Office hours today: MOVED to 4:30-5:30 on Zoom only.

Convex optimization problems

Lecture 15

November 16, 2021

Beck, sections 8.1-8.3

Convex optimization problems

An optimization problem is <u>convex</u> if it's of the form

$$\min_{x \in C} f(x), \qquad f \text{ is convex over } C$$

C is a convex set (typically in $\mathbb{R}^n$)

C: "feasible" set

f: "objective"

$x \in \mathbb{R}^n$ is "feasible" if $x \in C$.

Note: $\max_{x \in C} g(x)$ is also a convex problem if g is concave.

The above formulation is "implicit". "Explicit" formulations typically (not always) take the form:

$$\min_{\underline{x} \in \mathbb{R}^n} f(\underline{x}) \quad \text{"such that" subject to} \quad \begin{array}{l} g_i(x) \le 0, \; i = 1, .. m \\ h_j(x) = 0, \; j = 1, .. n \end{array}$$

for some given convex functions $\{g_i\}_{i=1}^{m}$ and affine functions $\{h_j\}_{j=1}^{n}$.

Feasible set: $\{\underline{x} \in \mathbb{R}^n \mid g_i(\underline{x}) \le 0, \; i = 1, .. m$
$h_j(\underline{x}) = 0, \; j = 1, .. n\} = C$ (convex)

Why are convex optimization problems nice?

We have "nice" characterizations of existence/uniquess to solutions.

Theorem: Suppose f is a convex function over a convex set $C \subset \mathbb{R}^n$. If $\underset{\sim}{x} \in C$ is a local minimum of f over C, then it solves:

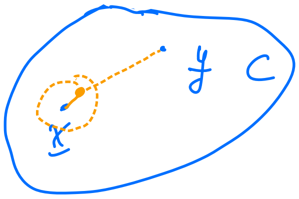
$$\underset{\underset{\sim}{x} \in C}{\operatorname{argmin}} f(\underset{\sim}{x}).$$

("Local minima are global minima")

Proof: Let $y \in C$. Since $\underset{\sim}{x}$ is a local minimum, then there's a small neighborhood around $\underset{\sim}{x}$ where f is greater than or equal to $f(\underline{x})$.

I.e., $\exists\, \varepsilon > 0$ s.t. $\forall\, \underline{z} \in B(\underline{x}, \varepsilon) \cap C$

$f(\underline{z}) \geq f(\underline{x})$.



Let $\delta = \varepsilon/2$ (assume $\delta \leq 1$).

Note: $\underline{x} + \delta(\underline{y} - \underline{x}) = \delta \underline{y} + (1-\delta)\underline{x} \in C$

$$f(\underline{x}) \leq f(\underline{x} + \delta(\underline{y} - \underline{x})) = f(\delta \underline{y} + (1-\delta)\underline{x})$$

$\underline{x} + \delta(\underline{y} - \underline{x}) \in C,\ B(\underline{x}, \varepsilon)$

$$\leq \delta f(\underline{y}) + (1-\delta) f(\underline{x})$$

$$\xRightarrow{\delta > 0} \quad f(\underline{x}) \leq f(\underline{y}). \qquad \blacksquare$$

Another property:

<u>Theorem</u>: Suppose f is convex over a convex set C.

Then:

$$S = \operatorname*{argmin}_{\underline{x} \in C} f(\underline{x}) \text{ is a convex set.}$$

(The set of all solutions is "nice".)

<u>Proof</u>: Assume $\underline{x}, \underline{y} \in S$. Let $\lambda \in (0,1)$

$$f(\lambda \underline{x} + (1-\lambda)\underline{y}) \leq \lambda f(\underline{x}) + (1-\lambda) f(\underline{y})$$

$$= [\lambda + (1-\lambda)] \min_{z \in C} f(z)$$

$$= \min_{z \in C} f(z)$$

$$\Rightarrow \lambda x + (1-\lambda) y \in S. \qquad \blacksquare$$

<u>Corollary</u>: If f is strictly convex over a convex set C, then a global minimizer is unique.

Rest of today: examples.

