$\boxed{\text{2020-09-28}}$

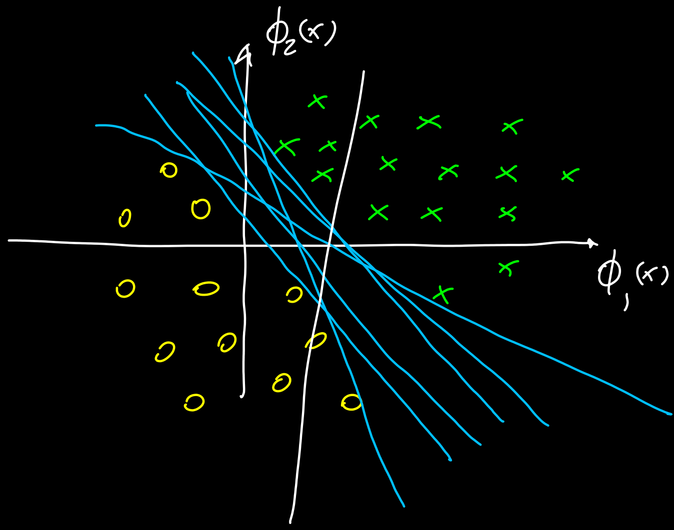
PRML Ch. 7

- sparse Kernel methods
- maximum margin classifier
- support vector machines
  - linearly separable data
  - overlapping data

Maximum margin classifier

two-class problem

$$y(x) = w^T \phi(x) + b$$



training data

$\{x_1 \dots x_N \; ; \; x_n \in \mathbb{R}^D\}$

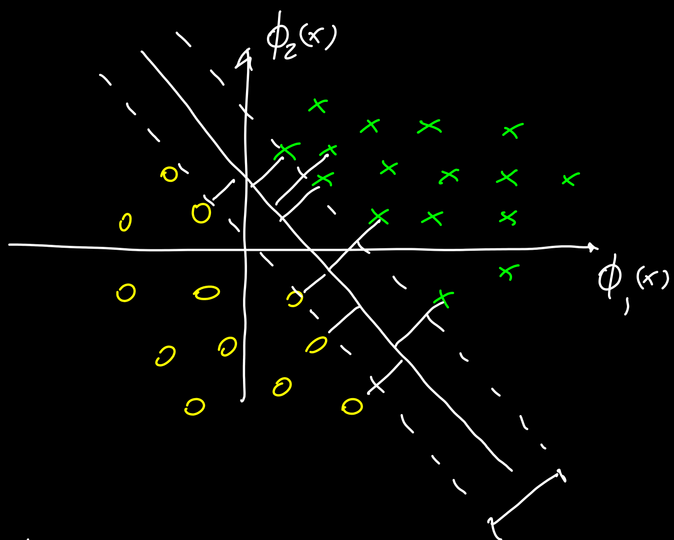
$\{t_1 \dots t_N \; ; \; t_n \in \{-1, +1\}\}$

$$\exists \; w, b \quad \text{s.t.} \quad \left.\begin{aligned} y(x_n) > 0 &\iff t_n = +1 \\ y(x_n) < 0 &\iff t_n = -1 \end{aligned}\right\} \Rightarrow y(x_n) \cdot t_n > 0 \quad \forall n$$

Goal: minimize generalization error

We only have the training data (do not know the underlying distribution)

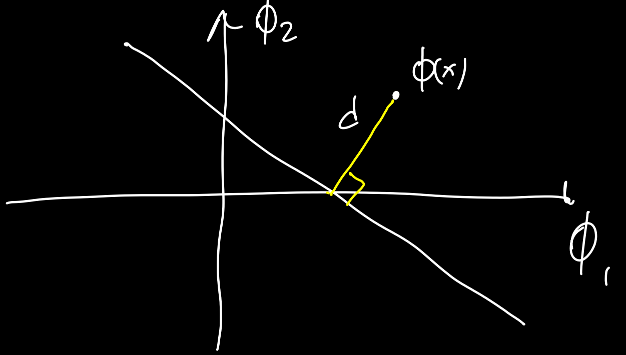
Heuristic solution: maximize margin



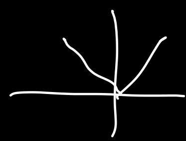
w, b

Recall:

