SVM

Let

- $S = (x_1, y_1) \dots (x_m, y_m)$
- $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$

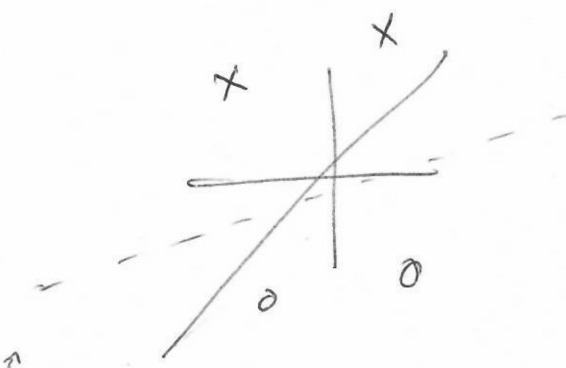
Def Training set S is linearly separable if $\exists$ a hyperplane (w, b) s.t.

$$y_i = \text{sign}(\langle w, x_i \rangle + b) \quad \forall i$$

Equivalently,

$$y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \in [m]$$

Q. There may be many separating halfspaces which one should we pick?



which hyperplane is preferable?

Idea: pick the hyperplane with the largest margin. We call this rule for picking a separating hyperplane "hard SVM".

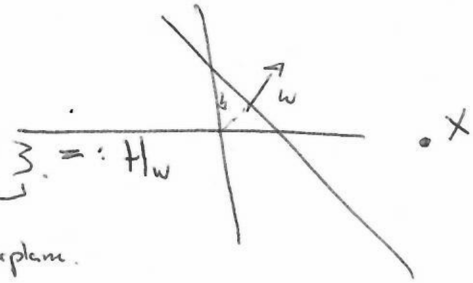
Preliminary result

Claim The distance between a point x and the hyperplane given by (w, b), where $\|w\| = 1$ is $|\langle w, x \rangle + b|$.

?.b?

pf 1. (Recenter) Let $y = x - bw$.

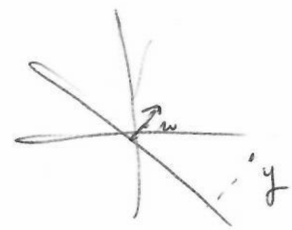
Goal: find distance from y to $\underbrace{\{v : \langle w, v\rangle = 0\}}_{\text{debiased hyperplane}} =: H_w$



2. Guess that ~~the~~ ~~minimal dis~~ ~~closest point~~ ~~is the projection of y onto halfspace~~

~~the minimal distance is given by~~ the closest point is

$$y - \underbrace{\langle w, y\rangle w}_{\text{projection of } y \text{ onto span } w} = p$$



We can formalize this using Hilbert projection thm. But we'll take a more brute force approach

~~The distance is given by~~ ~~$y - p$~~ ~~is~~ ~~$|y - p| = |\langle w, y\rangle w|$~~

Claim: $p \in H_w$.

pf.

wt $p^T w = 0$.

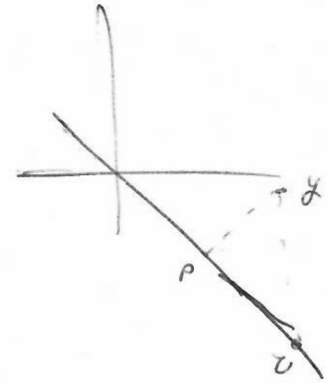
$$p^T w = (y - \langle y, w\rangle w)^T w = y^T w - \langle y^T, w\rangle \overbrace{\cancel{\|w\|^2}}^{=1} = 0 \quad \blacksquare$$

~~distance is given by~~

Thus we see that

$$d(y, H_w) \le |y - p| = |\langle w, y\rangle|$$

Claim $d(y, H_w) = |\langle w, y\rangle|$



pf Let $u \in H_w$. Trick: use inner product structure.

$$\|y-u\|^2 = \|y-p+p-u\|^2$$
$$= \|y-p\|^2 + \underbrace{\|p-u\|^2}_{\geq 0} + 2\langle y-p, p-u\rangle$$
$$\geq \|y-p\|^2 + 2\underbrace{\langle y-p, p-u\rangle}_{=0 \text{ b/c}}$$
$(y-p) \perp H_w$, & $p, u \in H_w$
$$= \|y-p\|^2$$

Hence, $d(y, H_w) = |\langle w, y\rangle|$

~~Finally, $y = x - bw$ $d(x$~~

claim $d(y, H_w) = d(x, H_{(w,b)})$

so $d(x, H_{w,b}) = |\langle w, y\rangle| = |\langle w, x-wb\rangle|$
$= |\langle w, x\rangle - b|$

Missed a minus sign somewhere

To Do
- Soft svm
- Experiments / demo in Jupyter notebook
- Example in image processing?

