Tik-61.231 Principles of Pattern Recognition

Answers to exercise 5: 23.10.2000

1. Let's denote the function slightly differently for the proof; let $d = l - 1$ and $O(N,l) = C(N,d)$, where

$$C(N,d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k}$$

In my opinion this is more illustrative, as we are actually using the dimension of the space d instead of constantly using l to denote $l-1$ dimensional space. $C(N,d)$ tells us the number of groupings that can be formed by d-dimensional hyperplanes to separate the N points into two classes.

First we need to prove that $C(N+1,d) = C(N,d) + C(N,d-1)$.

Let $C(N,d)$ be a separable set of dichotomies X. Let's take a new point x_{N+1} so that $X \cup \{x_{N+1}\}$ is in the general position (well distributed). Let there be a vector w that divides X into two sets $X = \{X^+, X^-\}$ so that $w \cdot x > t \Rightarrow x \in X^+$ and $w \cdot x < t \Rightarrow x \in X^-$, where t is a scalar.

If $\{X^+, X^-\}$ is separable, must also either $\{X^+ \cup \{x_{N+1}\}, X^-\}$ or $\{X^+, X^- \cup \{x_{N+1}\}\}$ be separable. However, they both are separable if and only if $\exists w$ that is a vector that separates $\{X^+, X^-\}$ in a $(d-1)$ dimensional space and is orthogonal to x_{N+1}

To prove the prior statement regarding w, let the set ova separating vectors $W = \{w : w \cdot x > t, x \in X^+; w \cdot x < t, x \in X^-\}$. The set $\{X^+ \cup \{x_{N+1}\}, X^-\}$ is homogeneously separable if and only if $\exists w \in W$ so that $w \cdot x_{N+1} > t$, and equivalently $\{X^+, X^- \cup \{x_{N+1}\}\}$ is homogeneously separable if and only if $\exists w \in W$ so that $w \cdot x_{N+1} < t$ Let the sets be linearly separable with w_1 and w_2, respectively. Then $w^* = (-w_2 \cdot x_{N+1})w_1 + (w_1 \cdot x_{N+1})w_2$ separates $\{X^+, X^-\}$ by the hyperplane $\{x : w^* \cdot x = t\}$ passing through x_{N+1}. Conversely, if the sets $\{X^+, X^-\}$ are homogeneously linearly separable by a hyperplane containing x_{N+1}, then $\exists w^* \in W$ so that $w^* \cdot x = t$. Since W is an open set, $\exists \epsilon > 0$ so that $w * + \epsilon x_{N+1}$ and $w * - \epsilon x_{N+1}$ are in W. Hence $\{X^+ \cup \{x_{N+1}\}, X^-\}$ and $\{X^+, X^- \cup \{x_{N+1}\}\}$ are homogeneously linearly separable by $w * + \epsilon x_{N+1}$ and $w * - \epsilon x_{N+1}$, respectively.

So the set can be separated if and only if $\exists w$ so that the projection onto a $(d-1)$ dimensional subspace is separable. By the induction hypothesis there are $C(N, d-1)$ such separable dichotomies. Hence,

$$C(N+1,d) = C(N,d) + C(N,d-1)$$

By repeatedly applying of this to the terms on the right we obtain

$$C(N,d) = \sum_{k=0}^{N-1} \binom{N-1}{k} C(1, d-k)$$

Now, as it is obvious that one point can be separated in two ways if the dimension is greater or equal to 1 and no separation can be made when the dimension is below one, or

$$C(1,m) = \begin{cases} 2, m \geq 1 \\ 0, m < 1 \end{cases}$$

The original theorem follows by separating the part of the sum where $d - k < 1 \Leftrightarrow k > d - 1$:

$$C(N,d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k} + 0 \cdot \sum_{k=d}^{N-1} \binom{N-1}{k} = 2\sum_{k=0}^{d-1} \binom{N-1}{k} \Leftrightarrow O(n,l) = 2\sum_{k=0}^{l} \binom{N-1}{k}$$

2. The SVM optimal hyperplane separates the space so that

$$\omega^T x_i + \omega_0 \geq +1, \text{ if } x_i \in \omega_1$$
$$\omega^T x_i + \omega_0 < -1, \text{ if } x_i \in \omega_2$$

Let ω_1 be on the positive side of the optimal hyperplane and ω_2 on the negative side, and d_+ and d_- be the distances from the optimal hyperplane and the nearest point in classes ω_1 and ω_2, respectively. Let $g(x) = \omega^T x_i + \omega_0$ be the distance from the optimal hyperplane ω. It can also be stated that

$$x = x_p + r\frac{\omega}{||\omega||}$$

where x_p is the projection of x onto the optimal hyperplane and r is the distance from the hyperplane. Since $g(x_p) = 0$ by definition (the point x_p lies on the optimal hyperplane),

$$g(x) = \omega^T x + \omega_0 = r||w|| \Leftrightarrow r = \frac{g(x)}{||\omega||}$$

Thus the algebraic distance for the support vectors is

$$r = \frac{g(x)}{||\omega||} = \begin{cases} \frac{1}{||\omega||} = d_+, \text{ when } x \text{ is the nearest point of } \omega_1 \\ -\frac{1}{||\omega||} = d_-, \text{ when } x \text{ is the nearest point of } \omega_2 \end{cases}$$

Here the negative sign denotes being on the negative side of the hyperplane. Thus the distance between the two points is $\frac{2}{||\omega||}$.