– if x is on decision boundary $\Rightarrow y(x) = 0$



$$d = \frac{|y(x)|}{\|w\|}$$



- we are only interested in no error $\Rightarrow t_n y(x_n) > 0$

$$d = \frac{t_n \, y(x_n)}{\|w\|} = \frac{t_n \left(w^T \phi(x_n) + b \right)}{\|w\|} \qquad \forall n \in [1, N]$$

Optimization:

$$\underset{w,b}{\arg\max} \left\{ \underbrace{\frac{1}{\|w\|}}_{\text{independent of } n} \min_n \left[t_n \left(w^T \phi(x_n) + b \right) \right] \right\}$$

hard to solve directly

$$w, b \rightarrow Kw, Kb$$

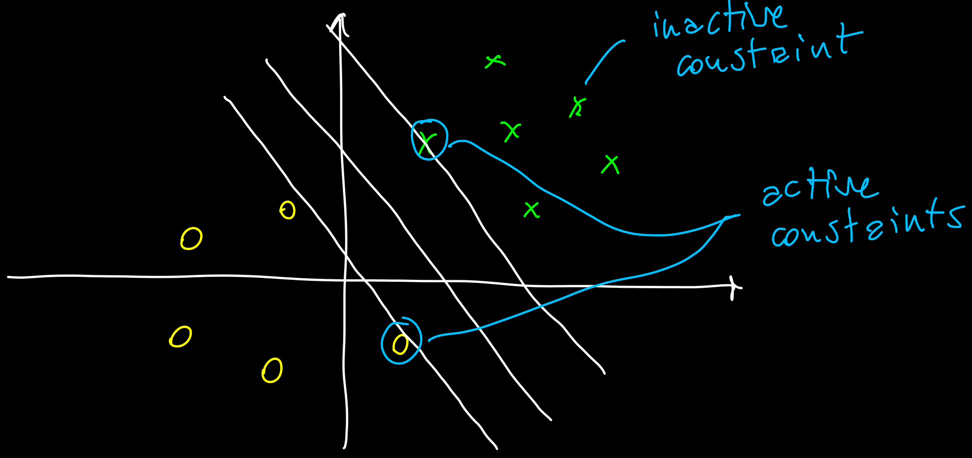
$$\frac{t_n \left(\cancel{K} w^T \phi(x) + \cancel{K} b \right)}{\|\cancel{K} w\|} = \frac{t_n \left(w^T \phi(x) + b \right)}{\|w\|}$$

Canonical representation of the decision hyperplane

$$t_n \left(w^T \phi(x_n) + b \right) = 1$$ for the point that is closest to the hiperplane

Then

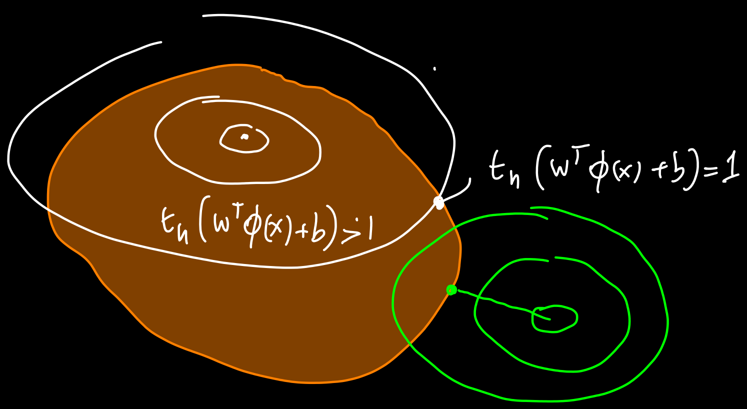
$$t_n \left(w^T \phi(x_n) + b \right) \geq 1 \qquad \forall n \in [1, N]$$



