

identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

In addition, there are so-called *semi-supervised learning* methods, which fall somewhere between supervised learning and unsupervised learning. Semi-supervised learning methods are often useful when only incomplete labels are available.

## Qualitative Examples of Machine Learning Applications

To make these ideas more concrete, let's take a look at a few very simple examples of a machine learning task. These examples are meant to give an intuitive, non-quantitative overview of the types of machine learning tasks we will be looking at in this chapter. In later sections, we will go into more depth regarding the particular models and how they are used. For a preview of these more technical aspects, you can find the Python source that generates the figures in the [online appendix](#).

### Classification: Predicting discrete labels

We will first take a look at a simple *classification* task, in which you 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 5-1](#) (the code used to generate this figure, and all figures in this section, is available in the [online appendix](#)).

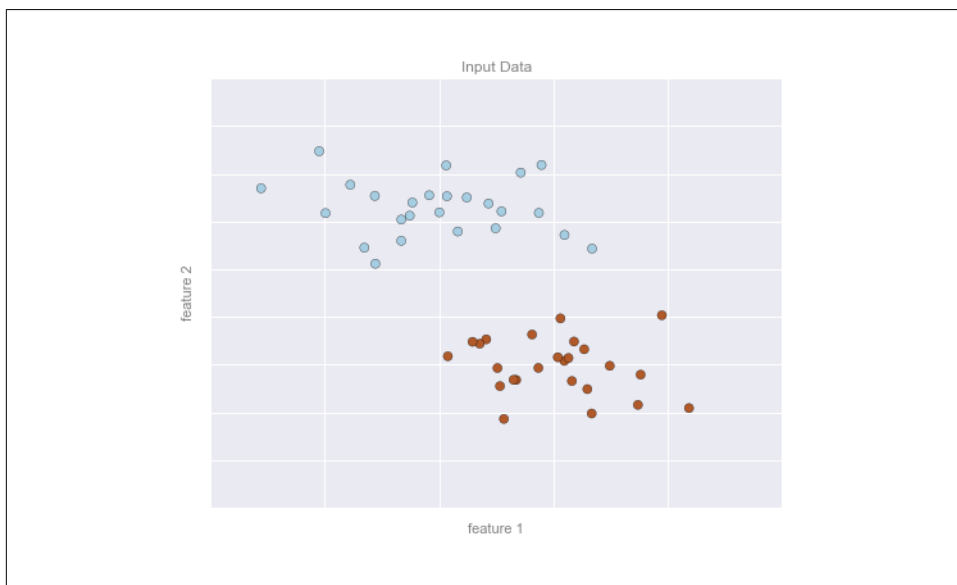


Figure 5-1. A simple data set for classification

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.”

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*.

Figure 5-2 is a visual representation of what the trained model looks like for this data.

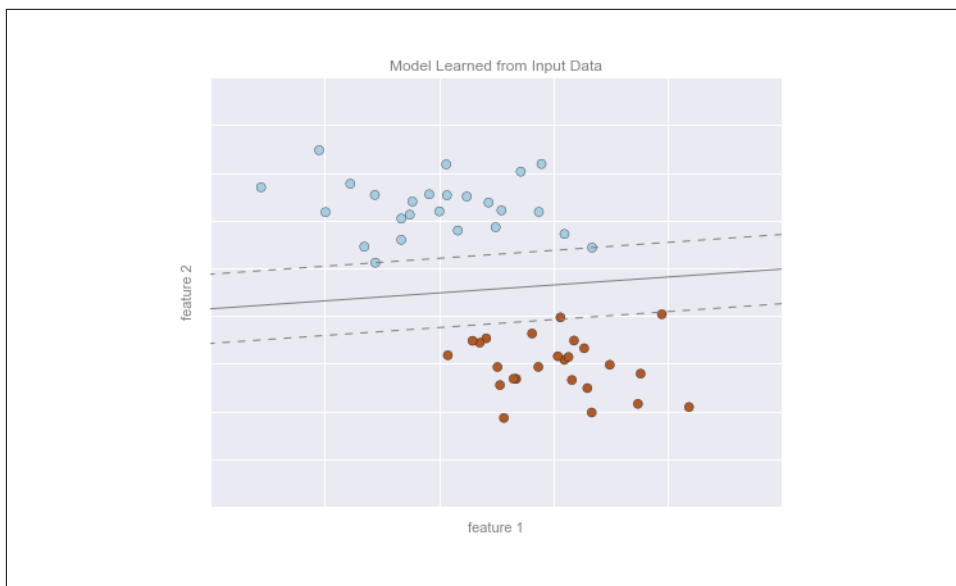


Figure 5-2. A simple classification model

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 Figure 5-3.

