Chapter 7

Support Vector Machines

Assume we have a learning set $\mathcal{L} = \{(x_i, y_i) : i = 1, \ldots, n\}$ where $x_i \in \mathbb{R}^r$ (and r-variate observation, r real valued random variables) and $y_i \in \{-1, 1\}$. Here y_i is a class variable, two classes, which we label $+1$ and -1. We would like to construct a function $f : \mathbb{R}^r \to \mathbb{R}$ such that $C(x) = \text{sign}(f(x))$ is a classifier. The separating function f then classifies a test set $\mathcal{T}$ into two classes, Π_+ and Π_- depending on whether $f(x)$ is positive or negative.

7.1 Linear Separability

The learning set $\mathcal{L}$ is *linearly separable* if and only if there is a $\beta_0 \in \mathbb{R}$ and a $\beta \in \mathbb{R}^r$ such that $f(x) = \beta_0 + x'\beta$ separates $\mathcal{L}$; for each $(y_i, x_i) \in \mathcal{L}$, $f(x_i) > 0$ if $y_i = 1$ and $f(x_i) < 0$ if $y_i = -1$. The hyperplane $f(x) = 0$ is said to separate $\mathcal{L}$.

If such a f exists then, by rescaling, we can find β_0 and β such that

$$\begin{cases} \beta_0 + x_i'\beta \geq +1 & y_i = +1 \\ \beta_0 + x_i'\beta \leq -1 & y_i = -1. \end{cases}$$

Now consider the two hyperplanes $H_{+1} : (\beta_0 - 1) + x'\beta = 0$ and $H_{-1} : (\beta_0 + 1) + x'\beta = 0$. Points of $\mathcal{L}$ that lie in either H_{+1} or H_{-1} are said to be *support vectors*.

If x_{-1} lies on H_{-1} and x_{+1} lies on H_{+1} then

$$\begin{cases} (x_{+1}' - x_{-1}')\beta = 2 \\ \beta_0 = -\frac{1}{2}(x_{+1}' + x_{-1}')\beta. \end{cases}$$

The perpendicular distances of the hyperplane $\beta_0 + x'\beta = 0$ to the points x_{-1} and x_{+1} are:

$$d_- = \frac{|\beta_0 + x_{-1}'\beta|}{\|\beta\|} = \frac{1}{\|\beta\|} \qquad d_+ = \frac{|\beta_0 + x_{+1}'\beta|}{\|\beta\|} = \frac{1}{\|\beta\|}.$$

The *margin* of the separating hyperplanes is: $d = \frac{2}{\|\beta\|}$. Note that:

$$y_i(\beta_0 + x_i'\beta) \geq +1, \qquad i = 1, \ldots, n$$

The problem is to find the optimal separating hyperplane, i.e. maximise the margin. That is:

$$\begin{array}{ll} \text{minimise} & \frac{1}{2}\|\beta\|^2 \\ \text{subject to} & y_i(\beta_0 + x_i'\beta) \geq 1 \qquad i = 1, \ldots, n \end{array}$$

This is a convex optimisation problem, hence we have a global minimum. The problem is solved using the Lagrange multiplier technique: set

$$\begin{array}{l} F_p(\beta_0, \beta, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^n \alpha_i \left(y_i(\beta_0 + x_i'\beta) - 1\right) \\ \alpha = (\alpha_1, \ldots, \alpha_n), \quad \alpha_i \geq 0 \end{array}$$

where α is the n-vector of Lagrange coeffficients. The Lagrange method is to find a global minimium for fixed α and then choose the value of α such that the constraint is satisfied. This boils down to:

$$\begin{cases} \frac{\partial F_P}{\partial \beta_0} = -\sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial F_P}{\partial \beta} = \beta - \sum_{i=1}^n \alpha_i y_i x_i = 0 \\ y_i(\beta_0 + x_i'\beta) - 1 \geq 0 \\ \alpha_i \geq 0 \\ \alpha_i(y_i(\beta_0 + x_i'\beta) - 1) = 0 \end{cases}$$

for $i = 1, \ldots, n$.