Linear programming

Applications: scheduling, allotment, cost.

Given $\underline{c} \in \mathbb{R}^n$, $\underline{\underline{A}} \in \mathbb{R}^{m \times n}$, $\underline{b} \in \mathbb{R}^m$, $\underline{\underline{C}} \in \mathbb{R}^{k \times n}$, $\underline{d} \in \mathbb{R}^k$,

$$\left.\begin{aligned} \min_{\underline{x} \in \mathbb{R}^n} \ \underline{c}^T \underline{x} \quad \text{subject to} \quad & \underline{\underline{A}}\,\underline{x} \leq \underline{b} \\ & \underline{\underline{C}}\,\underline{x} = \underline{d}. \end{aligned}\right) \text{ is convex.}$$

This is a prototypical linear programming problem.

Note: $\max_{\underline{x} \in \mathbb{R}^n} \underline{c}^T \underline{x}$ s.t. $\underline{\underline{A}}\,\underline{x} \leq \underline{b}$

$\underline{\underline{C}}\,\underline{x} = d$

is also convex since $f(\underline{x}) = \underline{c}^T \underline{x}$ is both convex and concave.

Quadratic programming

$$\min_{\underline{x}\in\mathbb{R}^n} \underline{x}^T \underline{\underline{A}}\,\underline{x} + 2\underline{b}^T\underline{x} + c, \qquad \underline{\underline{A}} \in \mathbb{R}^{n\times n}, \quad \underline{\underline{A}} \succeq \underline{\underline{0}}, \quad \underline{b}\in\mathbb{R}^n, \quad c\in\mathbb{R}$$

This problem is convex.

A related problem:

$$\min_{\underline{x}\in\mathbb{R}^n} \underline{x}^T \underline{\underline{A}}\,\underline{x} + 2\underline{b}^T\underline{x} + c \quad \text{subject to}$$

$$\underline{x}^T \underline{\underline{A}}_i\,\underline{x} + 2\underline{b}_i^T\underline{x} + c_i \leq 0, \quad i = 1, .. m$$

$$\underline{x}^T \underline{\underline{C}}_i\,\underline{x} + 2\underline{d}_i^T + e_i = 0, \quad i = 1 .. k$$

QCQP's and convexity

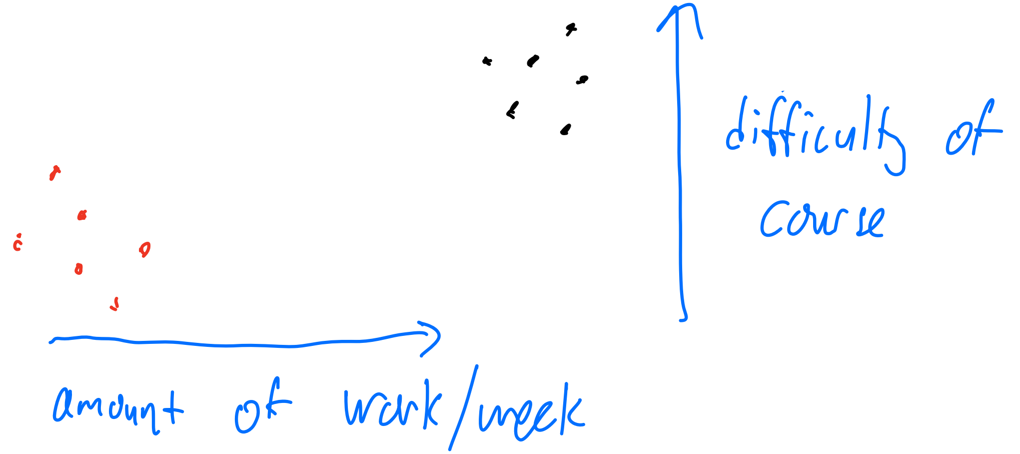
The above problem is a quadratically constrained quadratic program (QCQP).

If there are no equality constraints $(k=0)$ and $\underline{\underline{A}},\ \underline{\underline{A}}_i \succeq \underline{\underline{0}}$, then this is a convex problem.

