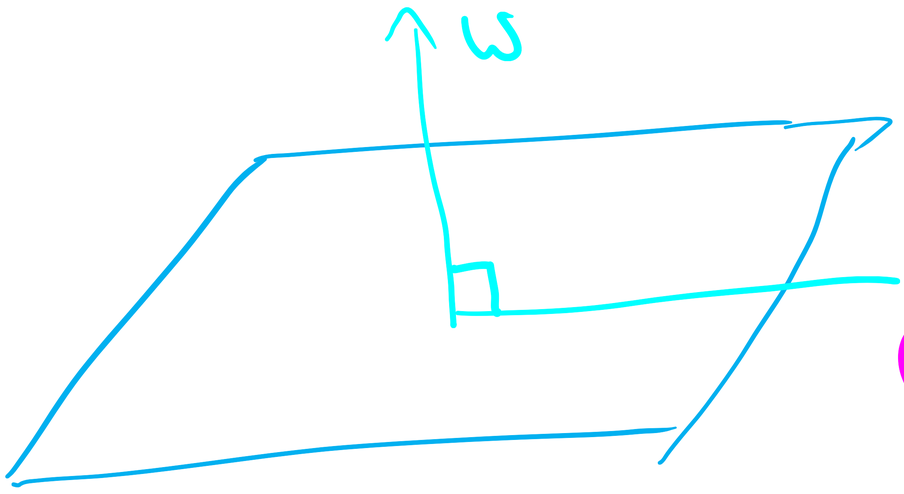
12 Support Vector Machines

- used for classification / regression
- $\rightarrow y_n \in \{-1, 1\}$
- linear model is considered, transformation ϕ for nonlinearity
- geometric Way to think about SV ML
- relies on inner products, projections
- different to Max. Likelihood:
  - Max. Likelihood: model based on probabilistic view of data
  - SVM: designs function to be optimized during training $\rightarrow$ loss function
- Hyperplane (within subspace) $\Rightarrow$ linear seperator

Ideas:

1) Hard margin
2) Soft margin
3) Dual SVM: convex hull of each class

12.1 Seperating Hyperplanes

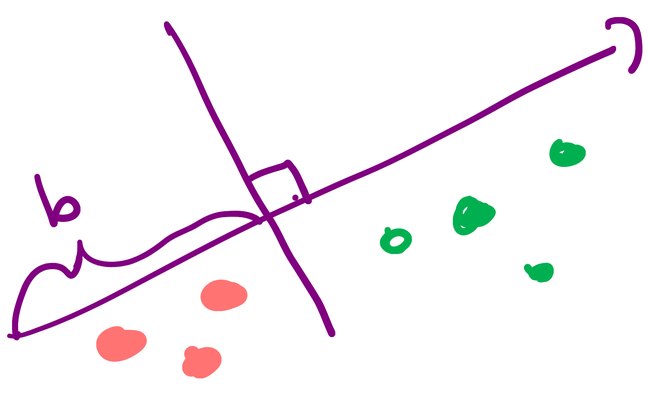


$\rightarrow$ seperates Classes:

$\exists x \in \mathbb{R}^D : f(x) = 0\}$

$\rightarrow$ hyperplane & direction

$\rightarrow$ normal to hyperplane:



$\rightarrow$ choose x_a, x_b on hyperplane:

$$f(x_a) - f(x_b) = \langle w, x_a \rangle + b - \langle w, x_b \rangle + b$$

$$= \langle w, x_a - x_b \rangle = 0 \quad (x_a, x_b = 0)$$

During Training

$\langle w, x_n \rangle + b \geq 0$, when $y_n = 1$

$\langle w, x_n \rangle + b < 0$, $y_n = -1$

12.2 Primal SVM

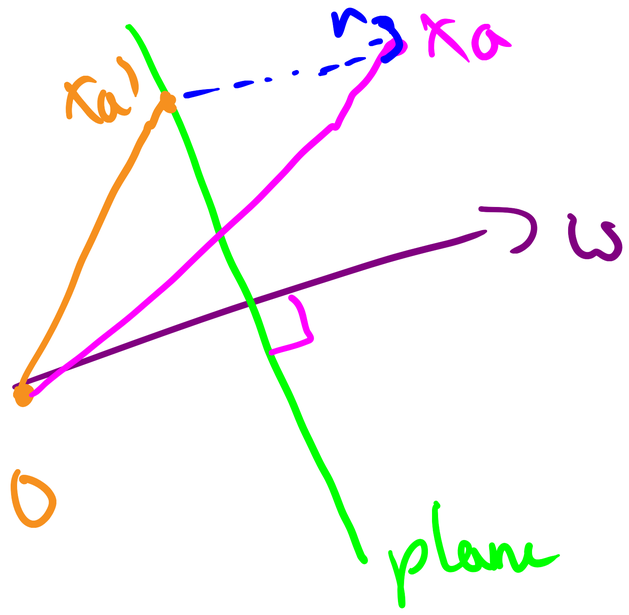
-> many hyperplanes (infinite) possible

=> chose plane w/ max. margin

-> Key: Scale @ which d is measured?

