More on Classification: Support Vector Machine

- The Support Vector Machine (SVM) is a classification method approach developed in the computer science field in the 1990s.
- It has shown good performance in classification of test observations for a variety of types of data.
- We first introduce a simple classifier, related to the SVM, called the maximal margin classifier.

Maximal Margin Classifier (MMC)

- A MMC depends on the notion of a *hyperplane*.
- In 2-dimensional space, a *hyperplane* is simply a line, and can be written with the equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

- In 3-dimensional space, a hyperplane is a flat plane (a flat surface cutting through the 3-D space).
- In q-dimensional space, a hyperplane is hard to visualize, but it has the equation

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q = 0$$

Classification Using a MMC

- Suppose we have a set of training observations having q numerical variables measured on them and falling into 2 categories.

- The training data is perfectly *separable* if a hyperplane (called a separating hyperplane) can be drawn so that all points from one category fall on one side of the hyperplane and all points from the other category fall on the other side of the hyperplane.

- See picture in two dimensions.

- In this case, there are actually *infinitely many* separating hyperplanes.

- The *maximal margin hyperplane* is the hyperplane that lies the farthest (in perpendicular distance) from any of the training data.

- If we imagine separating the two classes with a flat *slab* rather than a line, then the maximal margin hyperplane is the midline of that slab (see picture again).

The Classification Rule for the MMC

- Note that for a separating hyperplane, for any training observation, we have either

$$\beta_0 + \beta_{i1} X_1 + \cdots + \beta_q X_{iq} < 0 \text{ if } Y_i = 0$$

  and

$$\beta_0 + \beta_{i1} X_1 + \cdots + \beta_q X_{iq} > 0 \text{ if } Y_i = 1$$

  or else the same inequalities hold with the category labels 0 and 1 reversed.

- So for a new observation $\mathbf{x}_0$, we could classify it simply based on the sign of

$$\beta_0 + \beta_1 X_{01} + \cdots + \beta_q X_{0q}.$$

- And if this value is *far* from zero, we are *more confident* in our classification.

Support Vector Classifiers

- The major problem with the MMC is that in *most* training data sets, the categories are *not separable*.
- So no hyperplane can be constructed that perfectly separates the training data into classes.
- The MMC is also very sensitive to the position of individual observations near the separating hyperplane area.
- The *support vector classifier* (SVC) accepts the misclassification of a few training observations, but it is:
  - More robust to individual observations
  - Better at classification of test data

Details of the Support Vector Classifier

- The support vector classifier builds in a *margin* of width M around the hyperplane.
- It allows training observations to fall on the "wrong side" of the margin, or even on the wrong side of the hyperplane.
- The amount of the "margin violation" for each observation is measured by what are called "slack variables" $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$.
- If the i-th observation is on the "right side" of the margin, then $\epsilon_i = 0$.
- If the i-th observation is on the "wrong side" of the margin, then $\epsilon_i > 0$.
- If the i-th observation is on the wrong side of the hyperplane, then $\epsilon_i > 1$.
- The total allowed "margin violation" is controlled by a tuning parameter $C \geq 0$.

More Details of the Support Vector Classifier

- We seek to maximize the margin width M, for a given choice of C (see ISL page 346 for the mathematical details).
- If C is small (near zero), few violations are allowed, the margin will be narrow, and the classifier will be similar to a MMC.
- If C is larger, more violations are allowed, the margin will be wider, and the classifier will be more robust.
- Again, the classification of a new observation $\mathbf{x}_0$ is done by plugging $\mathbf{x}_0$ into the equation of the hyperplane and classifying based on the sign of the result.

Still More Details of the Support Vector Classifier

- Note that only the observations directly *on* the margin, or on the wrong side of the margin (these observations are called *support vectors*) affect the support vector classifier.

- When C is large and the margin is wide, there are many support vectors, and so the classifier is not as dependent on any single observation (it's a low-variance classifier).

- This robustness property of the support vector classifier is shared by the logistic classifier, but not by LDA, which is equally dependent on all observations.

Support Vector Machine

- The Support Vector Machine (SVM) generalizes the SVC by allowing a *nonlinear* decision boundary.
- This is accomplished through a *kernel function* that quantifies the similarity between observations.
- A *linear* kernel will simply produce the SVC, but a *polynomial* kernel or *radial* (exponential) kernel will allow a nonlinear boundary and will produce a SVM classifier.
- The `R` function `svm` in the `e1071` package implements the SVC and SVM methods.
- A `cost` argument that is inversely related to C can be specified.
- For the SVM with the polynomial kernel, the `degree` controls the departure of the decision boundary from linearity.
- With the radial kernel, we specify `gamma` to control the nonlinearity (this is often chosen via cross-validation).

Support Vector Machine with More than Two Classes

- The Support Vector Machine (SVM) does not work as naturally with more than two response categories.
- With more than two classes, the `R` function `svm` simply takes each possible pair of categories in turn, and attempts to classify a test observation into one of the two categories in that pair.
- It does this for all pairs of categories, and eventually classifies the observation to the category that was selected the most often, among the pairwise classifications.
- The ISL book called this approach “one-versus-one” classification.
- This is not the only way the classification could be done; Section 9.4.2 of the ISL book discusses the “one-versus-all” method of classification.