The boundary points of the respective classes which are known as the support vectors are essential in finding the optimal hyperplane. The support vectors are illustrated in Figure 22-4. The boundary points are called support vectors because they are used to determine the maximum distance between the class they belong to and the discriminant function separating the classes.

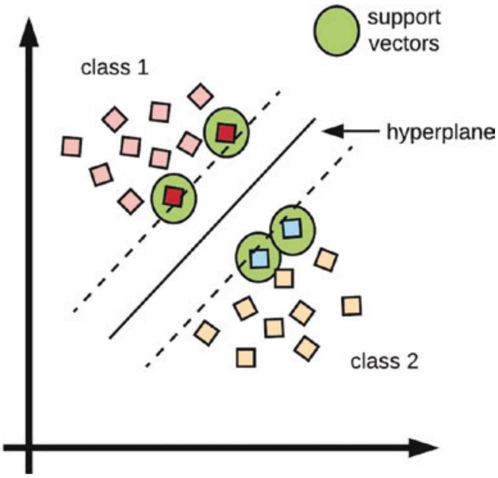

Figure 22-4. *Support vectors*

The mathematical formulation for finding the margin and consequently the hyperplane that maximizes the margin is beyond the scope of this book, but suffice to say this technique involves the Lagrange multiplier.

The Support Vector Classifier

In the real world, it is difficult to find data points that are precisely linearly separable and for which exists a large margin hyperplane. In Figure 22-5, the left image represents the data points for two classes in a dataset. Observe that there readily exists a linear separator between those two classes. Now, suppose we have an additional point from class 1 adjusted in such a way that it is much closer to class 2, we see that this point upsets the location of the hyperplane as seen in the right image of Figure 22-5. This reveals the sensitivity of the hyperplane to an additional data point that may result in a very narrow margin.

This sensitivity to data samples has significant drawbacks, the first being that the distance between the support vectors and the hyperplane reflects the confidence in the classification accuracy. Also, the drastic change in the position of the hyperplane due to a single additional point shows that the classifier is susceptible to high variability and can overfit the training data.

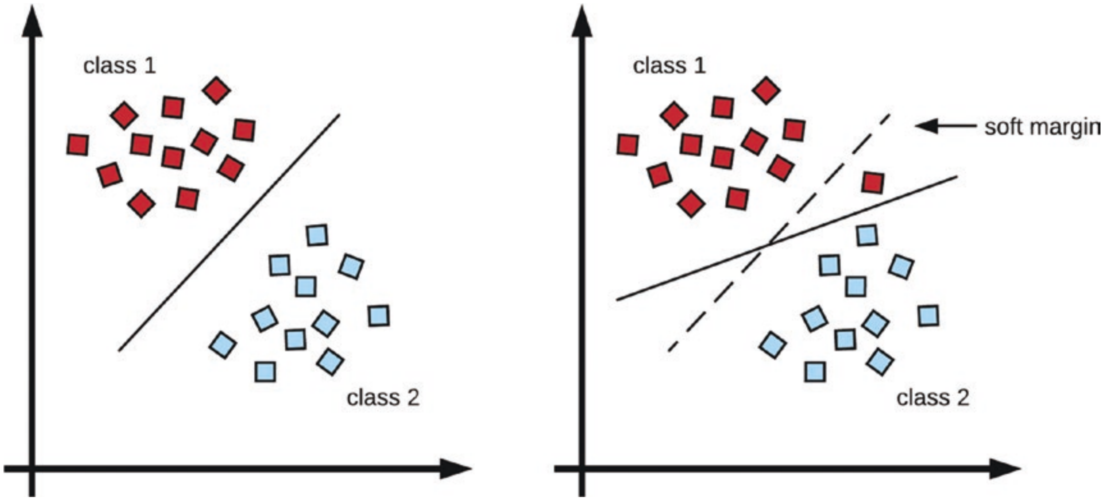

Figure 22-5. *Left: A linearly separable data distribution with a large margin. Right: The data point distribution makes it more difficult to find a large margin classifier that linearly separates the two classes*

The goal of the support vector classifier is to find a hyperplane that nearly discriminates between the two classes. This technique is also called a soft margin. A soft margin is tuned to ignore a degree of error when finding the separating hyperplane. This concept of a soft margin is how we generalize the support vector classifier to find a hyperplane in datasets that are not readily linearly separable. The margin is called soft because some examples are purposefully misclassified.