Example: Chebyshev centers

Goal: generate a region that "best" describes a collection of points.

Courses @ U.



Given all red data: try to fit the smallest circle possible around all the red dots. → "Chebyshev center" problem.

Goal: compute the radius and the center of the smallest enclosing ball.

Optimization:

$$\min_{\underline{c},\, r} \; r \quad \text{s.t.} \quad \|\underline{c} - \underline{x}_i\|_2 \le r \, , \quad i = 1 \ldots \underset{\substack{\uparrow \\ \text{\# of} \\ \text{data points.}}}{m}$$

This is a convex problem. (QCQP)

Hidden convexity

Hidden convexity: some problems are nonconvex, but under an equivalent "remapping", become convex.

Ex. $\min_{x \in [1,2]} \log x$ is not convex.

$$\downarrow \quad x = e^y$$

$\min_{y \in [0, \log 2]} y$. (this is convex)

"Real" example : QCQP

$$\min_{\underline{x} \in \mathbb{R}^n} \underline{x}^T \underline{\underline{A}} \underline{x} + 2\underline{b}^T \underline{x} + c \quad \text{subject to}$$

$$\|\underline{x}\|_2 \leq 1 .$$

$$\underline{\underline{A}} \in \mathbb{R}^{n \times n}, \quad \text{indefinite}$$

This is not convex.

Under an appropriate mapping, this *is* a convex problem.

Projections onto convex sets

Idea: find closest point in a convex set to a point outside that set.

$C \subset \mathbb{R}^n$, convex

$\underline{x}$ ~~y~~ $\in \mathbb{R}^n$, assume $\underline{x} \notin C$

Goal: compute $\operatorname*{argmin}_{y \in C} \|\underline{x} - y\|$

(if $\underline{x} \in C$, then the argmin is $\underline{x}$)

Theorem: Assume $\|\cdot\|$ is strictly convex (e.g. $\|\cdot\| = \|\cdot\|_2$). Then given $\underline{x} \in \mathbb{R}^n$, $C \subset \mathbb{R}^n$ that is convex, and non-empty, and closed, the problem

$$\underset{\underline{y} \in C}{\operatorname{argmin}} \|\underline{x} - \underline{y}\|$$

has a unique solution.

Proof (sketch): The problem is

$$\min_{y \in \mathbb{R}^n} \|\underline{x} - y\|$$
$$\text{s.t. } y \in C$$

We know: $f(y) = \|\underline{x} - y\|$ is strictly convex.
($\|\cdot\|$ is strictly convex, and $\Phi(y) = \underline{\underline{I}}\,\underline{y} - \underline{x}$ is affine)

So: this is optimization of a strictly convex function over a convex set.

$\Rightarrow$ local minima are global minima.

(can show that $\exists$ a single local minimum). $\square$

Since the above minimization problem is well-posed ($\exists$ a unique solution), we can call the solution to this problem as a projection.

Def: Given $C \subset \mathbb{R}^n$ that is closed, convex, then $P_C : \mathbb{R}^n \to C$ is the projection operator of the convex set C, defined by solving

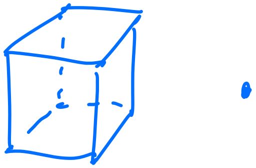
$$\underset{y \in C}{\operatorname{argmin}} \; \|\underline{x} - y\|$$

for any $\underline{x} \in \mathbb{R}^n$. Then $P_C(\underline{x})$ is called the projection (of $\underline{x}$) onto C.

Although we have nice theory for P_C, actually computing it is pretty difficult, except in special cases.

Projection examples

Ex 1 (Projection onto hypercube)



$$C = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$$

$a_i \le b_i$, $i = 1 \ldots n$

a_i could be $-\infty$

b_i could be $+\infty$

C is closed and convex. ($C = \mathbb{R}^n_+$ is one example)

In ℓ^2: The minimization problem is

$$\min_{y \in C} \sum_{i=1}^{n} (y_i - x_i)^2 \quad \text{s.t.} \quad a_i \le \overset{y_i}{\cancel{x_i}} \le b_i \quad \forall\, i = 1, \ldots n.$$

This problem is "separable": each dimension can be treated independently.

