We would like to find w and w_0 such that

$$w^T x^t + w_0 \geq +1 \quad \text{for} \quad y^t = +1$$
$$w^T x^t + w_0 \leq -1 \quad \text{for} \quad y^t = -1$$

which can be rewritten as

$$y^t (w^T x^t + w_0) \geq +1$$

↳ Note that this is not 0

Not only do we want the instances to be on the right side of the hyperplane, but we also want them some distance away for better generalization.

Margin: The distance from the hyperplane to the instances closest to it on either side is called margin, which we want to maximize for better generalization.

The distance of x^t to the discriminant is

$$\frac{|w^T x^t + w_0|}{\|w\|} = \frac{y^t (w^T x^t + w_0)}{\|w\|}$$

$$\frac{y^t (w^T x^t + w_0)}{\|w\|} \geq \rho, \quad \forall t$$

We would like this to be at least some value of ρ

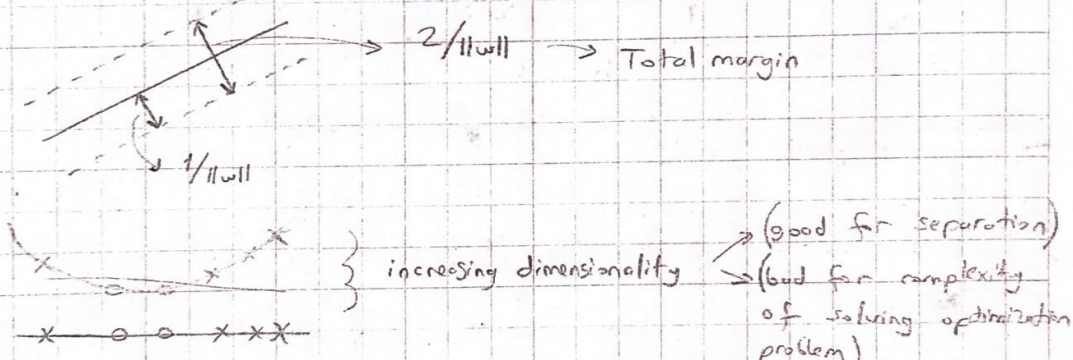
We would like to maximize ρ but then on an infinite number of solutions that we can get by scaling w and for a unique solution we fix $\rho \|w\| = 1$, thus to maximize the margin, we minimize $\|w\|$.

(See Cortes & Vapnik 1995, Vapnik 1995)

$$\min \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y^t (w^T x^t + w_0) \geq 1 \qquad (*)$$

This is a standard quadratic optimization problem whose complexity depends on d, and it can be solved directly to find w and w_0.
↳ (# of features of instances)

On both sides of the hyperplane, there will be instances that are $1/\|w\|$ away from the hyperplane.



To get the new formulation (an optimization problem of a form whose complexity depends on N) we first rewrite equation (*)

- svm light
- libSVM

} batch SVM software

- LASVM → by Antoine Bordes, Seyda Ertekin, Leon Bottou
online SVM softwares

as an unconstrained problem using Lagrange multipliers α^t

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{t=1}^{N} \alpha^t \left[y^t (w^T x^t + w_0) - 1 \right]$$

$$= \frac{1}{2}\|w\|^2 - \sum_t \alpha^t y^t (w^T x^t + w_0) + \sum_t \alpha^t \quad (**)$$

This should be minimized wrt w, w_0 and maximized wrt $\alpha^t \geq 0$
The saddle point gives the solution.

⇒ Convex quadratic optimization problem, because the main term is convex and the linear constraints are also convex.

The dual problem is to maximize L_p with respect to α^t, subject to the constraints that the gradient of L_p wrt w and w_0 are 0 and also that $\alpha^T \geq 0$

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_t \alpha^t y^t x^t \quad (+)$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_t \alpha^t y^t = 0$$

Plug them into equation (*) we get the dual

$$L_d = \frac{1}{2}(w^T w) - w^T \sum_t \alpha^t y^t x^t - w_0 \sum_t \alpha^t y^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}(w^T w) + \sum_t \alpha^t$$

$$= \frac{1}{2} \sum \sum \alpha^t \alpha^s y^t y^s (x^t)^T x^s + \sum_t \alpha^t$$

which we maximize wrt α^t only subject to constraints

$$\sum_t \alpha^t y^t = 0 \quad \text{and} \quad \alpha^T \geq 0 \quad \forall t$$

The size of the dual depends on N (# of training examples). The upper bound for space complexity is $O(N^2)$, for time complexity is $O(N^3)$.