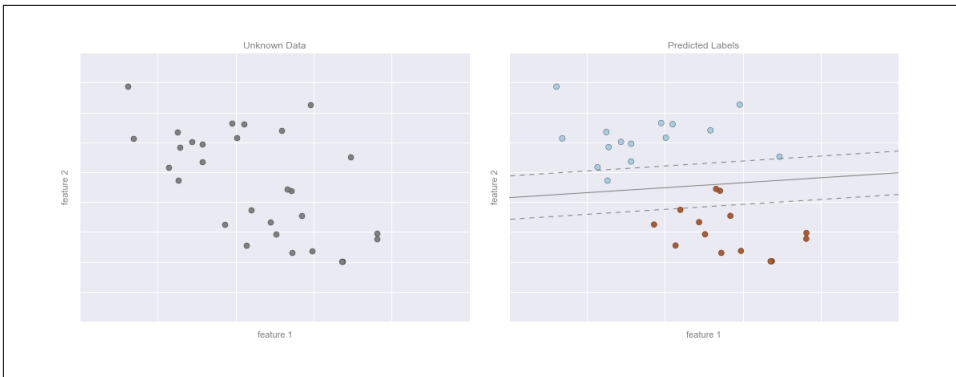


Figure 5-3. Applying a classification model to new data

This is the basic idea of a classification task in machine learning, where “classification” indicates that the data has discrete class labels. At first glance 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, however, 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 (“Viagra,” “Nigerian prince,” etc.)
- *label* → “spam” or “not spam”

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. We will see an example of such text-based classification in [“In Depth: Naive Bayes Classification” on page 382](#).

Some important classification algorithms that we will discuss in more detail are Gaussian naive Bayes (see [“In Depth: Naive Bayes Classification” on page 382](#)), support vector machines (see [“In-Depth: Support Vector Machines” on page 405](#)), and random forest classification (see [“In-Depth: Decision Trees and Random Forests” on page 421](#)).

## Regression: Predicting continuous labels

In contrast 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 5-4](#), which consists of a set of points, each with a continuous label.



*Figure 5-4. A simple dataset for regression*

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.

We can visualize this setup as shown in [Figure 5-5](#).

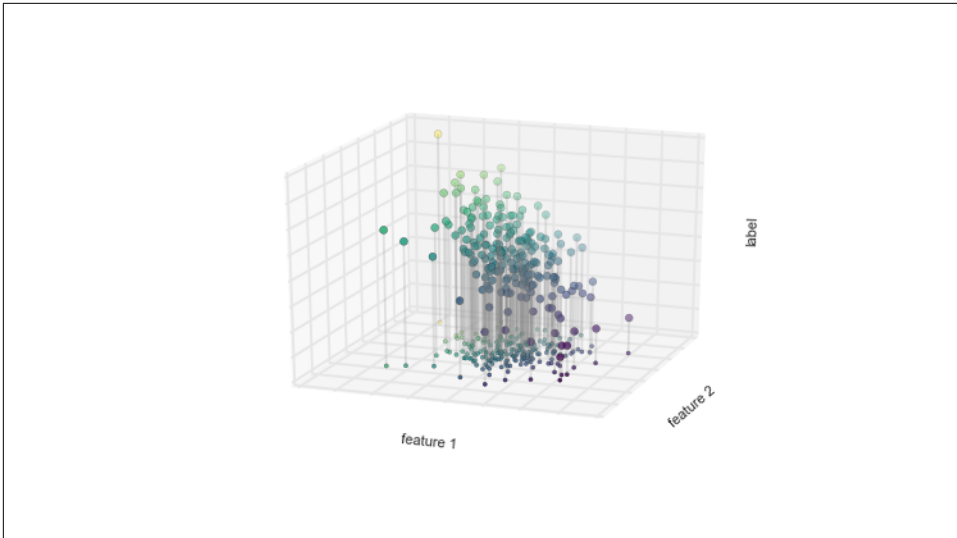


Figure 5-5. A three-dimensional view of the regression data

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 5-6](#).

