10.3 Lagrangian Formulation of the SVM

Having introduced some elements of statistical learning and demonstrated the potential of SVMs for company rating we can now give a Lagrangian formulation of an SVM for the linear classification problem and generalize this approach to a nonlinear case.

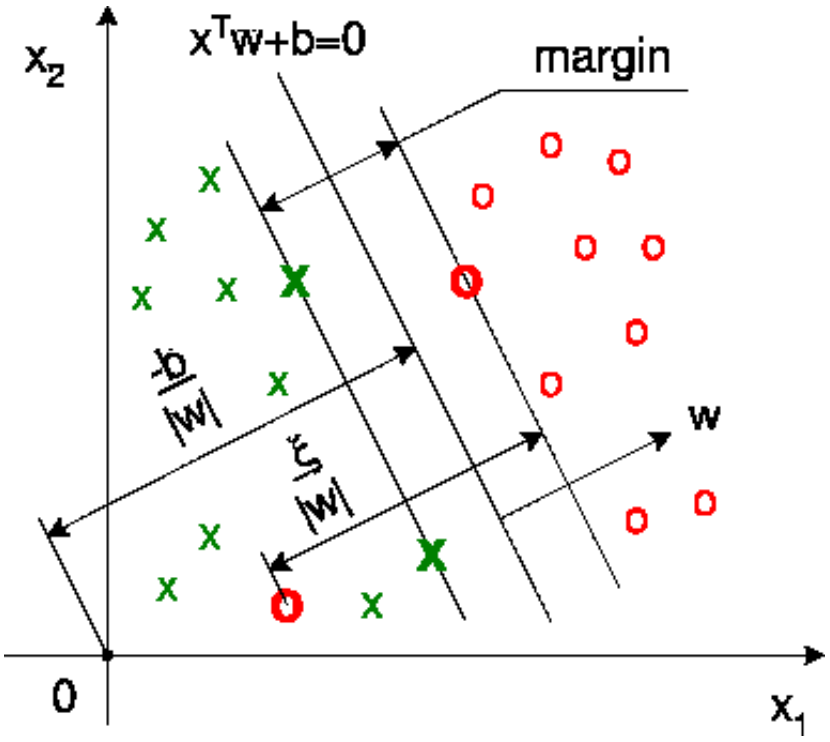

Figure 10.3: The separating hyperplane $x^\top w + b = 0$ and the margin in a non-separable case.

In the linear case the following inequalities hold for all n points of the training set:

$$x_i^\top w + b \geq 1 - \xi_i \text{ for } y_i = 1,$$

$$x_i^\top w + b \leq -1 + \xi_i \text{ for } y_i = -1,$$

$$\xi_i \geq 0,$$

which can be combined into two constraints:

$$y_i(x_i^\top w + b) \geq 1 - \xi_i \tag{10.9}$$

$$\xi_i \geq 0. \tag{10.10}$$

The basic idea of the SVM classification is to find such a separating hyperplane that corresponds to the

largest possible margin between the points of different classes, see Figure 10.3. Some penalty for misclassification must also be introduced. The classification error ξ_i is related to the distance from a misclassified point x_i to the canonical hyperplane bounding its class. If $\xi_i > 0$, an error in separating the two sets occurs. The objective function corresponding to penalized margin maximization is formulated as:

$$\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right)^{\upsilon}, \qquad (10.11)$$

where the parameter C characterizes the generalization ability of the machine and $\upsilon \geq 1$ is a positive integer controlling the sensitivity of the machine to outliers. The conditional minimization of the objective function with constraint (10.9) and (10.10) provides the highest possible margin in the case when classification errors are inevitable due to the linearity of the separating hyperplane. Under such a formulation the problem is convex. One can show that margin maximization reduces the VC dimension.

The Lagrange functional for the primal problem for $\upsilon = 1$ is:

$$L_P = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i\left(x_i^{\top}w + b\right) - 1 + \xi_i\} - \sum_{i=1}^{n}\mu_i\xi_i, \qquad (10.12)$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are Lagrange multipliers. The primal problem is formulated as:

$$\min_{w_k, b\xi_i}\max_{\alpha_i} L_P.$$

After substituting the Karush-Kuhn-Tucker conditions (Gale et al.; 1951) into the primal Lagrangian, we derive the dual Lagrangian as:

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^{\top}x_j, \qquad (10.13)$$

and the dual problem is posed as:

$$\max_{\alpha_i} L_D,$$

subject to:

$$0 \leq \alpha_i \leq C,$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0.$$

Those points i for which the equation $y_i(x_i^\top w + b) \leq 1$ holds are called support vectors. After training the support vector machine and deriving Lagrange multipliers (they are equal to 0 for non-support vectors) one can classify a company described by the vector of parameters x using the classification rule:

$$g(x) = \text{sign}\left(x^\top w + b\right), \qquad (10.14)$$

where $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ and $b = \frac{1}{2}\left(x_{+1} + x_{-1}\right) w$. x_{+1} and x_{-1} are two support vectors belonging to different classes for which $y(x^\top w + b) = 1$. The value of the classification function (the score of a company) can be computed as

$$f(x) = x^\top w + b. \qquad (10.15)$$

Each value of $f(x)$ uniquely corresponds to a default probability (PD).

The SVMs can also be easily generalized to the nonlinear case. It is worth noting that all the training vectors appear in the dual Lagrangian formulation only as scalar products. This means that we can apply kernels to transform all the data into a high dimensional Hilbert feature space and use linear algorithms there:

$$\Psi : \mathbb{R}^d \mapsto \mathbb{H}. \qquad (10.16)$$

If a kernel function K exists such that $K(x_i, x_j) = \Psi(x_i)^\top \Psi(x_j)$, then it can be used without knowing the transformation Ψ explicitly. A necessary and sufficient condition for a symmetric function $K(x_i, x_j)$ to be a kernel is given by Mercer's (1909) theorem. It requires positive definiteness, i.e. for any data set $x_1, ..., x_n$ and any real numbers $\lambda_1, ..., \lambda_n$ the function K must satisfy

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0. \qquad (10.17)$$

Some examples of kernel functions are:

- $K(x_i, x_j) = e^{-\|x_i - x_j\|/2\sigma^2}$ - the isotropic Gaussian kernel;
- $K(x_i, x_j) = e^{-(x_i - x_j)^\top r^{-2}\Sigma^{-1}(x_i - x_j)/2}$ - the stationary Gaussian kernel with an anisotropic

radial basis; we will apply this kernel in our study taking Σ equal to the variance matrix of the training set; r is a constant;

- $K(x_i, x_j) = (x_i^\top x_j + 1)^P$ - the polynomial kernel;
- $K(x_i, x_j) = \tanh(kx_i^\top x_j - \delta)$ - the hyperbolic tangent kernel.
