**Chapter 9 – Support vector machines**

* SVMs have been shown to perform well in a variety of settings, and are often considered one of the best “out of the box” classifiers.
* Support vector machines are intended for the binary classification setting in which there are two classes.

*Part 1: Maximal Margin Classifier*

* In a p-dimensional space, a hyperplane is a flat affine subspace of dimension p − 1.1 For instance, in two dimensions, a hyperplane is a flat one-dimensional subspace—in other words, a line. In three dimensions, a hyperplane is a flat two-dimensional subspace—that is, a plane.
* We have seen a number of approaches to classification, we will now see a new approach that is based upon the concept of a *separating hyperplane*.
* If a separating hyperplane exists, we can use it to construct a very natural
* classifier: a test observation is assigned a class depending on which side of the hyperplane it is located
* Not surprisingly, and as we see in Figure 9.2, a classifier that is based on a separating hyperplane leads to a linear decision boundary.
* In general, if our data can be perfectly separated using a hyperplane, then there will in fact exist an infinite number of such hyperplanes. This is because a given separating hyperplane can usually be shifted a tiny bit up or down, or rotated, without coming into contact with any of the observations. Three possible separating hyperplanes are shown in the left-hand panel of Figure 9.2. In order to construct a classifier based upon a separating hyperplane, we must have a reasonable way to decide which of the infinite possible separating hyperplanes to use.
* A natural choice is the maximal margin hyperplane (also known as the optimal separating hyperplane), which is the separating hyperplane that is farthest from the training observations.
* The maximal margin hyperplane is the separating hyperplane for which the margin is largest—that is, it is the hyperplane that has the farthest minimum dis- tance to the training observations.
* We can then classify a test observation based on which side of the maximal margin hyperplane it lies. This is known as the maximal margin classifier.
* We hope that a classifier that has a large margin on the training data will also have a large margin on the test data, and hence will classify the test observations correctly.
* Although the maximal margin classifier is often successful, it can also lead to overfitting when p is large.
* Examining Figure 9.3, we see that three training observations are equidis- tant from the maximal margin hyperplane and lie along the dashed lines indicating the width of the margin. These three observations are known as support vectors, since they are vectors in *p*-dimensional space, and they ‘support’ the maximal margin hyperplane in the sense that if these points were moved slightly than the maximal margin hyper-plane would move as well.
* The distance of an observation from the hyperplane can be seen as a measure of our confidence that the observation was correctly classified

*The non-separable model*

* The maximal margin classifier is a very natural way to perform classification *if a separating hyperplane exists.*
* In some cases, however, we cannot *exactly* separate the two classes. However, as we will see in the next section, we can extend the concept of a separating hyperplane in order to develop a hyperplane that almost separates the classes, using a so-called soft margin. The generalization of the maximal margin classifier to the non-separable case is known as the *support vector classifier.*

*Support vector classifiers*

* We might be willing to consider a classifier based on a hyperplane that does not perfectly separate the two classes, in the interest of (a) greater robustness to individual observations, and (b) better classification of *most* of the training observations.
* The support vector classifier, sometimes called a soft margin classifier, does exactly this.
* Rather than seeking the largest possible margin so that every observation is not only on the correct side of the hyperplane but also on the correct side of the margin, we instead allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane
* We now consider the role of the tuning parameter C. In (9.15), C bounds the sum of the εi’s, and so it determines the number and severity of the violations to the margin (and to the hyperplane) that we will tolerate. We can think of C as a budget for the amount that the margin can be violated by the n observations
* As the budget C increases, we become more tolerant of violations to the margin, and so the margin will widen. Conversely, as C decreases, we become less tolerant of violations to the margin and so the margin narrows.
* In practice, C is treated as a tuning parameter that is generally chosen via cross-validation. As with the tuning parameters that we have seen through- out this book, C controls the bias-variance trade-off of the statistical learn- ing technique. When C is small, we seek narrow margins that are rarely violated; this amounts to a classifier that is highly fit to the data, which may have low bias but high variance. On the other hand, when C is larger, the margin is wider and we allow more violations to it; this amounts to fitting the data less hard and obtaining a classifier that is potentially more biased but may have lower variance.

*Support vector machines*

* The support vector machine (SVM) is an extension of the support vector classifier that results from enlarging the feature space in a specific way, using kernels.
* However, the main idea is described in Section 9.3.1: we may want to enlarge our feature space in order to accommodate a non-linear boundary between the classes.
* To summarize, in representing the linear classifier f (x), and in computing
* its coefficients, all we need are inner products.
* Using such a kernel with d > 1, instead of the standard linear kernel (9.21), in the support vector classifier algorithm leads to a much more flexible decision boundary. It essentially amounts to fitting a support vector classifier in a higher-dimensional space involving polynomials of degree d, rather than in the original feature space.
* Another popular choice is the radial kernel,

*SVMs with more than two classes*

* Though a number of proposals for extending SVMs to the K-class case have been made, the two most popular are the one-versus-one and one-versus-all approaches.
* A one-versus-one or all-pairs approach constructs K SVMs, each of which compares a pair of classes. For example, one such SVM might compare the kth class, coded as +1, to the k′th class, coded as −1. We classify a test observation using each of the K classifiers, and 2 we tally the number of times that the test observation is assigned to each of the K classes.
* The one-versus-all approach is an alternative procedure for applying SVMs in the case of K > 2 classes. We fit K SVMs, each time comparing one of the K classes to the remaining K − 1 classes.

*Relationship with logistic regression*

* When the classes are well separated, SVMs tend to behave better than logistic regression; in more overlapping regimes, logistic regression is often preferred.
* The choice of tuning parameter is very important and determines the extent to which the model underfits or over- fits the data, as illustrated, for example, in Figure 9.7.
* Is the SVM unique in its use of kernels to enlarge the feature space to accommodate non-linear class boundaries? The answer to this question is “no”
* However, for historical reasons, the use of non-linear kernels is much more widespread in the context of SVMs than in the context of logistic regression or other methods.