I.e., can consider $\min_{y_i \in \mathbb{R}} (y_i - x_i)^2$ s.t. $a_i \le y_i \le b_i$

which is solved by:

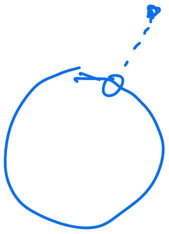
$$x_i^* = \begin{cases} x_i, & x_i \in [a_i, b_i] \\ a_i, & x_i < a_i \\ b_i, & x_i > b_i \end{cases}$$

("clamping")

The full solution is $\underline{x}^* = (x_1^*, \dots x_n^*) = P_C(\underline{x})$

Ex. 2: $\|\cdot\| = \|\cdot\|_2$

$C = B[0, r]$, $r > 0$



$$\operatorname{argmin}_{y \in \mathbb{R}^n} \|\underline{x} - y\|_2 \text{ s.t. } y \in C$$

$$\Updownarrow$$

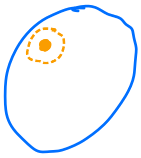
$$\operatorname{argmin}_{y \in C} \|\underline{x} - y\|_2^2$$

($t \mapsto t^2$ is strictly monotone on $[0, \infty)$)

Note: $f(\underline{y}) = \|\underline{x} - \underline{y}\|_2^2$ is quadratic.

Suppose that $P_C(\underline{x})$ lies in the <u>interior</u> of C

Since f is strictly convex, then $P_C(\underline{x})$ must be a <u>local</u> minimum in some neighborhood.



$$\implies \nabla f(P_C(\underline{x})) = \underline{0}$$

$$\nabla f(\underline{y}) = \underline{0} \implies \underline{y} = \underline{x}$$

If $\underline{x} \notin C$, this is not possible.
(because $P_C(\underline{x}) \notin C$)

$\xRightarrow{\text{contradiction}}$ $P_C(\underline{x})$ cannot lie in the interior of C.

I.e, $P_C(\underline{x})$ lies on the boundary, $\|P_C(\underline{x})\|_2 = r$.

I.e., $$\underset{\underline{y} \in C}{\operatorname{argmin}} \|\underline{x} - \underline{y}\|_2^2 = \underset{\|\underline{y}\|_2 = r}{\operatorname{argmin}} \|\underline{x} - \underline{y}\|_2^2$$

$$= \underset{\|\underline{y}\|_2 = r}{\operatorname{argmin}} \underbrace{\|\underline{x}\|_2^2 + \|\underline{y}\|_2^2}_{\text{constant}} - 2\langle \underline{x}, \underline{y} \rangle.$$

$$= \underset{\|\underline{y}\|_2 = r}{\operatorname{argmin}} -2\langle \underline{x}, \underline{y} \rangle$$

$$= \underset{\|\underline{y}\|_2 = r}{\operatorname{argmax}} \langle \underline{x}, \underline{y} \rangle$$

this is maximized when $\underline{y}$ is parallel to $\underline{x}$ because of Cauchy-Schwarz.

I.e., the max is achieved when $\underline{y} = \underline{x}\frac{r}{\|\underline{x}\|_2}$

$$\Longrightarrow P_C(\underline{x}) = \begin{cases} \underline{x}, & \|\underline{x}\|_2 \le r \\ \frac{r}{\|\underline{x}\|_2}\underline{x}, & \|\underline{x}\|_2 > r \end{cases}$$

<u>Ex. 3</u>: Let $\underline{\underline{A}} \in \mathbb{R}^{n\times n}$ be symmetric.

$S_+^n = \{\underline{\underline{B}} \in \mathbb{R}^{n\times n} \mid \underline{\underline{B}}$ symmetric, $\underline{\underline{B}} \succeq \underline{\underline{0}}\}$. (set of positive semi-definite matrices)

S_+^n is closed and convex.