$$\arg\max_{w,b}\left\{\frac{1}{\|w\|}\underbrace{\min_n\left[t_n\left(w^T\phi(x)+b\right)\right]}_{1}\right\} =$$

$$= \arg\max_{w,b}\left\{\frac{1}{\|w\|}\right\} = \underline{\arg\min_{w,b}\frac{1}{2}\|w\|^2}$$

subject to the constraint $\underline{t_n\left(w^T\phi(x)+b\right)\geq 1}$ $\forall n\in[1,N]$



Appendix E

quadratic programming

$a=(a_1 \dots a_N)^T$

Lagrange multipliers $a_n \geq 0$ 1 for each data point

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n\left\{t_n\left(w^T\phi(x_n)+b\right)-1\right\}$$

minimize ≥ 0

$$\frac{\partial L}{\partial w}=0 \qquad w=\sum_{n=1}^{N} a_n t_n \phi(x_n) \longleftarrow$$

$$\frac{\partial L}{\partial b} = 0 \qquad \boxed{0 = \sum_{n=1}^{N} a_n t_n}$$

$$\tilde{L}(a) = \frac{1}{2} \left\| \sum_{n=1}^{N} a_n t_n \phi(x_n) \right\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n \left(\overbrace{\sum_{m=1}^{N} a_m t_m \phi^T(x_m)}^{w^T} \right) \phi(x_n) \right\}$$

$$- b \cancel{\sum_{n=1}^{N} a_n t_n} + \sum_{n=1}^{N} a_n =$$

$$= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \phi^T(x_n) \phi(x_m) - \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \phi^T(x_n) \phi(x_m)$$

$$\underline{\tilde{J}(a)} \qquad + \sum_{n=1}^{N} a_n = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \underbrace{\phi^T(x_n) \phi(x_m)}_{K(x_m, x_n)} + \sum_{n=1}^{N} a_n$$

dual representation for the optimality criterion

$$a_n \geq 0 \qquad n \in [1, N]$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

Quadratic problem in M variables $\Rightarrow$ $\underline{O(M^3)}$

$$L(\overbrace{w}^{M}, b, a)$$

$\tilde{L}(a)$ N variables $\Rightarrow O(N^3)$

$$y(x) = w^T \phi(x) + b = \sum_{n=1}^{N} a_n t_n \underbrace{\phi^T(x_n) \phi(x)}_{K(x, x_n)} + b$$

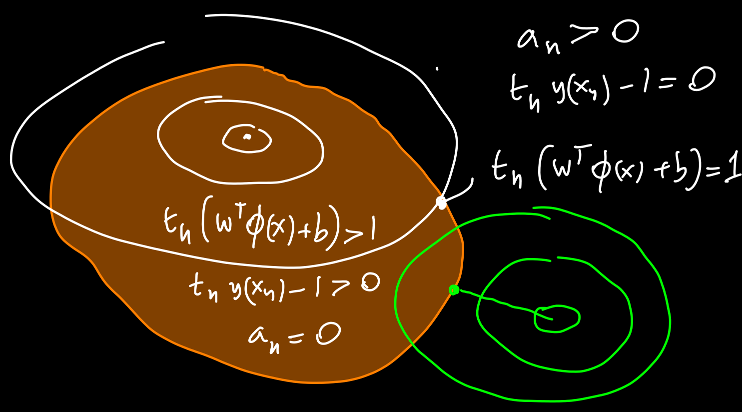
$$= \sum_{n \in N_S} a_n t_n K(x, x_n) + b$$