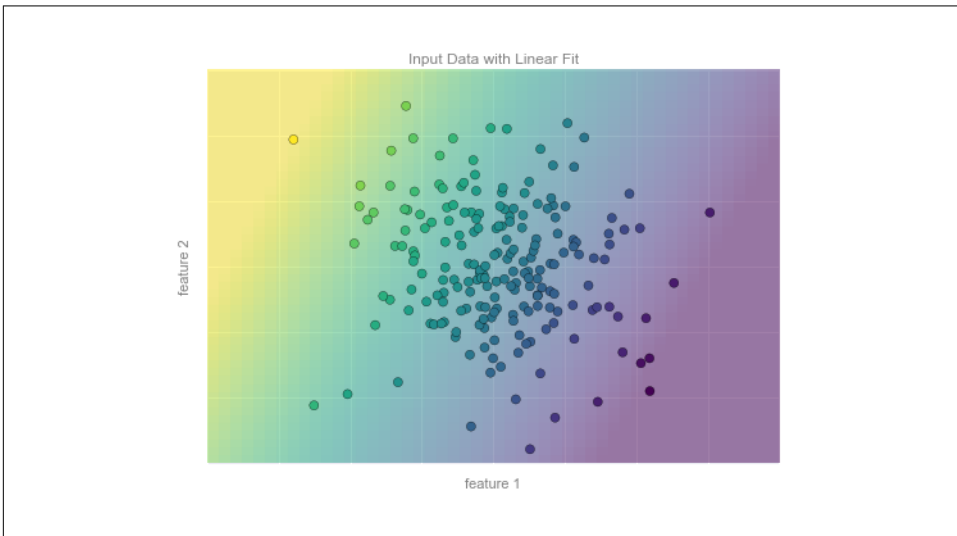
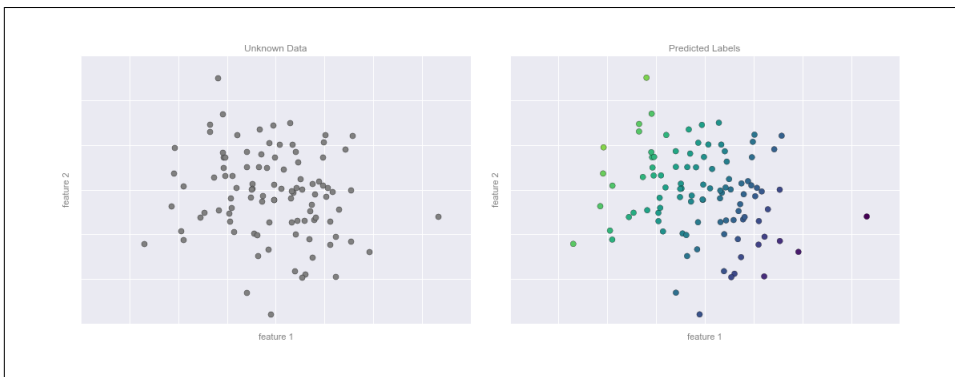


Figure 5-6. A representation of the regression model

This plane of fit gives us what we need to predict labels for new points. Visually, we find the results shown in [Figure 5-7](#).



*Figure 5-7. Applying the regression model to new data*

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

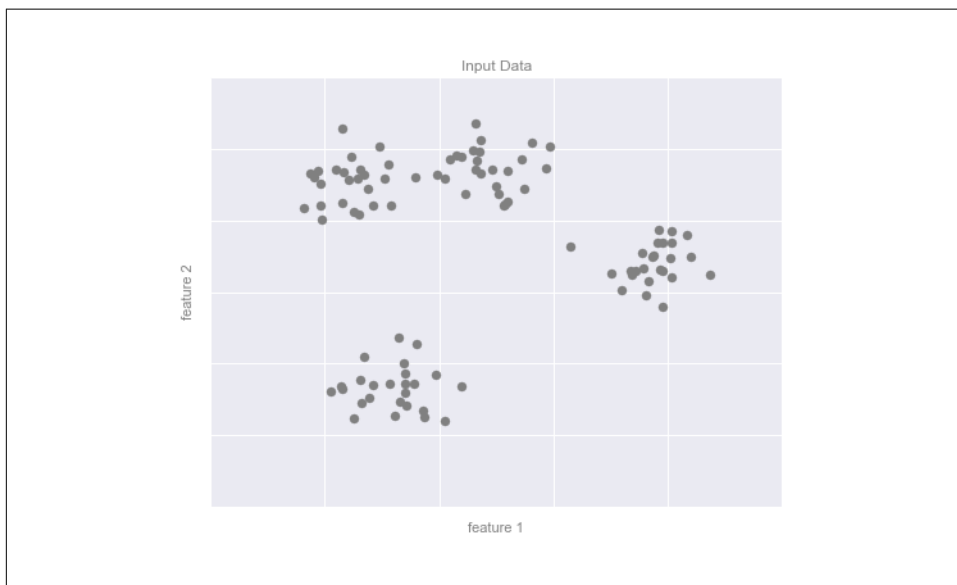
The distances for a small number of these galaxies might be determined through an independent set of (typically more expensive) observations. We could then estimate distances to remaining galaxies using a suitable regression model, without the need to employ the more expensive observation across the entire set. In astronomy circles, this is known as the “photometric redshift” problem.

