An interesting dual - the S.V.M. ①

Problem: given N d-dim data points with 2 classes, $(\underline{x}_i, y_i)$
$\llcorner$ 1 or -1

- wish to ~~classify~~ learn classifier for new points
- assume there is relatively little training data (otherwise we'd do something else)
- Two cases:
  - linearly separable
    = there is a hyperplane separating +ve or -ve examples
  - Not linearly separable
    = there isn't

linearly separable:

- we want a linear classifier (so $\text{sign}(a^T x + b)$ gives class)
- it must get training data right so $y_i(\underline{a}^T \underline{x}_i + b) \geq 1$, for each i
- if data is separable, we expect multiple $\underline{a}$, b will work.
- Notice that if $\underline{a}, b$ are feasible, then so are $\lambda \underline{a}, \lambda b$, for $\lambda > 0$
- Notice that the distance from x_i to the hyperplane

$$a^T x_i + b$$

is

$$\frac{|a^T x_i + b|}{\|a\|}$$

$\sqrt{a^T a}$; length of $\underline{a}$

This suggests choice of hyperplane s.t all points are as far away as possible

$$\min \quad \frac{a^T a}{2}$$

$$\text{st} \quad y_i(a^T x_i + b) \geq 1$$

<u>Dual:</u>

$$L_p(\underline{a}, b, \lambda) = \frac{\underline{a}^T \underline{a}}{2} - \sum_i \lambda_i \left\{ \left[y_i(\underline{a}^T \underline{x}_i + b) - 1 \right] \right\}$$

$$\nabla_{\underline{a}} L = 0 = a - \sum_i \lambda_i [y_i \underline{x}_i]$$

$$\nabla_b L = 0 = -\sum_i \lambda_i y_i$$

Substitute these into lagrangian (notice coeff of b is $\sum_i \lambda_i y_i = 0$) to get

$$\underline{\max} \quad \sum_i \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j \left[y_i y_j (x_i^T x_j) \right]$$

$$\underline{st}: \quad \sum_i \lambda_i y_i = 0$$

$$\lambda_i \geq 0 \qquad \text{(inequality constraints)}$$

This problem is well studied

- consider constraints.
- complementarity means that either

$$y_i(\underline{a}^T x_i + b) = 1 \quad \underline{\text{OR}} \quad \lambda_i = 0$$

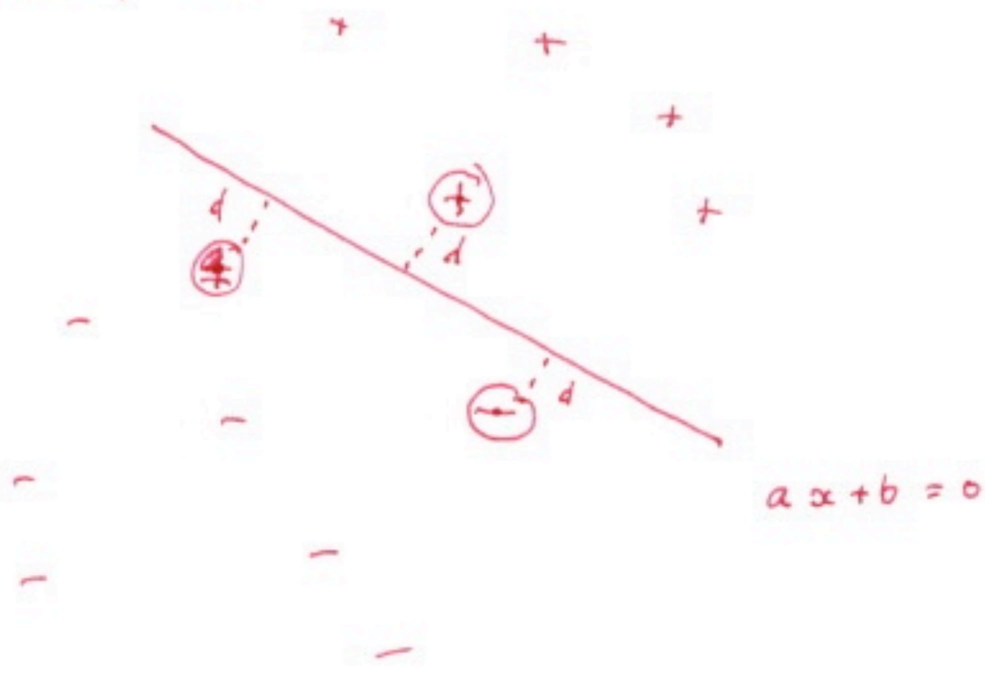
- Recall remarks on distance all points where