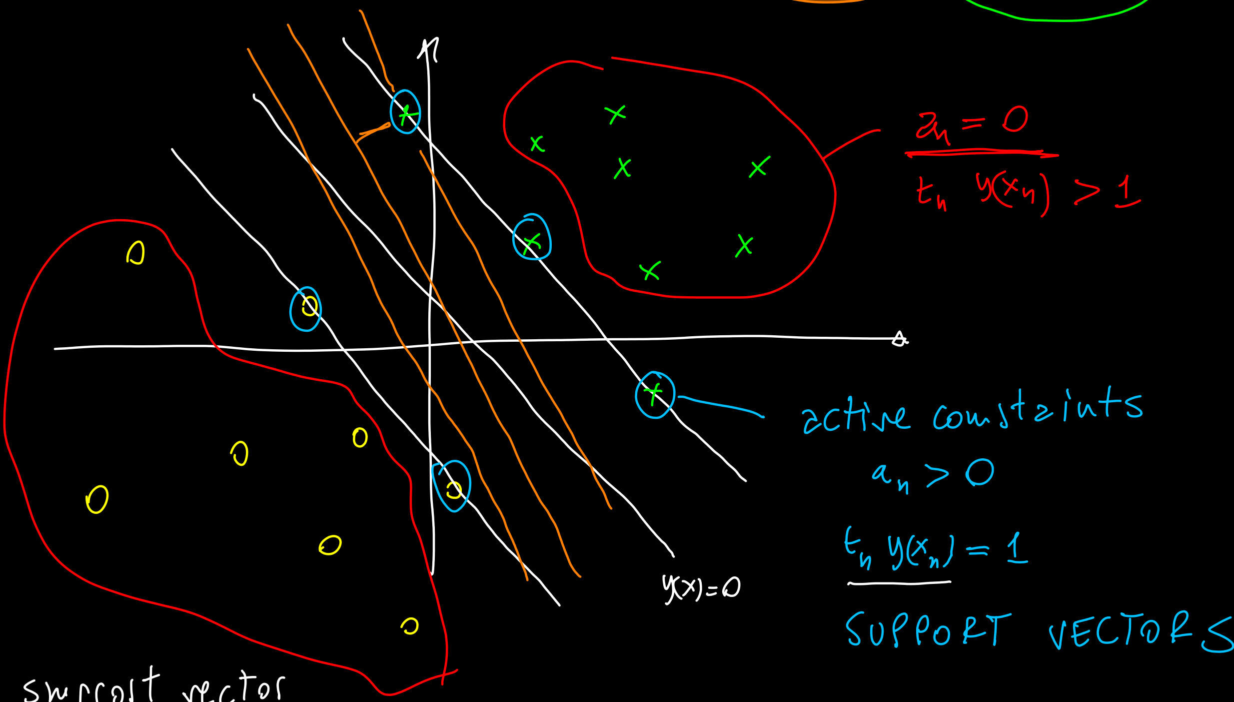
Karush-Kuhn-Tucker KKT conditions:

$$a_n \geq 0$$

$$t_n \underline{y(x_n)} - 1 \geq 0$$

$$\underline{a_n}\{t_n y(x_n) - 1\} = 0$$





if x_n is support vector

$$\Rightarrow 1 = \underset{+1,-1}{t_n}\, y(x_n) = t_n\left(\sum_{m\in S} a_m t_m K(x_n, x_m) + b\right)$$

$$t_n = \underset{\underset{1}{\parallel}}{t_n^2}\left(\qquad\qquad\right)$$

$N_S b$

$$\sum_{n\in S} t_n = \sum_{n\in S}\sum_{m\in S} a_m t_m K(x_n, x_m) + \sum_{n\in S} b$$

$$w = \sum_{n \in S} a_n t_n \phi(x_n)$$

$$b = \frac{1}{N_S} \sum_{n \in S} \left[t_n - \sum_{m \in S} a_m t_m k(x_n, x_m) \right]$$