$P_{S_+^n}(\underline{\underline{A}}) = ?$ (use matrix 2-norm $\|\cdot\|_2$)

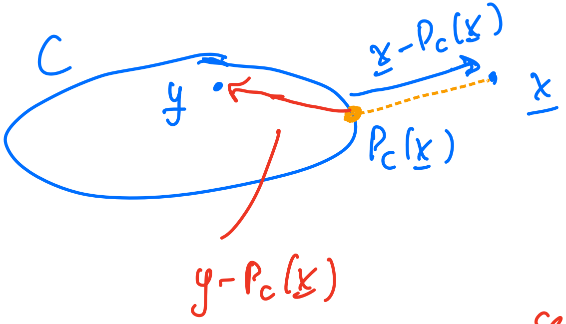
Let $\underline{\underline{A}} = \underline{\underline{U}}\,\underline{\underline{\Lambda}}\,\underline{\underline{U}}^T$ (eigenvalue decomp.)

$\underline{\underline{\Sigma}}$: n×n diagonal matrix, with $\sigma_{i,i} = \begin{cases} \lambda_{i,i} & \text{if } \lambda_{i,i} \ge 0 \\ 0 & \text{if } \lambda_{i,i} < 0 \end{cases}$

Then $P_{S_+^n}(\underline{\underline{A}}) = \underline{\underline{U}}\,\underline{\underline{\Sigma}}\,\underline{\underline{U}}^T$.

Orthogonal projection properties

There's a nice geometric property of this projection operator.



This picture suggests that

$$\langle \underline{x} - P_c(\underline{x}) , y - P_c(\underline{x}) \rangle \le 0 \quad \forall y \in C.$$

Theorem: Let C be a convex, closed set, let $\underline{x} \notin C$. Then $\underline{x}^* \in C$ satisfies

$$\underline{x}^* = P_c(\underline{x}) \text{ iff } \forall y \in C, \quad \langle \underline{x} - \underline{x}^*, y - \underline{x}^* \rangle \le 0$$

Theorem (P_c is norm non-expansive)

Let C be closed convex in $\mathbb{R}^n$. Then $\forall \underline{x}, y \in \mathbb{R}^n$, we have

$$\|P_C(\underline{x}) - P_C(\underline{y})\| \le \|\underline{x} - \underline{y}\|.$$

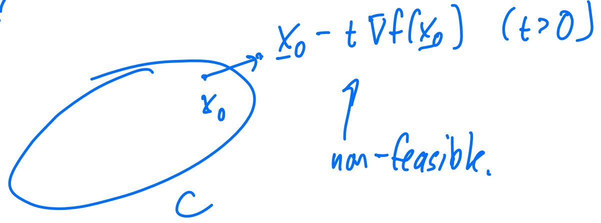
Why do we care about projection onto convex sets?

It's a subproblem in "projected gradient"/"gradient projection" methods.

Suppose we want to numerically compute a solution to

$$\min_{\underline{x} \in C} f(\underline{x}). \quad , C \text{ is convex}, \quad$$ ~~f is convex.~~

Maybe gradient descent?



Projected gradient method: initialize ~~at~~ $\underline{x}_0$

```
For $i = 1, 2..$
    compute $\nabla f(\underline{x}_{i-1})$
    compute stepsize $t_i$
    $\underline{x}_i = P_C(\underline{x}_{i-1} - t_i \nabla f(\underline{x}_{i-1}))$
end
```

This allows us to solve (convex)-constrained optimization problems.

Support vector machines

Goal: classification (binary)

Given: $\{(\underline{x}_i, y_i)\}_{i=1}^{m}$ data pairs

$\underline{x}_i$: "features" of an item

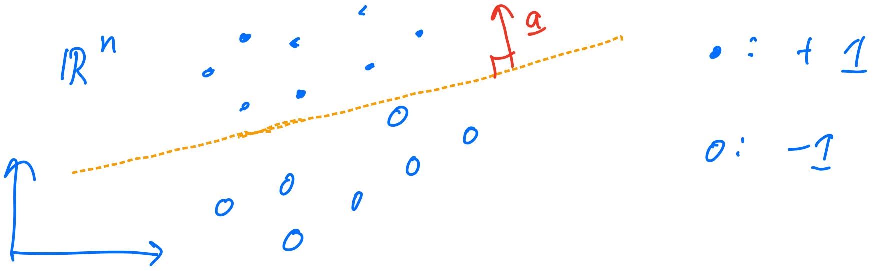
y_i: ± 1, a known classification.

