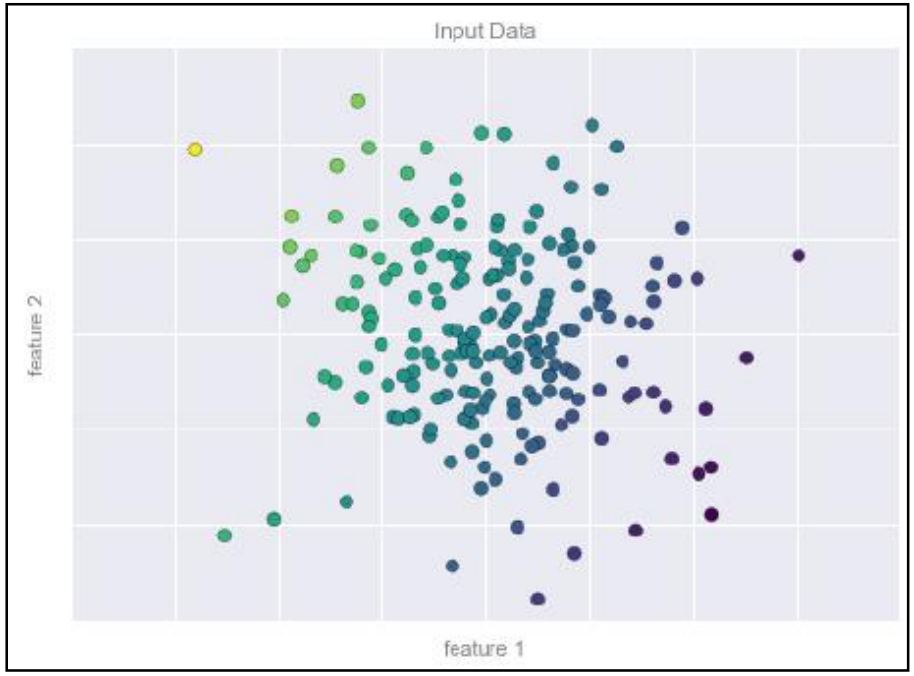


Figure 6-4. A simple dataset for regression

Introduction To Machine Learning

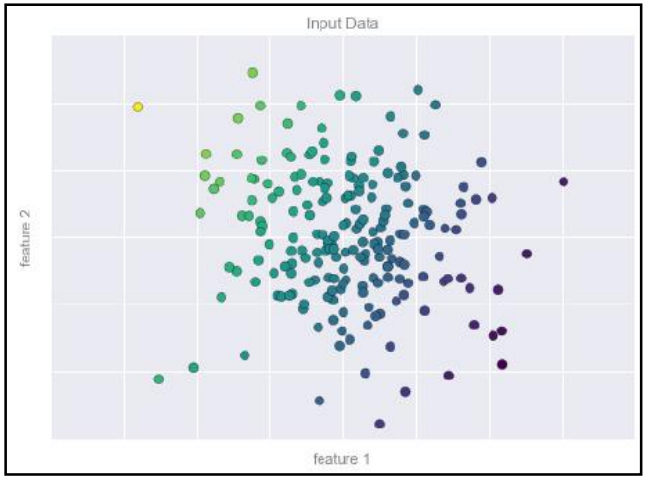
GlobNet HARNESSING SKILLS

Regression: Predicting continuous labels

As with the classification example, we have two-dimensional data; that is, there are two features describing each data point. The color of each point represents the continuous label for that point.

There are a number of possible regression models we might use for this type of data, but here we will use a simple linear regression to predict the points. This simple linear regression model assumes that if we treat the label as a third spatial dimension, we can fit a plane to the data.

This is a higher-level generalization of the well-known problem of fitting a line to data with two coordinates.



Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Regression: Predicting continuous labels

We can visualize this setup as shown in Figure 6-6.