This may be expressed in the dual form; the minimiser (β_0^*, β) satisfies:

$$\sum_{i=1}^n \alpha_i y_i = 0 \qquad \beta^* = \sum_{i=1}^n \alpha_i y_i x_i$$

and, putting this into so the equation for F_P gives the dual:

$$\begin{aligned} F_D(\alpha) &= \frac{1}{2}\|\beta^*\| - \sum_{i=1}^n \alpha_i \left(y_i(\beta_0^* + x_i'\beta^*) - 1\right) \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i' x_j). \end{aligned}$$

The primal variables have been removed from the problem; F_D is referred to as the dual functional of the optimisation problem. The problem may therefore be expressed as:

$$\begin{array}{ll} \text{maximise} & F_D(\alpha) = \mathbf{1}_n'\alpha - \frac{1}{2}\alpha' H \alpha \\ \text{subject to} & \alpha \geq 0, \quad \alpha' y = 0 \end{array}$$

where $y = (y_1, \ldots, y_n)'$, H is an $n \times n$ matrix with entries: $H_{ij} y_i y_j (x_i' x_j)$. Let $\widehat{\alpha}$ solve the optimisation problem, then

$$\widehat{\beta} = \sum_{i=1}^{n} \widehat{\alpha}_i y_i x_i$$

gives the optimal vector of weights. For $\widehat{\alpha}_i > 0$, we have $y_i(\beta_0^* + x_i'\beta^*) = 1$ and x_i is a support vector; for all observations that are not support vectors, $\widehat{\alpha}_i = 0$. Let $sv \subset \{1, \ldots, n\}$ be the subset of indices that identify support vectors, then any optimal β is:

$$\widehat{\beta} = \sum_{i \in sv} \widehat{\alpha}_i y_i x_i.$$

The primal and dual optimisation problems yield the same solution, the dual is easier to compute. The optimal bias $\widehat{\beta}_0$ is not determined explicitly from the optimisation problem, but is computed from $\alpha_i(y_i(\beta_0 + x_i'\beta) - 1) = 0$ for each support vector and averaging the results.

$$\widehat{\beta}_0 = \frac{1}{|sv|} \sum_{i \in sv} \left(\frac{1 - y_i x_i' \widehat{\beta}}{y_i} \right).$$

Hence the optimal hyperplane is:

$$\widehat{f}(x) = \widehat{\beta}_0 + x'\widehat{\beta} = \widehat{\beta}_0 + \sum_{i \in sv} \widehat{\alpha}_i y_i (x_i' x_i)$$

The classification rule is:

$$C(x) = \text{sign}(\widehat{f}(x))$$

For $j \in sv$,

$$y_j \widehat{f}(x_j) = y_j \widehat{\beta}_0 + \sum_{i \in sv} \widehat{\alpha}_i y_i y_j (x_j' x_i) = 1$$

so that the squared-norm of the weight vector $\widehat{\beta}$ satisfies:

$$\|\widehat{\beta}\|^2 = \sum_{i \in sv} \sum_{j \in sv} \widehat{\alpha}_i \widehat{\alpha}_j y_i y_j (x_i' x_j) = \sum_{j \in sv} \widehat{\alpha}_j y_j \sum_{i \in sv} \widehat{\alpha}_i y_i (x_i' x_j) = \sum_{j \in sv} \widehat{\alpha}_j (1 - y_j \widehat{\beta}_0) = \sum_{j \in sv} \widehat{\alpha}_j$$

7.2 Linearly Non-Separable

Now suppose that observations are noisy, so that they do not necessarily split into two distinct classes; there is some overlap. We introduce the concept of a non-negative *slack variable* ξ_i for each observation (x_i, y_i). Let $\xi = (\xi_1, \ldots, \xi_n)'$. The constraint now becomes:

$$y_i(\beta_0 + x_i'\beta) + \xi_i \geq 1 \qquad i = 1, 2, \ldots, n$$

We now find the optimal hyperplane that controls both the margin $\frac{2}{\|\beta\|}$ and some computationally simple function of the slack variables such as

$$g_\sigma(\xi) = \sum_{j=1}^{n} \xi_j^\sigma.$$

The usual values are either $\sigma = 1$ or $\sigma = 2$. We consider $\sigma = 1$ (the other case can be done as an exercise). The 1-norm soft-margin optimisation problem is to find β_0, β and ξ to:

$$\begin{array}{ll} \text{minimise} & \frac{1}{2}\|\beta\|^2 + C\sum_{j=1}^n \xi_i \\ \text{subject to} & \xi_i \geq 0, \qquad y_i(\beta_0 + x_i'\beta) \geq 1 - \xi_i \qquad i = 1, \ldots, n \end{array}$$

where C is a *cost* parameter, the cost of misclassification. The primal function for the Lagrange multiplier problem is:

$$F_P(\beta_0, \beta, \xi, \alpha, \eta) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i(\beta_0 + x_i'\beta) - (1 - \xi_i)\right) - \sum_{i=1}^{n}\eta_i\xi_i$$

where $\alpha \geq 0$ and $\eta \geq 0$. For fixed α and η, differentiating with respect to β_0, β and ξ gives:

$$\begin{cases} \frac{\partial F_P}{\partial \beta_0} = -\sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial F_P}{\partial \beta} = \beta - \sum_{i=1} \alpha_i y_i x_i = 0 \\ \frac{\partial F_P}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \qquad i = 1, \ldots, n \end{cases}$$

so that

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \qquad \beta^* = \sum_{i=1}^{n}\alpha_i y_i x_i \qquad \eta_i = C - \alpha_i$$

The solution to the optimisation problem is obtained by fixing α and η so that the constraints are satisfied.

The dual functional may be obtained by plugging in appropriately:

$$F_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i' x_j)$$

which is the same as for linear separated.

We now have the Karush-Kuhn-Tucker conditions:

$$\begin{aligned}
y_i(\beta_0 + x_i'\beta) - (1-\xi_i) &\geq 0 \\
\xi_i &\geq 0 \\
\alpha_i &\geq 0 \\
\eta_i &\geq 0 \\
\alpha_i(y_i(\beta_0 + x_i'\beta) - (1-\xi_i)) &= 0 \\
\xi_i(\alpha_i - C) &= 0
\end{aligned}$$

A slack variable ξ_i can be zero if and only if $\alpha_i = C$. The last two equations are used to compute the optimal bias β_0.

As before, the dual problem can be written as: find α to:

$$\begin{array}{ll}
\text{maximise} & F_D(\alpha) = \mathbf{1}_n'\alpha - \frac{1}{2}\alpha' H\alpha \\
\text{subject to} & \alpha' y = 0, \qquad 0 \leq \alpha \leq C\mathbf{1}_n.
\end{array}$$

The feasible region is the intersection of $\alpha' y = 0$ with the box constraint $0 \leq \alpha \leq C\mathbf{1}_n$. As before, if $\widehat{\alpha}$ solves the optimisation problem, then

$$\widehat{\beta} = \sum_{i \in sv} \widehat{\alpha}_i y_i x_i.$$

7.3 NonLinear Support Vector Machines

The observations x_i only enter into the dual problem via their inner products $\langle x_i, x_j \rangle = x_i' x_j$ and this observation is the crux of extending to nonlinear SVMs.

Let $\Phi : \mathbb{R}^r \to \mathcal{H}$ be a linear mapping from observation space to a space known as *feature* space. This may be a Hilbert space, which is what we will use. Let

$$\Phi(x_i) = (\phi_1(x_1), \ldots, \phi_{N(\mathcal{H})}(x_1))$$

where $N(\mathcal{H})$ is the dimension of $\mathcal{H}$. The transformed sample is $(\Phi(x_i), y_i)$, $i = 1, \ldots, n$. If we substitute $\Phi(x_i)$ for x_i, we need the inner products $\langle \Phi(x_i), \Phi(x_j) \rangle$.

7.3.1 The Kernel Trick

We compute these inner products using a non-linear kernel function $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$. We require a kernel to satisfy:

- $K(x, y) = K(y, x)$ (symmetry)
- $|K(x, y)|^2 \leq K(x, x)K(y, y)$ (derived from Cauchy Schwartz inequality)

We would like a *reproducing kernel*; that is, for any function $f \in \mathcal{H}$

$$\langle f(.), K(x,.)\rangle = f(x)$$

Note, if K is a reproducing kernel, then $\langle K(x,.), K(y,.)\rangle = K(x,y)$.

7.3.2 Examples of Kernels

Some standard exaimples are:

- Polynomial of degree d: $K(x,y) = (\langle x,y\rangle + c)^d$
- Gaussian radial: $K(x,y) = \exp\left\{-\frac{1}{2\sigma^2}\|x-y\|^2\right\}$
- Laplace $K(x,y) = \exp\left\{-\frac{1}{\sigma}\|x-y\|\right\}$
- Thin-plate spline $K(x,y) = \left(\frac{\|x-y\|}{\sigma}\right)^2 \log(\frac{\|x-y\|}{\sigma})$
- Sigmoid $K(x,y) = \tanh(a\langle x,y\rangle + b)$

For example, consider $r = 2$ and $d = 2$, $x = (x_1, x_2)'$, $y = (y_1, y_2)'$ and

$$K(x,y) = (\langle x,y\rangle + c)^2 = (x_1y_1 + x_2y_2 + c)^2 = \langle \Phi(x), \Phi(y)\rangle.$$

Here

$$\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1, x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c)'$$

The function $\Phi(x)$ consists of six features and $\mathcal{H} = \mathbb{R}^6$.