"label"

Output: predictor $f(\underline{x}) \to \pm 1$, $f: \mathbb{R}^n \to \{+1, -1\}$

Support vector machines is a family of classification techniques for building f.

Hypothesis: feature data $(\underline{x}_i)$ are "linearly" separated in $\mathbb{R}^n$ based on their labels

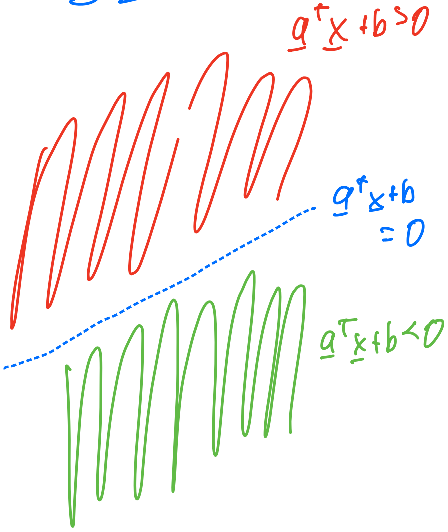


"Linearly separated": $\exists$ a hyperplane in $\mathbb{R}^n$ that divides $\mathbb{R}^n$ into two half-spaces, each of which contains only data of a single label

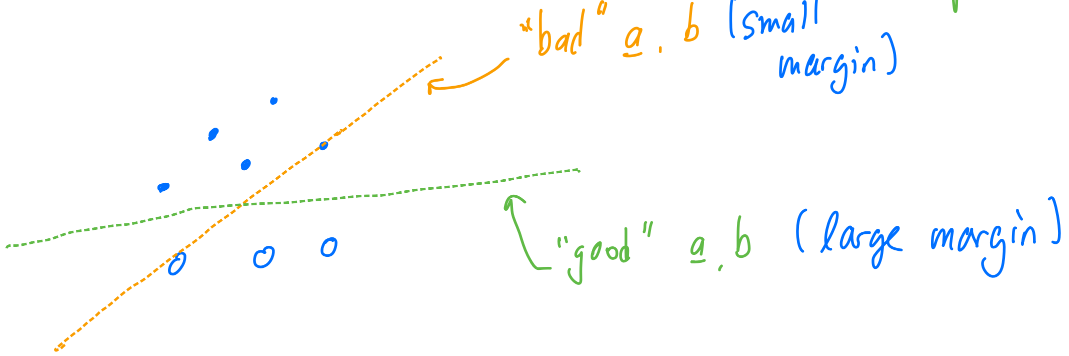
Equations of hyperplanes take the form $\underline{a}^T \underline{x} + b = 0$ for some $\underline{a} \in \mathbb{R}^n \setminus \{\underline{0}\}$ and $b \in \mathbb{R}$.

$\underline{a}$: normal vector to hyperplane.

Predictor/classifier: $f(\underline{x}) = \mathrm{sgn}(\underline{a}^T \underline{x} + b)$



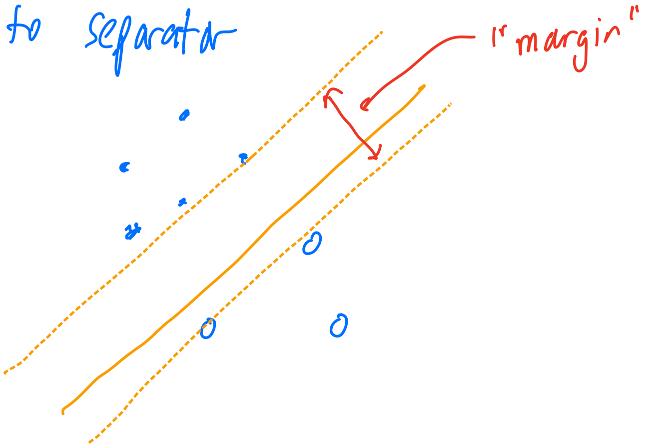
This problem doesn't have a unique solution: find the "best" $\underline{a}, b$.