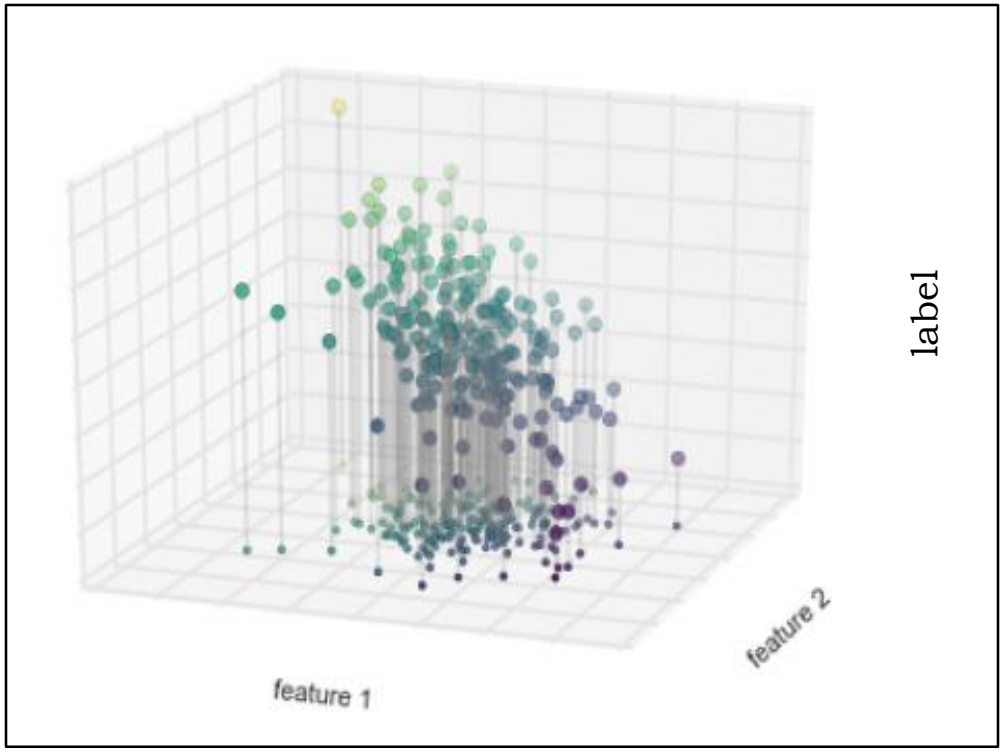


Figure 6-5. A three-dimensional view of the regression data

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Regression: Predicting continuous labels

Notice that the feature 1–feature 2 plane here is the same as in the two-dimensional plot from before; in this case, however, we have represented the labels by both color and three-dimensional axis position. From this view, it seems reasonable that fitting a plane through this three-dimensional data would allow us to predict the expected label for any set of input parameters.

Returning to the two-dimensional projection, when we fit such a plane we get the result shown in Figure 6-6.

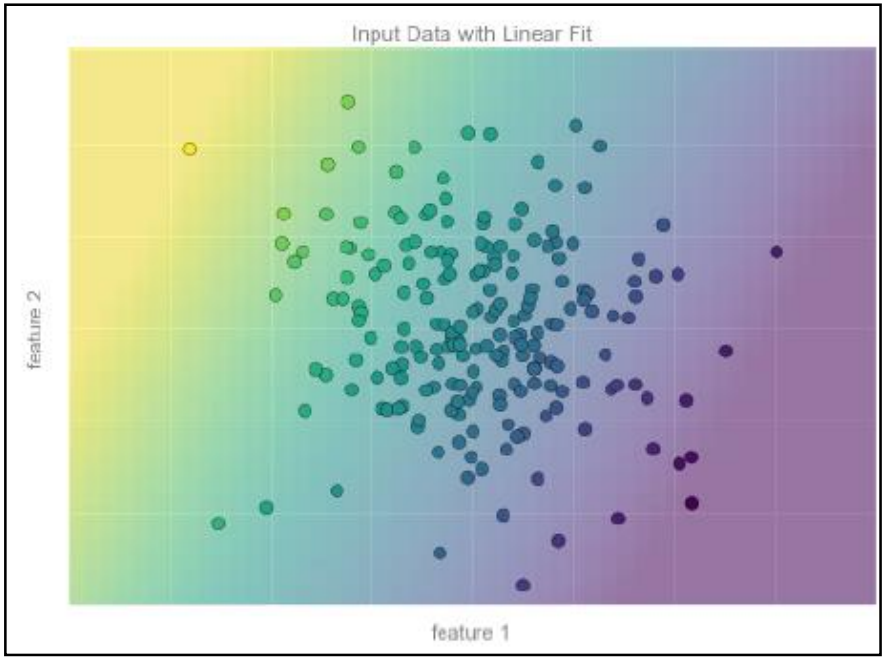


Figure 6-6. A representation of the regression model.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Regression: Predicting continuous labels

This plane of fit gives us what we need to predict labels for new points. Visually, we find the results shown in Figure 6-7.