-> Scale as based on hyperplane

r: dist case from hyperplane



r is scaling of w:

$x_a = x_a' + r \frac{w}{\|w\|}$

-> normalize by $\|w\| = \sqrt{w^T w}$

positive examples:
more than r from plane

negative examples:
more than r from plane

$y_n(\langle w_n, x_n \rangle + b) \geq r$

Hard Constrained Optimization Problem:

$$\max_{w,b} \underbrace{r}_{\text{margin}}$$

subject to $y_n(\langle w, x_n \rangle + b) \geq r$, $\underbrace{\|w\| = 1}_{\text{normalization}}$, $r > 0$

-> max r, s.t. data is on right side of the plane

-> Vapnik: large margin => "complexity" low

Margin Max.

$\langle w, x_a' \rangle + b = 0$ / subs.

$\langle w, x_a - r \frac{w}{\|w\|} \rangle + b = 0$ / bilinearity

$\langle w, x_a \rangle + b - r \frac{\langle w, w \rangle}{\|w\|} = 0$

$\downarrow$

$\langle w, w \rangle = \|w\|^2$

$r = \frac{1}{\|w\|}$

→ r derived but we do not know yet

$$\max_{w,b} \frac{1}{\|w\|}$$

Subject to $y_n(\langle w, x_n \rangle + b) \geq 1$ for all $n \in \{1, \dots N\}$

often: squared norm is used:

$$\max_{w,b} \frac{1}{2}\|w\|^2$$

Subject to — " —

→ hard margin SVM (can be relaxed)

Set margin to 1

$$\max_{w,b} \underbrace{\frac{1}{\text{margin}}}$$

Subject to $y_n(\langle w, x_n \rangle + b) \geq r$, $\|w\| = 1$ $r > 0$

equivalent to:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

Subject to $y_n(\langle w, x_n \rangle + b) \geq 1$

Proof → book

Soft Margin SVM

-> allow for outliers

-> Key idea: slack variable ξ_n -> each label pair (x_n, y_n)

↳ substract from margin

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

Subject to $y_n(\langle w, x_n\rangle + b) \geq 1 - \xi_n$

$$\xi_n \geq 0$$

C: Trade off size of margin & slack

$\|w\|^2$: regularizer

C↑ -> regularization ↓ (slack has more weight)

Loss:

Hinge loss:

$l(t) = \max\{0, 1-t\}$ $t = y\,f(x) = y(\langle w, x_n\rangle + b)$

Loss:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\max\{0, 1-y_n(\langle w, x_n\rangle + b)\}$$

↳ regularizer

"Margin max. can be interpreted as regularization"

Dual SVM

- independent on # of features (unlike primal SVM)
  ↳ # of params ↑ when # training examples ↑
- convex duality
- allows for kernels to be applied

$$L(w, b, \xi, \alpha, \gamma) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\alpha_n(y_n(\langle w, x_n\rangle + b) - 1 + \xi_n) - \sum_{n=1}^{N}\gamma_n\xi_n$$

→ partial derivatives to 0:

$$w = \sum_{n=1}^{N}\alpha_n y_n x_n \quad \text{(representer theorem)}$$

→ solution lies in span of the data

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle - \sum_{i=1}^{N}\alpha_i$$

$$\text{subject to } \sum_{i=1}^{N} y_i \alpha_i = 0 \qquad 0 \le \alpha_i \le C \text{ for all } i = 1, \dots, N$$

1) obtain α
2) get w through representer theorem

$$b^* = y_n - \langle w^*, x_n\rangle \quad (x_n \text{ on the boundary})$$

→ if no boundary: median of SVs

// Convex Hull View

- build convex set w/ examples ⇒ smallest set possible (convex hull)
- triangle formed by edges of region

$$\text{conv}(X) = \left\{\sum_{n=1}^{N}\alpha_n x_n\right\} \quad \sum_{n=1}^{N}\alpha_n = 1 \qquad \alpha_n \ge 0$$

- do not overlap when classes are separated

1) pick points c & d in each class:

$$w := c - d \rightarrow \text{require closeness}$$

$$= \arg\min_{w} \|w\| = \arg\min_{w} \frac{1}{2}\|w\|^2$$

$$c = \sum_{n:y_n=+1} \alpha_n^+ x_n \quad \rightarrow \text{must be in positive convex hull}$$

$$d = \sum_{n:y_n=-1} \alpha_n^- x_n$$

substitute

$$\Rightarrow \min_{\alpha} \frac{1}{2} \left\| \sum_{n:y_n=+1} \alpha_n^{+1} x_n - \sum_{n:y_n=-1} \alpha_n^{-1} x_n \right\|^2$$

require: $\sum_{n:y_n=+1} \alpha_n^+ = 1$ $\quad \Rightarrow$ implies: $\sum_{n=1}^{N} y_n \alpha_n = 0$

$\sum_{n:y_n=-1} \alpha_n^- = 1$

$\Rightarrow$ constrained optimization

$\Rightarrow$ dual hard margin SVM

to obtain soft margin: reduced hull (upper bound of convex hull)

2.4 Kernels

→ dual SVM: no inner products between examples & params

→ consider features $\phi(x_i)$ to represent x_i → only change inner product in dual SVM

⇒ classification method & representation method are separate

$\phi(x)$ can be non-linear ⇒ Kernel

Kernel: function K

$K: X \times X \to \mathbb{R}$ for which there exists a Hilbert Space $\mathcal{H}$ & $\phi: X \to \mathcal{H}$ a feature map:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$$

examples ↑ ↑ Similarity measure

⇒ every Kernel has unique Hilbert Space

→ Generalized idea from inner product → feature map:

KERNEL TRICK