"bad" $\underline{a}, b$ (small margin)

"good" $\underline{a}, b$ (large margin)

Margin: distance between hyperplanes defining data closest to separator

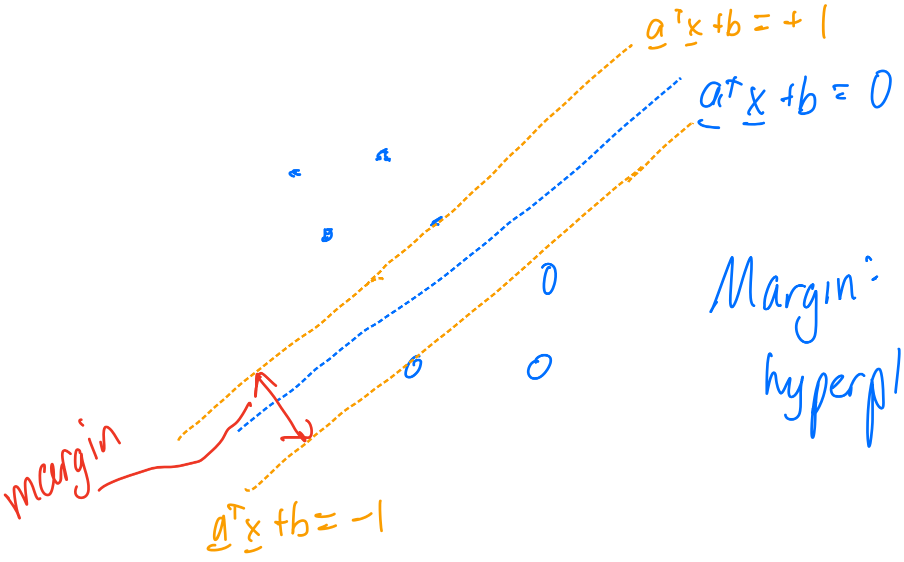


Simplify computation of the margin: can always choose $\underline{a}$, b s.t. for all (-1) labelled data,

$\underline{a}^T\underline{x}_i + b = -1$ is the max value

s.t. for all $(+1)$ labelled data,

$\underline{a}^T\underline{x}_i + b = +1$ is the min value



Margin: distance between hyperplanes

The distance between hyperplanes defined by

$\underline{a}^T\underline{x}+b=+1$ and $\underline{a}^T\underline{x}+b=-1$

is $\frac{2}{\|\underline{a}\|_2}$ ($\underline{x}$ in $\underline{a}^T x+b=-1$, replace by $\tilde{\underline{x}}=\underline{x}+\frac{2\underline{a}}{\|\underline{a}\|^2}$, then

$\underline{a}^T\tilde{\underline{x}}+b=+1$)

Margin

Optimization for $\underline{a}, b$:

$\max_{\underline{a},b}$ (Margin) subject to

$f(\underline{x}_i)=y_i \quad \forall\ i=1..M.$

$\Big\downarrow$ simplification: $|\underline{a}^T\underline{x}_i+b|\geq 1$

Margin is $2/\|\underline{a}\|_2$

$$\max_{\underline{a},b} \frac{2}{\|\underline{a}\|_2} \quad \text{s.t.} \quad y_i(\underline{a}^T\underline{x}_i+b)\geq 1$$

$$\forall\ i=1\ldots M$$

Recall: $\underline{x}_i, y_i$ are given

$\Updownarrow$ max $t \leftrightarrow \min 1/t$ $(t \geq 0)$

$s \mapsto s^2$ is monotonic on $[0, \infty)$

$$\min_{\underline{a}, b} \|\underline{a}\|_2^2 \quad \text{s.t.} \quad y_i(\underline{a}^T \underline{x}_i + b) \geq 1$$

$$\forall\ i = 1, \dots M$$

(Maximizing margin subject to data fidelity)

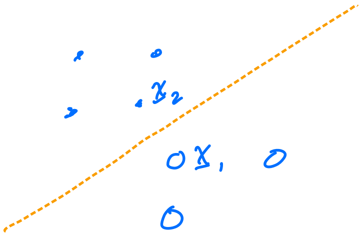
Note: objective $\|\underline{a}\|_2^2$ is convex.

