For example, if $p = 2$ and $q = 3$, we have the matrix

$$\mathbf{K} = \begin{pmatrix} \kappa(u_1, u_1) & \kappa(u_1, u_2) & -\kappa(u_1, v_1) & -\kappa(u_1, v_2) & -\kappa(u_1, v_3) \\ \kappa(u_2, u_1) & \kappa(u_2, u_2) & -\kappa(u_2, v_1) & -\kappa(u_2, v_2) & -\kappa(u_2, v_3) \\ -\kappa(v_1, u_1) & -\kappa(v_1, u_2) & \kappa(v_1, v_1) & \kappa(v_1, v_2) & \kappa(v_1, v_3) \\ -\kappa(v_2, u_1) & -\kappa(v_2, u_2) & \kappa(v_2, v_1) & \kappa(v_2, v_2) & \kappa(v_2, v_3) \\ -\kappa(v_3, u_1) & -\kappa(v_3, u_2) & \kappa(v_3, v_1) & \kappa(v_3, v_2) & \kappa(v_3, v_3) \end{pmatrix}.$$

Then the classification function

$$f(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle - b)$$

for points in the original data space $\mathbb{R}^n$ is also expressed solely in terms of the matrix $\mathbf{K}$ and the inner products $\kappa(u_i, x) = \langle \varphi(u_i), \varphi(x)\rangle$ and $\kappa(v_j, x) = \langle \varphi(v_j), \varphi(x)\rangle$. As a consequence, in the original data space $\mathbb{R}^n$, the hypersurface

$$\mathcal{S} = \{x \in \mathbb{R}^n \mid \langle w, \varphi(x)\rangle - b = 0\}$$

separates the data points u_i and v_j, but it is not an affine subspace of $\mathbb{R}^n$. The classification function f tells us on which "side" of $\mathcal{S}$ is a new data point $x \in \mathbb{R}^n$. Thus, we managed to separate the data points u_i and v_j that are not separable by an affine hyperplane, by a *nonaffine hypersurface* $\mathcal{S}$, by assuming that an embdedding $\varphi\colon \mathbb{R}^n \to F$ exists, even though we don't know what it is, but having access to F through the kernel function $\kappa\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by the inner products $\kappa(x, y) = \langle \varphi(x), \varphi(y)\rangle$.

In practice the art of using the kernel method is to choose the right kernel (as the knight says in Indiana Jones, to "choose wisely.").

The method of kernels is very flexible. It also applies to the soft margin versions of SVM, but also to regression problems, to principal component analysis (PCA), and to other problems arising in machine learning.

We discussed the method of kernels in Chapter 53. Other comprehensive presentations of the method of kernels are found in Schölkopf and Smola [145] and Shawe–Taylor and Christianini [159]. See also Bishop [23].

We first consider the soft margin SVM arising from Problem (SVM_{h1}).

54.1 Soft Margin Support Vector Machines; (SVM_{s1})

In this section we derive the dual function G associated with the following version of the soft margin SVM coming from Problem (SVM_{h1}), where the maximization of the margin δ has been replaced by the minimization of $-\delta$, and where we added a "regularizing term" $K\Big(\sum_{i=1}^p \epsilon_i + \sum_{j=1}^q \xi_j\Big)$ whose purpose is to make $\epsilon \in \mathbb{R}^p$ and $\xi \in \mathbb{R}^q$ *sparse* (that is, try to make ϵ_i and ξ_j have as many zeros as possible), where $K > 0$ is a fixed constant that can be