Let K be a kernel and suppose that the observations of $\mathcal{L}$ are linearly separable in the *feature* space corresponding to kernel K. Then the dual optimisation problem is as before, but with the matrix H:

$$H_{ij} = y_iy_jK(x_i,x_j) = y_iy_jK_{ij}.$$

Since K is a kernel, the matrix K defined by entries K_{ij} is non-negative definite so that the optimisation problem can be solved as before.

The non-separable setting (for the dual problem) also follows through as before.

Grid search for parameters A *reproducing kernel Hilbert space* is a Hilbert space such that there is a Kernel K satisfying $f(x) = \langle f, K_x \rangle$. Consider the Gaussian reproducing kernel. We need to determine two parameters: C, the cost of violating the constraints and the parameter $\gamma = \frac{1}{\sigma^2}$. The parameter C for the box constraints is usually chosen by searching through a wide range of possible values using cross validation (usually 10-fold) on $\mathcal{L}$ An initial grid rather crude grid of possible values for γ, say 0.00001, 0.001, 0.01, 0.1, 1 can be used to get a 'ball park' figure and then refined. In this way, we make a two-way grid for (C, γ).

7.3.3 SVM as a Regularisation Method

Let $f \in \mathcal{H}_K$, the reproducing Hilbert space associated with K. Let $\|f\|^2_{\mathcal{H}_K}$ denote the squared norm of f in $\mathcal{H}_K$. We consider the *hinge loss function*:

$$L = (1 - y_i f(x_i))_+$$

Note that $L = 0$ if $y_i f(x_i) \geq 1$. That is, $L = 0$ for $y_i = 1$ and $f(x_i) \geq 1$ or $y_i = -1$ and $f(x_i) < -1$ (the situations where $f(x_i)$ gives the correct classification).

Consider the problem of finding $f \in \mathcal{H}_K$ to:

$$\text{minimise} \qquad \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(x_i))_+ + \lambda \|f\|^2_{\mathcal{H}_K}$$

where $\lambda > 0$. The first term measures the distance of the data from separability, while the second penalises overfitting. THe tuning parameter λ balances the trade-off.

The optimisation criterion is not differentiable, but we can consider it as follows:

$$f(.) = f^{\|}(.) + f^{\perp}(.) = \sum_{i=1}^{n} \alpha_i K(x_i, .) + f^{\perp}(.)$$

where $f^{\|}$ denotes the projection of f onto the subspace of $\mathcal{H}_K$ generated by $(K(x_1, .), \ldots, K(x_n, .))$ and $f^{\perp}$ is the part perpendicular to this; i.e. $\langle f^{\perp}(.), K(x_i, .)\rangle = 0$ for $i = 1, \ldots, n$. Since

$$f(x_i) = \langle f(.), K(x_i, .)\rangle = \langle f^{\|}(.), K(x_i, .)\rangle + \langle f^{\perp}(.), K(x_i, .)\rangle$$

and the second term vanishes, we have:

$$f^{\|}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

independent of $f^{\perp}$ and hence

$$\|f\|^2_{\mathcal{H}_K} = \|\sum_i \alpha_i \alpha_i K(x_i, .)\|^2_{\mathcal{H}_K} + \|f^\perp\|^2_{\mathcal{H}_K} \geq \|\sum_i \alpha_i K(x_i, .)\|^2_{\mathcal{H}_K}$$

with equality if and only if $f^\perp = 0$.

Therefore

$$\|f^\|\|^2_{\mathcal{H}_K} = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) = \|\beta\|^2$$

where $\beta = \sum_{i=1}^n \alpha_i \Phi(x_i)$.

If the space $\mathcal{H}_K$ consists of linear functions of the form $f(x) = \beta_0 + \Phi(x)'\beta$, with $\|f\|^2_{\mathcal{H}_k} = \|\beta\|^2$, then the problem of finding f is equivalent ot finding β_0, β which solves:

$$\text{minimise} \qquad \frac{1}{n}\sum_{i=1}^n (1 - y_i(\beta_0 + \Phi(x_i)'\beta))_+ + \lambda\|\beta\|^2$$

so that the problem with non-differentiability due to the hinge loss function can be reformulated in terms of the 1-norm soft-margin optimisation problem.