Support Vector Machines (SVMs)

Ref: Chapter 5 of FoML book by Mohri et al.

<u>Setting:</u> Two-class classification problem

Input space $\mathcal{X} \subseteq \mathbb{R}^d$, class labels $\mathcal{Y} \in \{+1, -1\}$

Target function $f : \mathcal{X} \to \mathcal{Y}$

Dataset $S = \{ (x_i, y_i), i = 1 \ldots m \}$, $y_i = f(x_i)$ $\forall i$

↳ The samples drawn from some unknown distribution "D" in an iid fashion.

$R(h) = \underset{x \sim D}{P} \left(h(x) \neq f(x) \right)$ is the generalization error of hypothesis h.

Goal: Minimize the generalization error

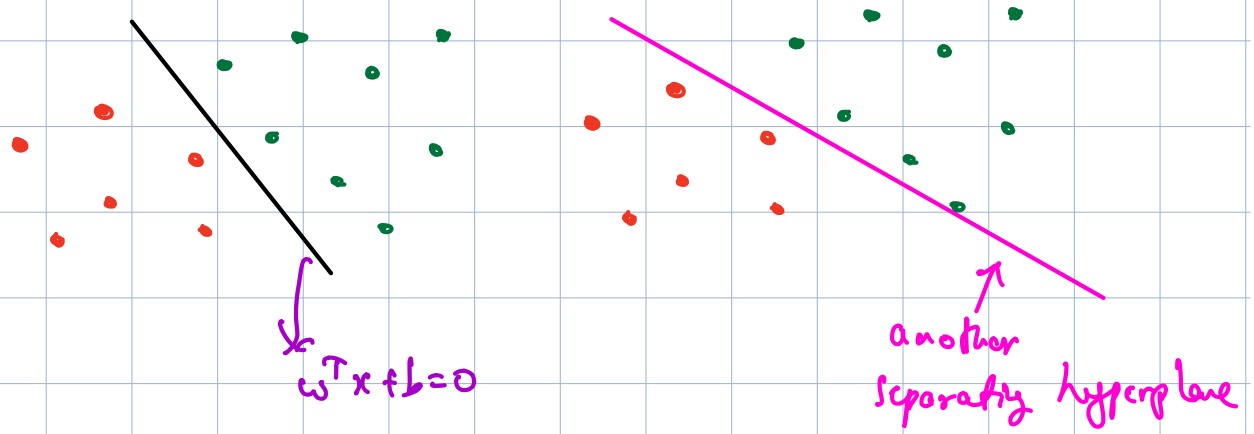
The hypothesis set we consider is

$$\mathcal{H} = \{ x \to \text{sign}(w^T x + b) : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

"Linear classification problem".

$w^T x + b = 0$ is a hyperplane & a hypothesis

$h : x \mapsto \text{sgn}(w^T x + b)$



Linearly separable dataset:

A dataset $S = \{(x_i, y_i), i = 1 \ldots m\}$ is linearly separable if $\exists\, w, b$ s.t.

$$\left.\begin{array}{ll} w^T x_i + b > 0 & \forall i \text{ with } y_i = +1, \text{ and} \\ w^T x_i + b < 0 & \forall i \text{ with } y_i = -1. \end{array}\right\} (*)$$

Note: $(*)$ is equivalent to saying $\exists\, c > 0$ s.t.

$$w^T x_i + b > c \quad \forall i,\ y_i = +1$$
$$w^T x_i + b < -c \quad \forall i,\ y_i = -1$$

$\boxed{\text{Since we have a finite \# of points in } S}$

$\Rightarrow$ we can scale w, b to have

$$w^T x_i + b \geq +1 \quad \forall i \text{ with } y_i = +1$$
$$w^T x_i + b \leq -1 \quad \forall i \text{ with } y_i = -1$$

(or) equivalently, $\forall i$, $y_i (w^T x_i + b) \geq 1$ ← Condition for linear separability

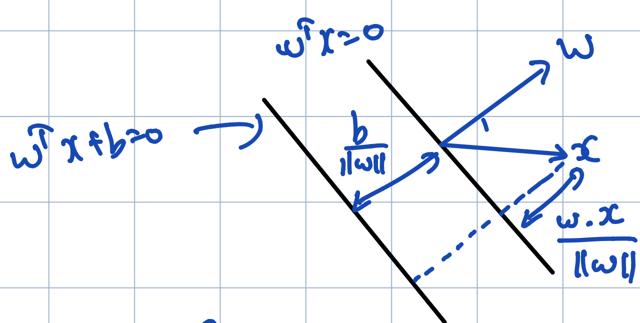
"Assume S is linearly separable".