~~Soft svm~~

~~Hard svm requires datasets to be separable. Soft svm relaxes this.~~
~~The following is an equivalent~~

~~Since~~ We see that the distance to the closest point in the training set to a hyperplane is given by $\min_{i\in[m]} |\langle w, x_i\rangle + b|$. The hard svm rule is given by

$$\operatorname*{argmax}_{(w,b):\|w\|=1} \min_{i\in[m]} |\langle w, x_i\rangle + b| \quad \text{s.t.} \quad y_i(\langle w, x_i\rangle + b) > 0 \quad \forall i$$

The Hard SVM rule can be reformulated as a quadratic program (quadratic objective w/ cvx constraints).

QHSvm

> input: $(x_1, y_1) \dots (x_m, y_m)$
>
> Solve:
>
> $$(w_0, b_0) = \underset{(w,b)}{\arg\min} \|w\|^2 \text{ s.t. } y_i(\langle w, x_i \rangle + b) \geq 1 \ \forall i$$
>
> output: $\hat{w} = \frac{w_0}{\|w_0\|}$, $\hat{b} = \frac{b_0}{\|w_0\|}$

Claim:

The output of $\mathcal{A}$ is an optimal solution to the previous hard svm formulation.

Pf. Would like to show that a solution of QHSvm yields a margin at least as small as the margin in H-svm. Let w^* be a soln to H-svm & let γ^* be the associated margin; $\gamma^* := \min_{i \in [m]} y_i(\langle w, x_i \rangle + b)$.

We have that

$$y_i(\langle w^*, x_i \rangle + b^*) \geq \gamma^* \quad \forall i$$

equivalently

$$y_i\left(\langle \frac{w^*}{\gamma^*}, x_i \rangle + \frac{b^*}{\gamma^*}\right) \geq 1$$

Hence, $\left(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*}\right)$ satisfies constraints of QHsvm, so $\|w_0\| \leq \|\frac{w^*}{\gamma^*}\| = \frac{1}{\gamma^*}$

For all i we get

$$(*) \quad y_i\left(\langle \hat{w}, x_i \rangle + \hat{b}\right) = \frac{1}{\|w_0\|} y_i\left(\langle w_0, x_i \rangle + b_0\right) \geq \frac{1}{\|w_0\|} \geq \gamma^*.$$

So, $\hat{w}$ is an optimal soln to H-svm.

Proof structure: Logically, the flow of the pf as you'd think of it for the first time goes a little different. 1) Want to show margin achieved by $(\hat{w}, \hat{b})$ is as good as H-svm margin. 2) Define γ to be H-svm margin 3) Then, deduce that QHsvm margin satisfies bound in $(*)$. 4) Finally, deduce that $\|w_0\| \leq \frac{1}{\gamma^*}$.

↳ Their way is cleaner, but this is easier to see logical flow.

Soft SVM

Hard SVM assumes training set is linearly separable. Lets relax this.

~~Allow Some~~ H-SVM (quad. version) enforces the constraint

$$y_i\left(\langle w, x_i\rangle + b\right) \geq 1 \qquad \forall i$$

Let's relax this with slack variables ξ_i, $i=1,\dots,m$, so the constraint becomes

$$y_i\left(\langle w, x_i\rangle + b_i\right) \geq 1 - \xi_i \qquad \forall i.$$

Soft SVM

input: $(x_1, y_1)\dots(x_m, y_m)$

Parameter: λ

Solve:

$$\min_{w,b,\xi}\left(\lambda\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i\right)$$

$$\text{s.t.} \quad y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \xi_i \qquad \forall i$$

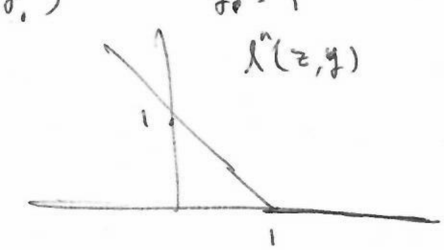
$$\xi_i \geq 0 \qquad \forall i$$

We may reformulate soft SVM into a more convenient form as follows.

Define the hinge loss $\cancel{\ell^{hinge}(z) =}\ z \mapsto \ell^h(\overset{\in\mathbb{R}}{z}, \overset{\in\{-1,1\} = \text{training label}}{y_0})$

$y_0 = 1$

$\ell^h(z,y)$



Soft svm is equivalent to

$$\min_{(w,b)} \frac{1}{m}\sum_{i=1}^{m} l^h(\langle w, x_i\rangle + b, y_i) + \lambda \|w\|^2$$

pf Consider soft svm.
Since $\xi_i \geq 0$, the optimal choice for ξ_i given the objective is to pick ξ_i as close to 0 as possible. How close you can get ~~This~~ depends on the other constraint:

~~if $y_i(\langle w, x_i\rangle + b)$~~

$$y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$$

If $y_i(\langle w, x_i\rangle + b) \geq 1$, choose $\xi_i = 0$. otherwise, the best choice of ξ_i is $1 - y_i(\langle w, x_i\rangle + b)$. Substituting these optimal choices for ξ_i back into the soft svm objective gives exactly the hinge loss formulation.