Some important regression algorithms that we will discuss are linear regression (see [“In Depth: Linear Regression” on page 390](#)), support vector machines (see [“In-Depth: Support Vector Machines” on page 405](#)), and random forest regression (see [“In-Depth: Decision Trees and Random Forests” on page 421](#)).

### **Clustering: Inferring labels on unlabeled data**

The classification and regression illustrations we just 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.

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 5-8](#).



*Figure 5-8. Example data for clustering*

By eye, 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 (see [“In Depth:  \$k\$ -Means Clustering” on page 462](#)), we find the clusters shown in [Figure 5-9](#).

$k$ -means fits a model consisting of  $k$  cluster centers; the optimal centers are assumed to be those that minimize the distance of each point from its assigned center. Again, this might seem like a trivial exercise in two dimensions, but as our data becomes larger and more complex, such clustering algorithms can be employed to extract useful information from the dataset.

We will discuss the  $k$ -means algorithm in more depth in [“In Depth:  \$k\$ -Means Clustering” on page 462](#). Other important clustering algorithms include Gaussian mixture models (see [“In Depth: Gaussian Mixture Models” on page 476](#)) and spectral clustering (see [Scikit-Learn’s clustering documentation](#)).

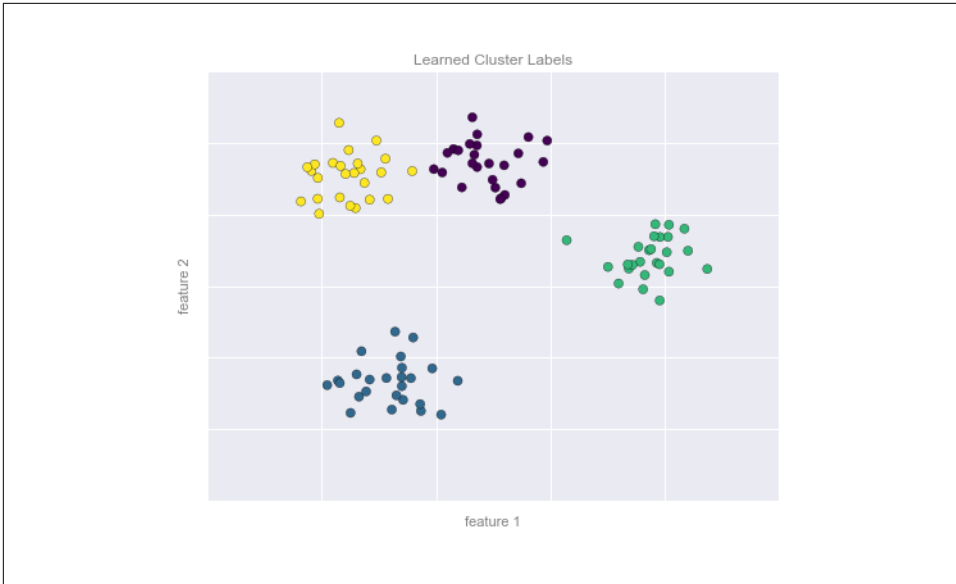


Figure 5-9. Data labeled with a *k*-means clustering model

### 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. Different dimensionality reduction routines measure these relevant qualities in different ways, as we will see in [“In-Depth: Manifold Learning” on page 445](#).

As an example of this, consider the data shown in [Figure 5-10](#).

Visually, 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, you could 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.