3. The main idea behind finding the optimal SVM decision hyperplane is to maximize the marginal $\frac{2}{||\omega||}$. In the basic, separable case this is done through taking positive (because of the form $\ldots \geq 0$) Lagrange multipliers $\alpha_i, i = 1, \ldots, l$, where l is the number of points, for each inequity

$$y_i(\omega^T x_i + \omega_0) - 1 \geq 0$$

where y_i denotes class membership, $y_i = 1$ if $x_i \in \omega_1$ and $y_i = -1$ if $x_i \in \omega_2$. Thus the objective function to minimize is

$$L_P = \frac{||\omega||^2}{2} - \sum_{i=1}^{l} \alpha_i y_i(\omega^T x_i + \omega_0) + \sum_{i=1}^{l} \alpha_i$$

The objective is to minimize L_P with respect to ω and ω_0 and simultaneously require that the derivatives of L_P with respect to all α_i vanish, all subject to the constraints $\alpha_i \geq 0$. This is a convex quadratic programming problem, since both the objective function is convex and the points satisfying the constraints form a convex set. This means that it is equivalently possible to solve the dual problem, maximize L_P subject to the constraints that the gradient of L_P with respect to ω and ω_0 vanish and again all $\alpha_i \geq 0$. Requiring the gradient of L_P to vanish with respect to ω and ω_0 gives the additional constraints:

$$\frac{\delta L_P}{\delta \omega} = \omega - \sum_i \alpha_i y_i x_i = 0 \Rightarrow \omega = \sum_i \alpha_i y_i x_i$$
$$\frac{\delta L_P}{\delta \omega_0} = -\sum_i \alpha_i y_i = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

By substituting these conditions into the equation for L_P we obtain

$$L_D = \tfrac{1}{2}(\textstyle\sum_i \alpha_i y_i x_i)^2 + -(\sum_i \alpha_i y_i x_i)^2 - 0 * b + \sum_i \alpha_i$$
$$= \textstyle\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

Both formulations produce the same result. The latter formulation is called the Wolfe dual.

For the non-separable case the basic algorithm provides no feasible solution as the objective function grows arbitrarily large. In order to handle the non-separable case an additional cost must be introduced to loosen the original constraints when necessary. This can be done by introducing positive slack variables $\xi_i \geq 0, i = 1, \ldots, l$ into the constraints, which then become

$$\omega^T x_i + \omega_0 \geq 1 - \xi_i, \text{ if } x \in \omega_1$$
$$\omega^T x_i + \omega_0 \leq -1 - \xi_i, \text{ if } x \in \omega_2$$

Thus for an error to occur $\xi_i > 1$, so $\sum_i \xi_i$ is an upper bound for training errors. The costs can be added to the objective function so that the objective function becomes $\frac{||\omega||^2}{2} + C(\sum_i \xi_i)^k$, where C is a cost parameter to be freely chosen (larger C is equivalent to a larger cost for making a mistake). It can be seen that when $k = 1$ L_P becomes

$$L_P = \frac{||w||^2}{2} C \sum_i \xi_i - \sum_i \alpha_i [y_i(x_i \cdot \omega + \omega_0) - 1 + \xi_i] - \sum_i \mu_i \xi_i$$

In this case neither ξ_i nor their Lagrange multipliers μ_i appear in the Wolfe dual L_D, which can be seen by requiring the gradient of L_P to vanish with respect to ω, ω_0 and all ξ_i:

$$\frac{\delta L_P}{\delta \omega} = \omega - \sum_i \alpha_i y_i x_i = 0$$
$$\frac{\delta L_P}{\delta \omega_0} = -\sum_i \alpha_i y_i = 0$$
$$\frac{\delta L_P}{\delta \xi_i} = C - \alpha_i - \mu_i = 0$$

By substituting these into the equation for L_P we get the Wolfe dual for the non-separable case

$$\begin{aligned} L_D &= \tfrac{1}{2}(\textstyle\sum_i \alpha_i y_i x_i)^2 + \sum_i (\alpha_i + \mu_i)\xi_i - (\sum_i \alpha_i y_i x_i)^2 - 0 * b + \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \mu_i \xi_i \\ &= -\tfrac{1}{2}(\textstyle\sum_i \alpha_i y_i x_i)^2 + \sum_i \alpha_i \\ &= \textstyle\sum_i \alpha_i - \tfrac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \end{aligned}$$

So the problem is to maximize L_D subject to the constraints $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$, and $w = \sum_{i=1}^{N_s} \alpha_i y_i x_i$, where N_s is the amount of support vectors. As we can see, the form of L_D is actually identical to that of the separable case. The only difference is in the costraints.

So, in the situation where we have the points $x_1 = [1\ \ 1]^T \in \omega_1$, $x_2 = [2\ \ 1]^T \in \omega_2$, $x_3 = [3\ \ 2]^T \in \omega_1$ and $x_4 = [2\ \ 3]^T \in \omega_2$, the dual problem L_D can be written as

$$\begin{aligned} L_D &= \textstyle\sum_i \alpha_i - \tfrac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \tfrac{1}{2}(2\alpha_1^2 + 7\alpha_2^2 + 13\alpha_3^2 + 13\alpha_4^2 - 6\alpha_1\alpha_2 + 10\alpha_1\alpha_3 \\ &\quad -10\alpha_1\alpha_4 - 14\alpha_2\alpha_3 + 14\alpha_2\alpha 4 - 23\alpha_3\alpha_4) \end{aligned}$$

and the constraints as

$$\alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0$$
$$0 \leq \alpha_i \leq C \ \ \forall i$$

where C is the cost parameter to be chosen.