In such cases, as outlined in Figure 22-5, a soft margin classifier is preferred as it is more insensitive to individual data points and overall will have a better chance of generalizing to new examples. Howbeit, this might misclassify a couple of examples while training, but this is overall beneficial to the quality of the classifier as it generalizes to new samples.

Again, the margin is called soft because some examples are allowed to violate the margin or even be misclassified by the hyperplane to preserve overall generalizability. This is illustrated in Figure 22-6.

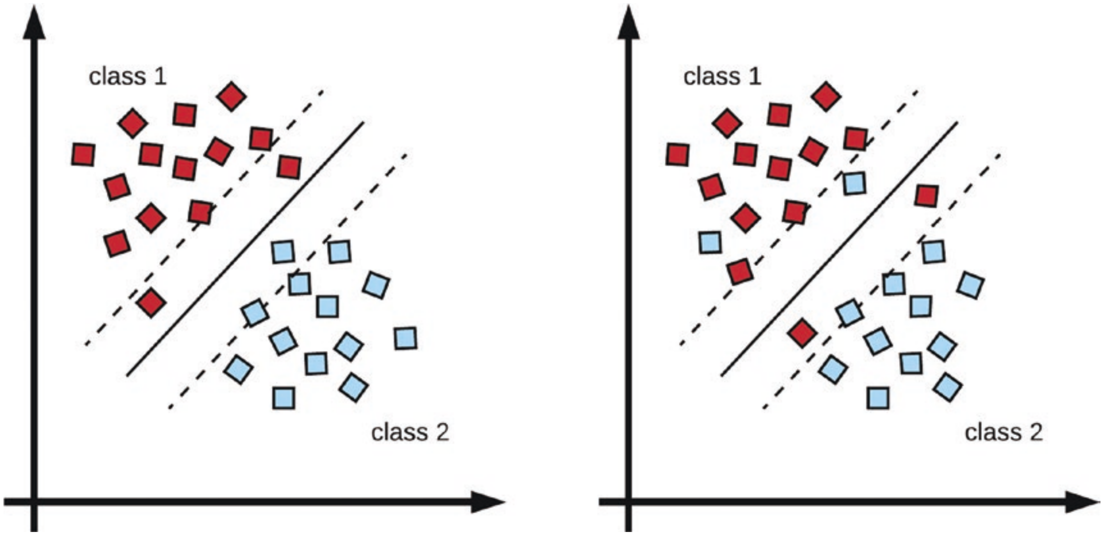

Figure 22-6. *Left: An example of a soft margin with points allowed to violate the margin. Right: An example with some points intentionally misclassified.*

The C Parameter

The C parameter is the hyper-parameter that is responsible for controlling the degree of violations to the margins or the number of intentionally misclassified points allowed by the support vector classifier. The C hyper-parameter is a non-negative real number. When this C parameter is set to 0, the classifier becomes the large margin classifier.

In a soft margin classifier, the C parameter is tuned by adjusting its values to control the tolerance of the margin. With larger values of C, the classifier margins become wider and more tolerant to violations and misclassifications. However, with smaller values of C, the margins become narrower and are less tolerant of violations and misclassified points.

Observe that the C hyper-parameter is vital for regulating the bias/variance trade-off of the support vector classifier. The higher the value of C, our classifier is more prone to variability in the data points and can under-simplify the learning problem. Also, if C is set closer to zero, it results in a much narrower margin, and this can overfit the classifier, leading to high variance – and this will likely fail to generalize to new examples (see Figure 22-7).