$y_i(\underline{a}^T \underline{x}_i + b) \geq 1$: level sets of convex (affine) functions

$\Longrightarrow$ feasible set is convex.

$\Longrightarrow$ Support vector machine opt. is a convex problem. (It's quadratic.)

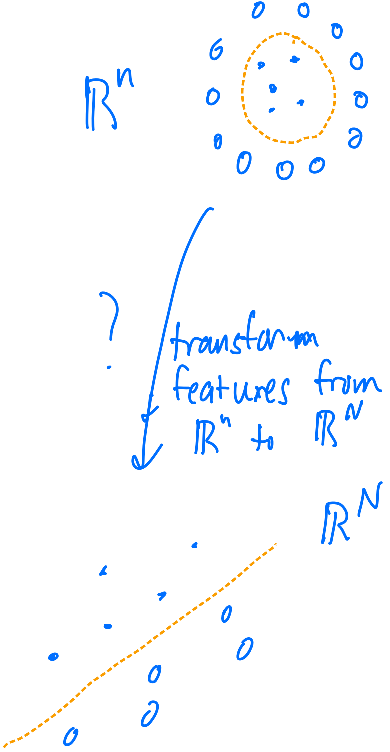
Why "support vector"?



Note: optimization result is unaffected by $\underline{x}_i$, $i \neq 1, 2$.

$\underline{x}_1, \underline{x}_2$ entirely define hyperplane → $\underline{x}_1, \underline{x}_2$ called "support vectors".

This procedures employs linear separability. This can't always be done.



"Modern" versions of SVM employ "nonlinear" separators via the "kernel trick": idea is to replace $\underline{a}^T\underline{x}$ by $K(\underline{a},\underline{x})$.

(e.g. $k(\underline{x},\underline{y}) = \underline{x}^T\underline{y}$ → linear SVM

$K(\underline{x},\underline{y}) = 1-\exp(\underline{x}^T\underline{y})$

$k(\underline{x},\underline{y}) = 1-\exp(\|\underline{x}-\underline{y}\|_2^2)$)

Another practical augmentation: $y_i(\underline{a}^T\underline{x}_i + b) \geq 1$ $\quad i=1 .. M$ is a "hard" cutoff condition.

Terminology: "hard margin"

This is unforgiving and the feasible set can be empty.

Instead: soften problem: hard requirement is $y_i(\underline{a}^T\underline{x}_i + b) \geq 1$

Alternative: $\min \underbrace{1 - y_i(\underline{a}^T \underline{x}_i + b)}_{\text{measure of violation of } y_i(\underline{a}^T\underline{x}_i+b)\geq 1.}$ when $1 - y_i(\underline{a}^T\underline{x}_i+b) \geq 0.$

Soft margin optimization:

$$\min_{\underline{a},b} \|\underline{a}\|_2^2 + \lambda \sum_{i=1}^{M} \max\{0, 1 - y_i(\underline{a}^T\underline{x}_i + b)\}$$

λ: regularization parameter balancing margin maximization against soft violations.

$(\lambda > 0)$, $\lambda \sim \frac{1}{M}$

This is still convex.

→ this soft margin SVM optimization is Tikhonov regularization in disguise.

$$\min: \sum_{i=1}^{M} \max\{0, 1 - y_i(\underline{a}^T\underline{x}_i + b)\} + \frac{1}{\lambda}\|\underline{a}\|_2^2$$

Tikhonov regularization.

— in machine learning, minimizing $\sum_{i=1}^{M} \max\{0, 1 - y_i(\underline{a}^T\underline{x}_i + b)\}$ is called minimizing the empirical risk.