Concept of margin:

Margin: $\rho_h(x)$ of a linear classifier $h: x \to w^T x + b$ at the point x is

$$\rho_h(x) = \frac{|w^T x + b|}{\|w\|_2}$$

distance of x to the hyperplane $w^T x + b = 0$



(Assuming $w^T x > 0$, $b > 0$)

distance of x to $w^T x + b = 0$ is

$$\frac{|w^T x + b|}{\|w\|}$$

SVM: Finds a hyperplane with the maximum margin.

For a given h & dataset $S = \{x_1, \dots x_m\}$

$$\rho_h = \min_{i=1 \sim m} \rho_h(x_i)$$

The max-margin ρ is

$$\rho = \max_{w,b:\ y_i(w^T x_i + b) \geq 0} \ \min_{i \in \{1..m\}} \frac{|w^T x_i + b|}{\|w\|}$$

$$= \max_{w,b} \ \min_{i \in \{1..m\}} \frac{y_i(w^T x_i + b)}{\|w\|}$$

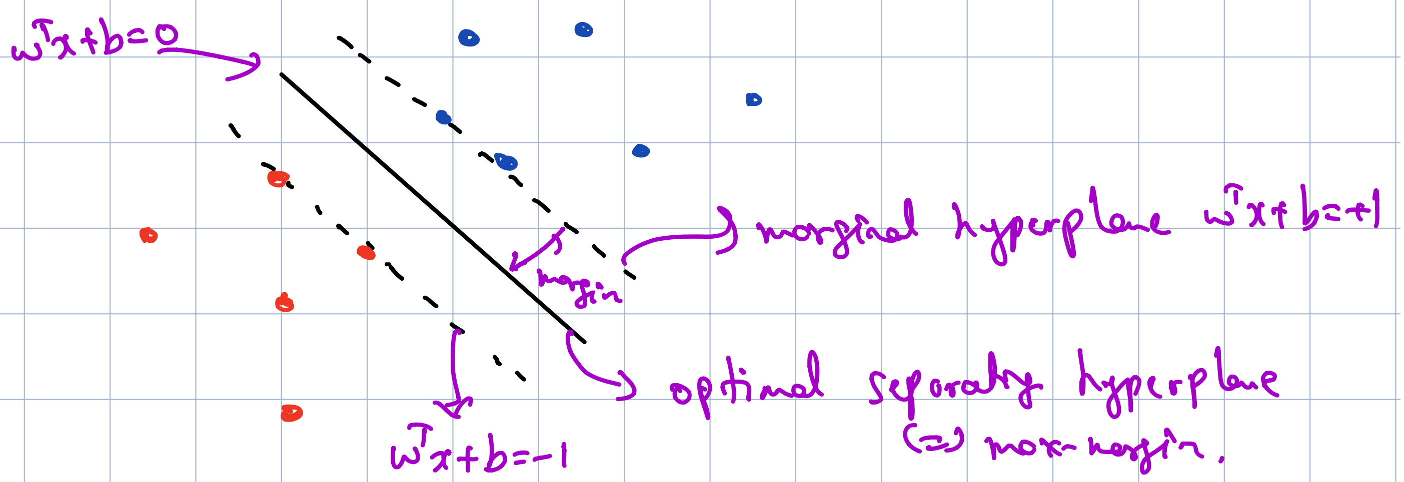
holds because the dataset is linearly separable, which implies $y_i(w^T x_i + b)$ is positive for the hyperplane defined by (w, b) that achieves the maximum.

It is enough to consider scaled (w,b) s.t.

$$\min_{i \in \{1,..m\}} y_i(w^T x_i + b) = 1$$

So,
$$\rho = \max_{\substack{w,b:\\ \min_{i \in \{1..m\}} y_i(w^T x_i + b) = 1}} \frac{1}{\|w\|}$$

$$= \max_{\substack{w,b:\\ \forall i,\ y_i(w^T x + b) \geq 1}} \frac{1}{\|w\|}$$



Optimization problem underlying the SVM method:

Maximizing $\rho \propto \frac{1}{\|w\|} \iff$ minimizing $\frac{1}{2}\|w\|^2$

s.t. $y_i(w^T x_i + b) \geq 1 \quad \forall i$

Primal problem:

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

Subject to $y_i(w^T x_i + b) \geq 1$, $i = 1 \dots m$

"Quadratic optimization problem with linear inequality constraints".

