Lecture #16: 11-1-18

Last time: Suppose data $\{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathbb{R}^D$, $y_i \in \{-1, 1\}$

are linearly separable by a hyperplane. This means there is some $w \in \mathbb{R}^d$ ("direction"), $b \in \mathbb{R}$ ("offset") such that

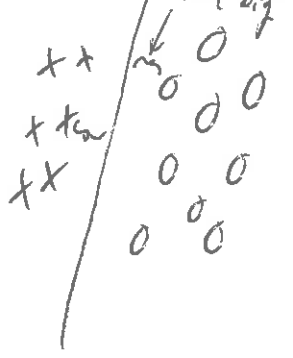
$$w x_i + b = \begin{cases} +, & y_i = 1 \\ -, & y_i = -1 \end{cases} \iff y_i(w x_i + b) > 0 \text{ for all } i.$$

↑ that any w, b exist

In such a case, one can find more (infinitely, in fact), by trying to find the feasible point u satisfying $Vu > (0, \dots, 0)$, where $V \in \mathbb{R}^{n \times d}$ is $V = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} y_1[w^T x_1 + b] \\ \vdots \\ y_n[w^T x_n + b] \end{pmatrix}$. This can be done using linear programming, but could be slow.

Obvious mathematical concern: if many separating hyperplanes exist, how to choose one?

The SVM approach suggests the one that maximizes the margin, i.e. maximize the separation between the classes:



- A calculation reveals the margin is maximized while separating by solving the optimization problem

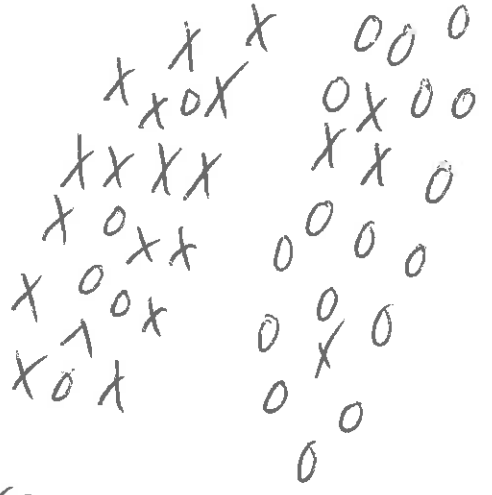
maximize margin ⇔ minimize $\|w\|_2$

$$\text{minimize } \|w\|_2 \quad \text{subject to } y_i(w^T x_i + b) \geq 1 \quad i = 1, \dots, n \qquad (\star)$$

stay separated.

- In a very idealized setting, the constraint set is non-empty and this problem can be solved (very slowly) with quadratic programming (need optimization theory to say more).

- However, for lots of real (and synthetic!) data, this feasible set is empty:



No separating hyperplane at all!

So, how to adjust $(\star)$ to give a reasonable solution when no exact separator exists?

- Replace the "hard" constraint $y_i(w^T x_i + b) \geq 1 \quad i = 1, \dots, n$ required with a "soft" penalty that encourages, but doesn't require $y_i(w^T x_i + b) \geq 1$.

minimize $F(w,t) = \|w\|^2 + \tilde{\lambda} \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + t))$ (★★)

($\tilde{\lambda}$: tuning parameter)

$\|w\|^2$: margin maximizing

$\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + t))$: try to separate

Remark: If a separating hyperplane does exist, then it is possible to make $\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + t)) = 0$. So, in this case, as $\lambda \to \infty$, we recover (★).

- The beauty of (★★) is it makes sense even if a separating hyperplane doesn't exist.

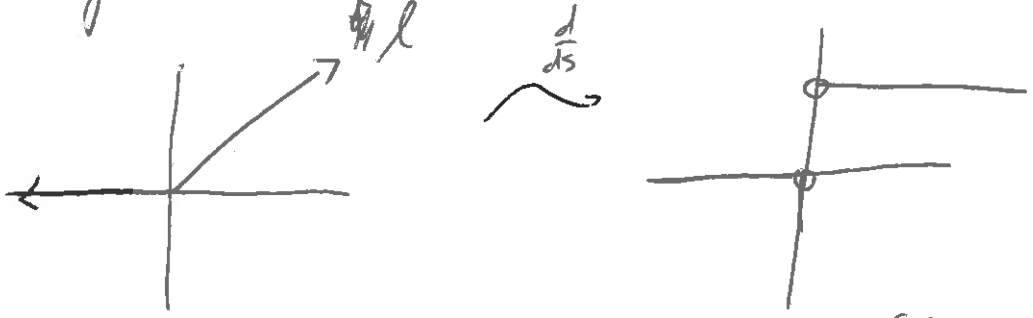
- How to think of the three pieces of (★★):
  1) $\|w\|^2$: regularization term — enforces the secondary property we care about, ie. margin maximizing
  2) $\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + t))$: loss/penalty term: enforces the primary property we care about.
  3) λ: tuning parameter — how to balance the loss and regularizer terms.

Remark: The loss function is of the form

$$\sum_{i=1}^{n} \ell\left(1 - y_i(w^T x_i + b)\right),$$

where $\ell(s) = \max(0, s)$ is the so-called "hinge loss" function. It basically penalizes linearly if you can't get $y_i(w^T x_i + b) \geq 1$, and stops if/when that is achieved, i.e. $y_i(w^T x_i + b)$ is no worse than $y_i(w^T x_i + b) \geq 2$.

- Note that the hinge loss has a "sharp point" at $s = 0$ where it is not differentiable:



- This can be corrected in a variety of ways if desired (for example, to make derivative methods possible). Perhaps the simplest way is to set

$$\tilde{\ell}(s) = \max(0, s^2)$$

$\Rightarrow$ loss is

$$\sum_{i=1}^{n} \tilde{\ell}\left(1 - y_i(w^T x_i + b)\right)$$

$$= \sum_{i=1}^{n} \max\left(0, \left(1 - y_i(w^T x_i + b)\right)^2\right).$$

- Regardless of which loss is used, one needs to solve an (unconstrained) optimization problem:

$$\arg\min_{(w,t)} \|w\|^2 + \lambda \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + t))$$

- There are many approaches to solving this (quadratic programming/convex optimization, sub-gradient descent, coordinate descent), but we will focus only on classical optimization methods.

- To do there, we will cast our optimization problem in its dual formulation.

Next time, we will discuss the theory of dual optimization, and derive the (simpler to solve) dual problem