SVM

Transform labels:

$\{0,1\} \longrightarrow \{-1,1\}$

$\tilde{y}_n \qquad y_n$

$y_n = 2\tilde{y}_n - 1$

$\tilde{y}_n = \frac{1}{2}(y_n + 1)$

Classification losses:

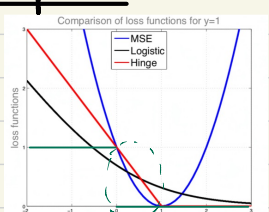
$[1-zy]_+ = \max\{0, 1-z\}$

↳ Hinge(z,y)

$MSE(z,y) = (1-yz)^2$

$\text{Logistic Loss}(z,y) = \log(1+e^{-yz})$

Classification:



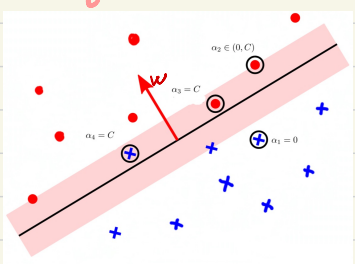
- As we get on the "right side" the hinge loss is 0.
- The cost of hinge loss increases linearly as we get away on the "wrong side"

Support vector Machines

SVM's correspond to the loss: (Hinge + Regul)

$$\min_w \underbrace{\sum_{n=1}^{N} [1 - y_n x_n^T w]_+}_{\text{Hinge}} + \underbrace{\frac{\lambda}{2}\|w\|_2^2}_{\text{L2 Reg}}$$

The region in pink is called **The margin:**

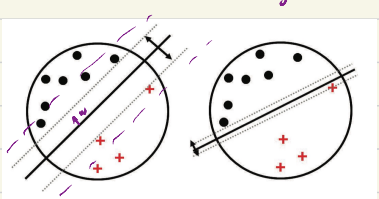


Take the normal vector w that defines the hyperplane, look at the feature vectors x so that $|x^T w| \leq 1$

→ This is the margin: it doesn't only depend on the direction of w, but also on its norm: Total width $= \frac{2}{\|w\|}$

- The goal is: 1. find a separating hyperplane
  (Hinge) ← 2. A scaling of w st no point of the data is in the margin
  (Reg) 3. that the separating hyperplane and scaling of w for which the margin is the largest.

- We have enough if no point in margin (narrow enough hinge = 0) + (Reg)



optimization:

- How to get w?

$$\min_w \sum_{n=1}^{N} [1 - y_n x_n^T w]_+ + \frac{\lambda}{2} \|w\|_2^2$$

→ convex problem!

(one small problem at 1, ⇒ use subgradient-Methods GD or SGD)

- There is another way to look at it:

Convex duality:

$$\min_w \mathcal{L}(w) = \min_w \max_\alpha G(\alpha, w)$$

1. What would be a suitable $G(\alpha, w)$?
2. When can we exchange the min max?
3. When is it useful?

Primal problem: $\min_w \max_\alpha G(\alpha, w)$

⇒ Sometimes the dual problem is simpler to solve

Dual problem: $\max_\alpha \min_w G(\alpha, w)$

1. $[z]_+ = \max\{0, z\}$

$$= \max_{\alpha \in [0,1]} \alpha z$$

$$\mathcal{L}(w) = \frac{1}{N} \sum_{n=1}^{N} \max_{\alpha_n \in [0,1]} \alpha_n (1 - y_n x_n^T w) + \frac{\lambda}{2} \|w\|_2^2$$

$$\Rightarrow \min_w \mathcal{L}(w) = \min_w \max_{\alpha \in [0,1]^N} \frac{1}{N} \sum_{n=1}^{N} \alpha_n (1 - y_n x_n^T w) + \frac{\lambda}{2} \|w\|_2^2$$

$$\Rightarrow G(\alpha, w) = \sum_{n=1}^{N} \alpha_n (1 - y_n x_n^T w) + \frac{\lambda}{2} \|w\|_2^2$$

G is convex in w and concave in α (linear in α)

- $\min_w G(\alpha, w) \leq G(\alpha, w') \quad \forall w'$

$\max_\alpha \min_w G(\alpha, w) \leq \max_\alpha G(\alpha, w')$ $\quad w' \in \mathbb{R}$, $G(\alpha, w')$ convex

$\Rightarrow \max_\alpha \min_w G(\alpha, w) \leq \min_{w'} \max_\alpha G(\alpha, w')$