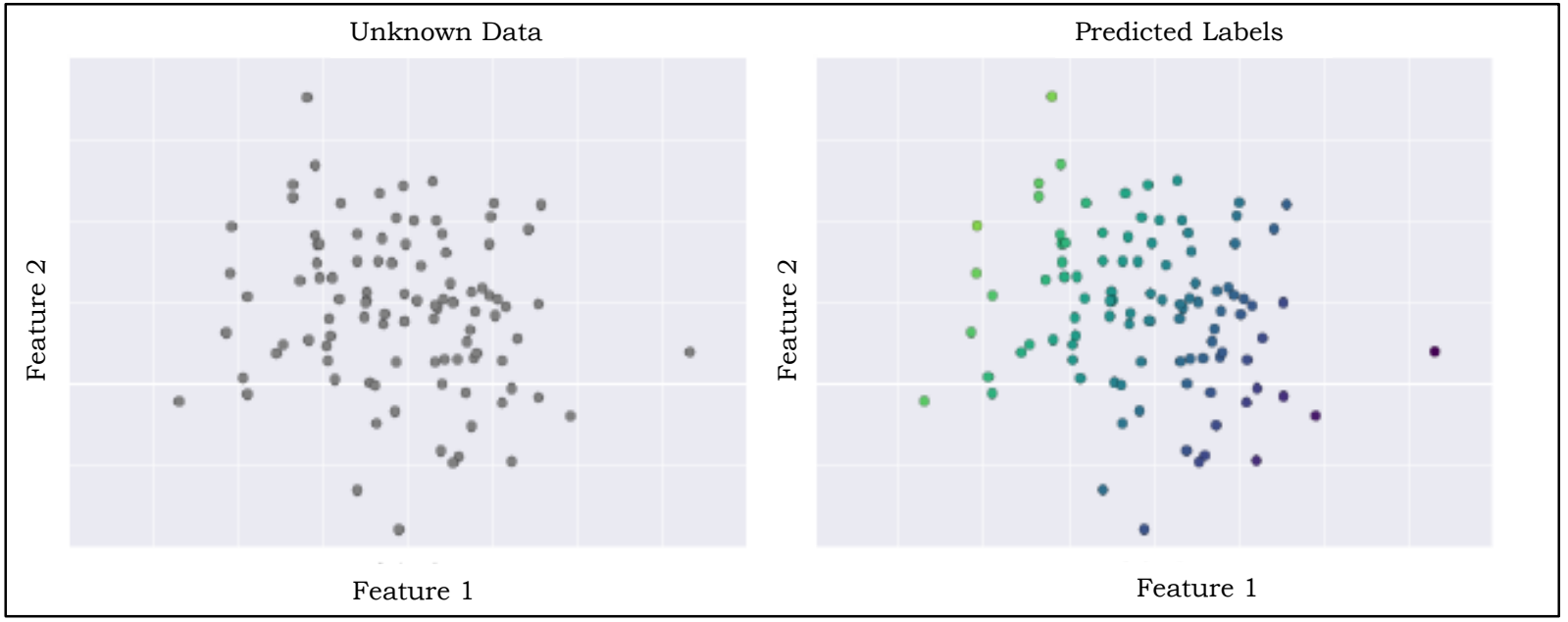


Figure 6-7. Applying the regression model to new data

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Regression: Predicting continuous labels

As with the classification example, this may seem rather trivial in a low number of dimensions. But the power of these methods is that they can be straightforwardly applied and evaluated in the case of data with many, many features.

For example, this is similar to the task of computing the distance to galaxies observed through a telescope—in this case, we might use the following features and labels:

- feature 1, feature 2, etc. brightness of each galaxy at one of several wavelengths or colors
- label distance or redshift of the galaxy

Introduction To Machine Learning

Clustering: Inferring labels on unlabeled data

The classification and regression illustrations we looked at are examples of supervised learning algorithms, in which we are trying to build a model that will predict labels for new data. Unsupervised learning involves models that describe data without reference to any known labels.

Introduction To Machine Learning

Clustering: Inferring labels on unlabeled data

One common case of unsupervised learning is "clustering" in which data is automatically assigned to some number of discrete groups. For example, we might have some two-dimensional data like that shown in Figure 6-8.