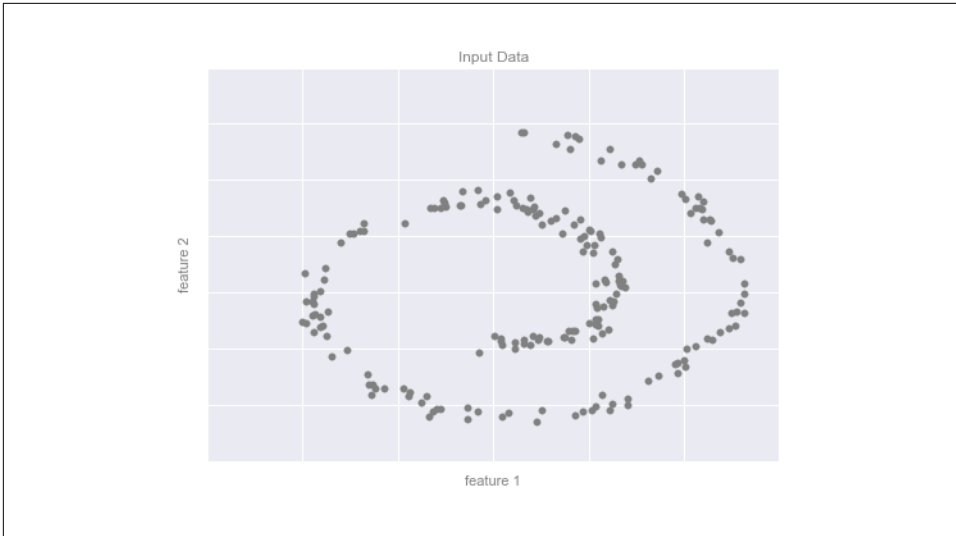


Figure 5-10. Example data for dimensionality reduction

Figure 5-11 presents a visualization of the results of the Isomap algorithm, a manifold learning algorithm that does exactly this.

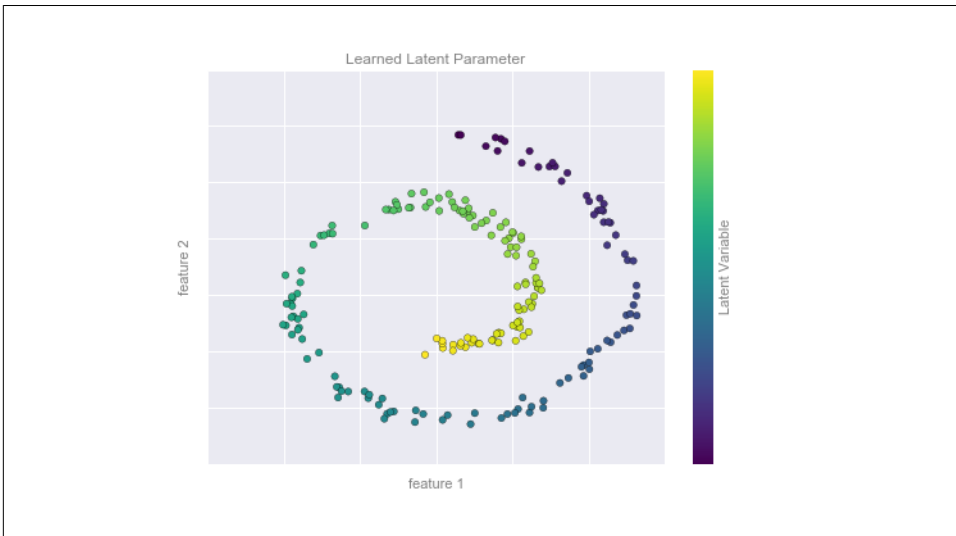


Figure 5-11. Data with a label learned via dimensionality reduction

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 by eye. 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.

Some important dimensionality reduction algorithms that we will discuss are principal component analysis (see “[In Depth: Principal Component Analysis](#)” on page 433) and various manifold learning algorithms, including Isomap and locally linear embedding (see “[In-Depth: Manifold Learning](#)” on page 445).

## Summary

Here we have seen a few simple examples of some of the basic types of machine learning approaches. Needless to say, there are a number of important practical details that we have glossed over, but I hope this section was enough to give you a basic idea of what types of problems machine learning approaches can solve.

In short, we saw the following:

### *Supervised learning*

Models that can predict labels based on labeled training data

### *Classification*

Models that predict labels as two or more discrete categories

### *Regression*

Models that predict continuous labels

### *Unsupervised learning*

Models that identify structure in unlabeled data

### *Clustering*

Models that detect and identify distinct groups in the data

### *Dimensionality reduction*

Models that detect and identify lower-dimensional structure in higher-dimensional data

In the following sections we will go into much greater depth within these categories, and see some more interesting examples of where these concepts can be useful.

All of the figures in the preceding discussion are generated based on actual machine learning computations; the code behind them can be found in the [online appendix](#).