2. We have: $\max_\alpha \min_w G(\alpha, w) \leq \min_w \max_\alpha G(\alpha, w)$

- Equality if G is convex and concave and domain is compact ("Bounded") and convex: $\max_\alpha \min_w G(\alpha, w) = \min_w \max_\alpha G(\alpha, w)$

$$\Rightarrow \min_w \mathcal{L}(w) = \max_{\alpha \in [0,1]^N} \min_w \frac{1}{N} \sum_{n=1}^{N} \alpha_n (1 - y_n x_n^T w) + \frac{\lambda}{2} \|w\|_2$$

Compute the min:

$$\nabla_w G(\alpha, w) = \sum_{i=1}^{N} (-\alpha_i y_i x_i) + \lambda w = 0$$

$$\Rightarrow w = \frac{1}{\lambda} \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$= \frac{1}{\lambda} \underset{(p \times N)}{X^T} \underset{(N \times N)}{Y} \underset{(N \times 1)}{\alpha} \quad ; Y = \text{diag}(y)$$

$$w(\alpha) = \frac{1}{\lambda} X^T Y \alpha$$

$$\Rightarrow \min_w \mathcal{L}(w) = \max_{\alpha \in [0,1]^N} \sum_{n=1}^{N} \alpha_n \left(1 - y_n x_n^T w(\alpha)\right) + \frac{\lambda}{2} \|w\|_2^2$$

($\frac{\lambda}{2}$ doesn't affect the max)

$$= \max_{\alpha \in [0,1]^N} \sum_{n=1}^{N} \alpha_n \left(1 - \frac{1}{\lambda} y_n x_n^T X^T Y \alpha\right) + \frac{\lambda}{2} \frac{1}{\lambda^2} \|X^T Y \alpha\|^2 \quad (\alpha_{N \times 1})$$

$$= \max_{\alpha \in [0,1]^N} \sum_{n=1}^{N} \alpha_n - \frac{1}{\lambda} \sum_{n=1}^{N} \underbrace{\alpha_n y_n x_n}_{\alpha^T Y X} \cdot X^T Y \alpha + \frac{1}{2\lambda} \|X^T Y \alpha\|_2^2$$

$$1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{N \times 1}$$

$$= \max_{\alpha \in [0,1]^n} \alpha^T \cdot \mathbf{1} - \frac{1}{\lambda} \|X^T Y \alpha\|_2^2 + \frac{1}{2\lambda} \|X^T Y \alpha\|_2^2$$

$$= \max_{\alpha \in [0,1]^n} \alpha^T \cdot 1 - \frac{1}{2\lambda} \|X^T Y \alpha\|_2^2$$

$$\min_w \mathcal{L}(w) = \max_{\alpha \in [0,1]^n} \alpha^T \cdot 1 - \frac{1}{2\lambda} \alpha^T \underbrace{Y X X^T Y}_{\text{PSD matrix}} \alpha$$

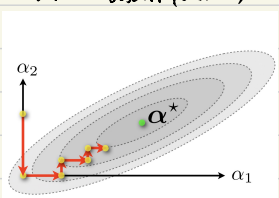
3. Why is this better?

$$\min_w 2(w) = \max_{\alpha \in [0,1]^N} \alpha^T \cdot 1 - \frac{1}{2\lambda} d^T y X X^T y \alpha$$

concave in α ($-(\cdot)^2$)

1) Non-diff prob. can be solved to solve
→ now we have a differentiable, concave problem:
- can be solved using quadr. prog. tools
- useful to use coordinate ascent (1 variable)

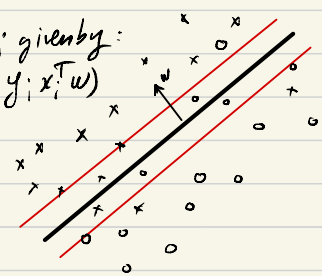
Coordinate ascent (descent)



2) The cost function is only depending on the data in the form:
$K = XX^T_{(N \times N)}$ (does not depend on D!) "Kernel"

3) It gives a good interpretation of SVM

- α is sparse, for any x_i, we have α_i given by:
$$\max_{\alpha_i \in [0,1]} \alpha_i (1 - y_i x_i^T w)$$
- if x_i is correct and outside the margin:
$1 - y_i x_i^T w \leq 0 \Rightarrow \alpha_i = 0$
→ non-support point
- correct point on the margin:
$1 - y_i x_i^T w = 0 \Rightarrow \alpha_i \in [0,1]$ → essential support vector
- Points that are strictly inside the margin: $1 - y_i x_i^T w > 0 \rightarrow \alpha_i = 1$
or on the wrong side
→ bound support vector



$$w = \frac{1}{\lambda} \sum_{n=1}^{N} x_i y_i \alpha_i$$; linear comb. of α_i
depends on support points

Issues with SVM:
- Difficult to to extend to multiclass / no obvious probab. interpretation