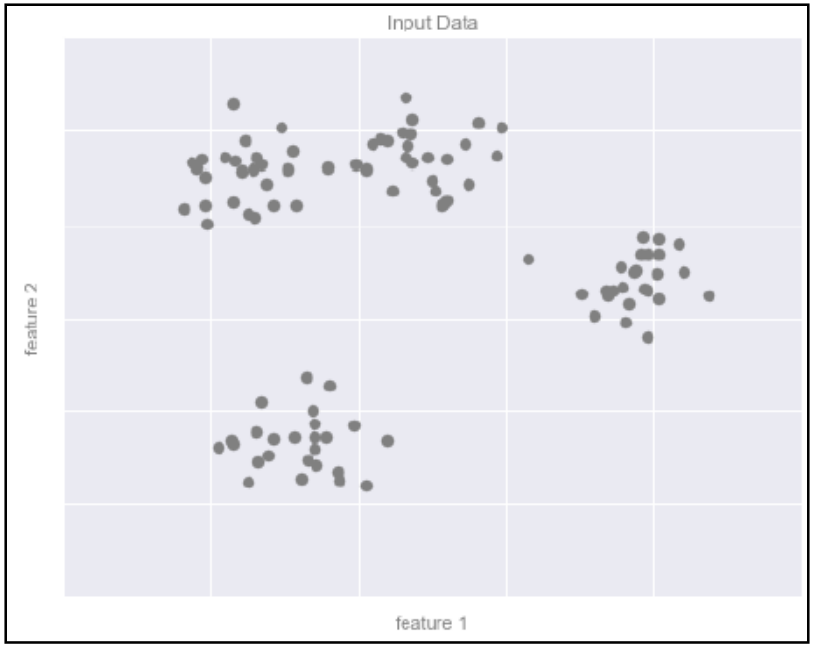


Figure 6-8. Example data for clustering

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Clustering: Inferring labels on unlabeled data

In normal view, it is clear that each of these points is part of a distinct group. Given this input, a clustering model will use the intrinsic structure of the data to determine which points are related. Using the very fast and intuitive k-means algorithm, we find the clusters shown in Figure 6-9.

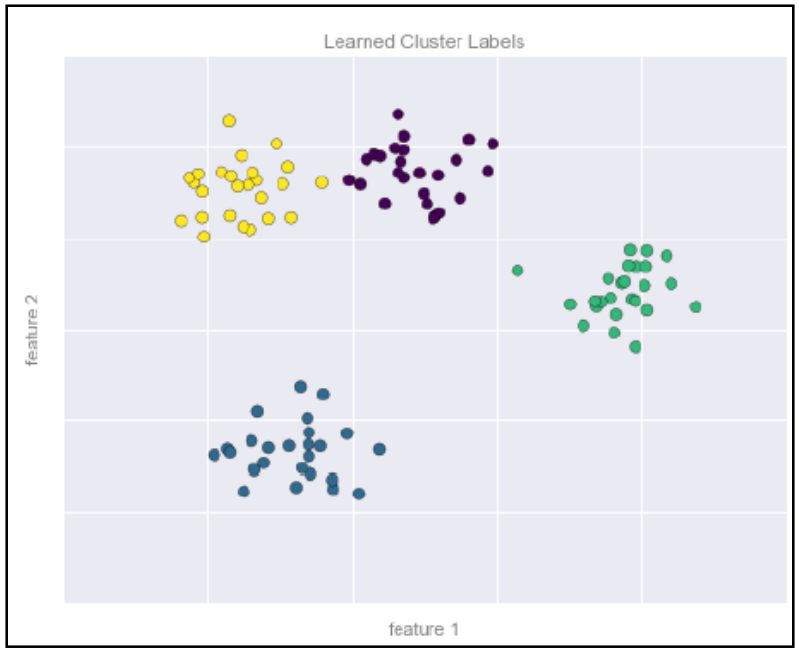


Figure 6-9. Data labeled with a k-means clustering model

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Dimensionality reduction: Inferring structure of unlabeled data

Dimensionality reduction is another example of an unsupervised algorithm, in which labels or other information are inferred from the structure of the dataset itself.

Dimensionality reduction is a bit more abstract than the examples we looked at before, but generally it seeks to pull out some low-dimensional representation of data that in some way preserves relevant qualities of the full dataset.

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Dimensionality reduction: Inferring structure of unlabeled data

As an example of this, consider the data shown in Figure 6-10.

As seen, it is clear that there is some structure in this data: it is drawn from a one-dimensional line that is arranged in a spiral within this two-dimensional space. In a sense, we can say that this data is "intrinsically" only one dimensional, though this one-dimensional data is embedded in higher-dimensional space.