Once we solve for α^t, we see that though there are N of them, most vanish with $\alpha^t = 0$ and only a small percentage have $\alpha^t > 0$

The set of x_t whose $\alpha^t > 0$ are the <u>support vectors</u>.

⊕ w is written as the weighted sum of those training instances that are selected as the support vectors.

$$y^t(w^T x^t + w_0) = 1 \quad \text{(examples that lie on the margin)}$$

$$w_0 = y^t - w^T x^t \quad \text{(for numerical stability, it is advised that this be done for all support vectors and average be taken).}$$

The Non-separable Case: Soft Margin Hyperplane

If there is no hyperplane which linearly separates data, we look for the one which incurs the least error. We define slack variables $\xi^t \geq 0$ which store the deviation from the margin.

There are 2 types of deviation

1) An instance may lie on the wrong side of the hyperplane and be misclassified.
2) or it may be on the right side but may lie in the margin (not sufficiently away from the hyperplane)

Relaxing the equation, we get

$$y^t(w^T x^t + w_0) \geq 1 - \xi^t$$

$0 < \xi^t < 1$: x^t is correctly classified but in the margin

$\xi \geq 1$: x^t is misclassified

The # of misclassifications is $\#\{\xi^t > 1\}$
The # of nonseparable points is $\#\{\xi^t > 0\}$

We define a soft error as $\sum_t \xi^t$

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_t \xi^t$$

→ Penalty factor for misclassifieds. It determines the complexity of the model.

$$L_p = \frac{1}{2}\|w^2\| + C\sum_t \xi^t - \sum \alpha^t [y^t(w^T x^t + w_0) - 1 + \xi^t] - \sum_t \mu^t \xi^t$$

new Lagrange multiplier to guarantee the positivity of ξ^t

When we take the derivatives wrt the parameters and set them to 0, we get

$$\frac{\partial L_p}{\partial w} = w - \sum_t \alpha^t y^t x^t = 0 \Rightarrow w = \sum_t \alpha^t y^t x^t$$

$$\frac{\partial L_p}{\partial w_0} = \sum \alpha^t y^t = 0$$

$$\frac{\partial L_p}{\partial \xi^t} = C - \alpha^t - \mu^t = 0$$

Since $\mu^t \geq 0$, this last implies that $0 \leq \alpha^t \leq C$

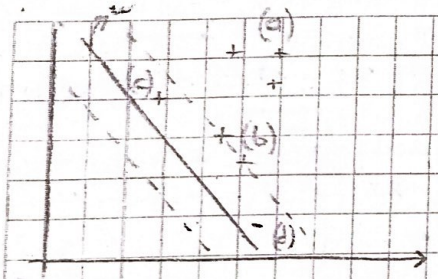
Plug these into equation ⊛ we get the dual that we maximize wrt α^t

$$L_d = \sum \alpha^t - \frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s y^t y^s (x^t)^T x^s \quad \text{subject to}$$

$$\sum_t \alpha^t y^t = 0 \quad \text{and} \quad 0 \leq \alpha^t \leq C \;, \; \forall t$$

T.C.
Ulaştırma Denizcilik ve Haberleşme Bakanlığı

__/__/20__



Cases b, c and d become support vectors.

a) The instances on the correct side and far away from the margin

$$y^t f(x^t) > 1 \qquad \xi^t = 0 \qquad \alpha^t = 0$$

b) It is on the right side and on the margin

$$y^t f(x^t) = 1 \qquad \xi^t = 0 \qquad \alpha^t \leq C$$

c) The instance is on the right side but it's in the margin and not sufficiently away

$$\alpha^t = C$$

$$\xi^t = 1 - f(x^t) \quad , \quad 0 < \xi < 1 \quad , \quad y^t f(x^t) < 1$$

d) The instance is on the wrong side. This is misclassification.

$$y^t f(x^t) < 0 \qquad \xi^t = 1 + f(x^t) > 1 \qquad \alpha^t = C$$

$$L_p = \frac{1}{2} \|w\|^2 + C \sum_t \xi^t$$

↳ Penalty for misclassifications