$$y_i(\underline{a}^T \underline{x}_i + b) = 1$$

are <u>the same distance</u> from hyperplane $\underline{a}^T \underline{x} + b = 0$ (but may be on different sides!)

- We do not expect many of these (geometry).

e.g. in 2D, expect 3



in dD, $d+1$

This means almost every λ_i is zero!

Obsolete algorithmic procedure:

- find the non-zero λ_i
- $\underline{a}$ is then a weighted sum of those examples
- b follows

Now assume problem isn't linearly separable.

- this means

$$\underbrace{y_i(\underline{a}^T \underline{x}_i + b) \geq 1}$$

↳ feasible set – it's empty.

- introduce slacks

$$y_i(\underline{a}^T \underline{x}_i + b) + \xi_i \geq 1$$
$$\xi_i \geq 0$$

and penalize these slacks, to get

$$\min \quad \frac{\underline{a}^T a}{2} + c \sum_i \xi_i$$

$$\text{st} \quad y_i(\underline{a}^T \underline{x}_i + b) + \xi_i \geq 1$$
$$\xi_i \geq 0$$

Dual :

$$\mathcal{L} = \frac{\underline{a}^T\underline{a}}{2} + C\sum_i \xi_i - \sum_i \lambda_i \left[y_i(\underline{a}^T\underline{x}_i + b) - 1 + \xi_i\right] - \sum_i \mu_i \xi_i$$

$$\nabla_a L = \underline{a} - \sum_i \lambda_i y_i \underline{x}_i = 0$$

$$\nabla_b L = 0 = -\sum_i \lambda_i y_i$$

$$\nabla \xi_i = C - \lambda_i - \mu_i = 0$$

Subst back

$$\text{Dual} = \sum_i \lambda_i - \frac{1}{2}\sum_{ij} y_i y_j \underline{x}_i^T \underline{x}_j \lambda_i \lambda_j$$

s.t. $\sum_i \lambda_i y_i = 0$

$0 \leq \lambda_i \leq C$

neat trick w/ μ_i

This problem is well studied:

- Notice at soln, if x_i is on the right side of boundary (i.e. correctly classified), $\mu_i > 0$ (complementarity!)
- It follows that for points far enough on the right side $\lambda_i = 0$, so $\mu_i = C$
  complementarity
- So if most datapoints are correctly classified with high margin, most $\lambda_i = 0$

Obsolete algorithmic thread: find the non-zero λ_i.

Notice we can interpret ξ_i as a loss

$$\xi_i = \underbrace{\max}_{\text{hinge loss}}(0,\ 1 - y_i(\underline{a}^T\underline{x}_i + b))$$

$h(\underline{a}, b; x_i, y_i)$.

then we have

$$\min_{\underline{a}, b} \quad \frac{\underline{a}^T\underline{a}}{2} + C\sum_i h(\underline{a}, b; \underline{x}_i, y_i)$$

(without constraints!).

Cleaner to write

$$\underbrace{\frac{1}{N}\sum_i h(\underline{a}, b; \underline{x}_i, y_i)}_{\text{empirical hinge loss}} + \lambda \underbrace{\frac{\underline{a}^T\underline{a}}{2}}_{\text{regularizer}}$$