A suitable dimensionality reduction model in this case would be sensitive to this nonlinear embedded structure, and be able to pull out this lower-dimensionality representation.

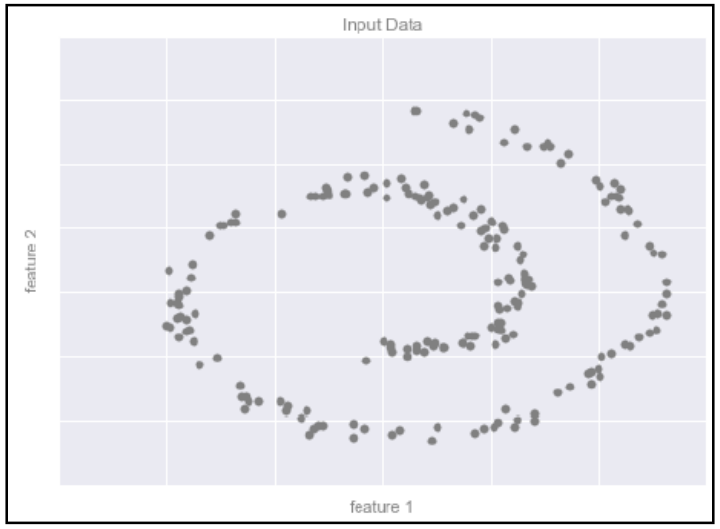


Figure 6-10. Example data for dimensionality reduction

Introduction To Machine Learning

GlobNet HARNESSING SKILLS

Dimensionality reduction: Inferring structure of unlabeled data

The figure shown next, Figure 6-11 presents a visualization of the results of the Isomap algorithm, a manifold learning algorithm that does exactly this.

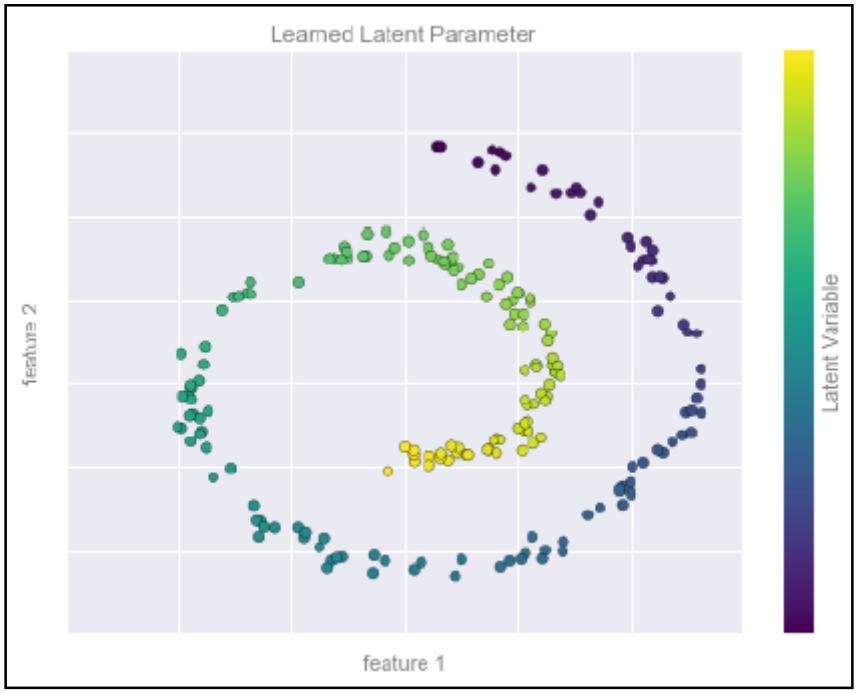


Figure 6-11. Data with a label learned via dimensionality reduction

Introduction To Machine Learning

Dimensionality reduction: Inferring structure of unlabeled data

Notice that the colors (which represent the extracted one-dimensional latent variable) change uniformly along the spiral, which indicates that the algorithm did in fact detect the structure we saw. As with the previous examples, the power of dimensionality reduction algorithms becomes clearer in higher-dimensional cases.

For example, we might wish to visualize important relationships within a dataset that has 100 or 1,000 features. Visualizing 1,000-dimensional data is a challenge, and one way we can make this more manageable is to use a dimensionality reduction technique to reduce the data to two or three dimensions.

Thank You