A brief tour of constrained optimization

Ref: Appendix B of FOML book

Constrained optimization problem:

$$\min_{x \in \mathcal{X}} f(x)$$

s.t. $g_i(x) \leq 0$, $i = 1 \dots m$

Lagrangian $L(x, \alpha_1, \ldots \alpha_m)$ is defined by

$$L(x, \alpha_1, \ldots \alpha_m) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x)$$

Dual function:

$$F(\alpha_1, \ldots \alpha_m) = \inf_{x \in X} L(x, \alpha_1, \ldots \alpha_m), \text{ for any } \alpha_1 \ldots \alpha_m \geq 0$$

$$= \inf_{x \in X} \left(f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) \right)$$

concave

Dual problem:-

$$\max_{\alpha_1 \ldots \alpha_m} F(\alpha_1 \ldots \alpha_m)$$

$$\text{s.t.} \quad \alpha_i \geq 0, \; i = 1 \ldots m$$

Karush-Kuhn-Tucker (KKT) conditions:

Assume $f, g_i, i = 1 \ldots m$ are convex & differentiable

" constraint qualification" holds ($\exists \bar{x} \in$ interior of X s.t. $g_i(\bar{x}) < 0 \; \forall i$)

$\bar{x}$ is a solution of the primal problem

if and only if

" $\exists (\bar{\lambda}_1, \dots \bar{\lambda}_m),\ \bar{\lambda}_i \geq 0, \forall i$ s.t.

$$\nabla_x L(\bar{x}, \bar{\lambda}_1, \dots \bar{\lambda}_m) = 0 \Leftrightarrow \nabla_x f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla_x g_i(\bar{x}) = 0$$

$g_i(\bar{x}) \leq 0,\ i = 1 \dots m$ "feasibility"

$\bar{\lambda}_i\, g_i(\bar{x}) = 0,\ i = 1 \dots m$ "complementary slackness"

Note: (1) Let p^* be the optimal value of the primal problem & d^* — " — " — dual problem

Weak duality: $d^* \leq p^*$

$(p^* - d^*) \leftarrow$ Duality gap

Strong duality: $d^* = p^*$ (= no duality gap)

If the primal problem is a convex optimization problem (= f, g_i are convex) & a constraint qualification holds (= $\exists\, \bar{x} \in \text{int}(X)$ s.t. $g_i(\bar{x}) < 0\ \forall i$), then "strong duality" holds.

(7) If the constraints are linear, then the constraint qualification holds.

Back to the SVM optimization problem in the linearly separable case:

Recall
$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1$$

Lagrangian:

$$L(w, b, \alpha_1, \dots, \alpha_m) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i (w^T x_i + b) - 1 \right)$$

KKT conditions:

$$(*)\begin{cases} \nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \implies w = \sum_{i=1}^{m} \alpha_i y_i x_i \\ \nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \implies \sum_{i=1}^{m} \alpha_i y_i = 0 \\ \forall i, \ \alpha_i \left(y_i (w^T x_i + b) - 1 \right) = 0 \implies \alpha_i = 0 \ \text{(or)} \ y_i(w^T x_i + b) = 1 \end{cases}$$

Observe that

optimal w is a linear combination of the data $x_1, \dots, x_m$.

In fact, let $S = \{ i \mid \alpha_i \neq 0 \}$

Then, from (*)

$$w = \sum_{i \in S} \alpha_i y_i x_i$$

Vectors x_i with $\alpha_i \neq 0$ are the support vectors.

From (*), for the support vectors $y_i (w^T x_i + b) = 1$.

(or) the support vectors lie on the marginal hyperplanes.

Also, $b = y_i - w^T x_i$, where x_i is a support vector

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j^T x_i)$$

Dual optimization problem:

$$F(\alpha_1, \dots, \alpha_m) = \inf_{w, b} \left[\frac{1}{2} \|w\|^2 - \sum_{i=1}^{m} \left(\alpha_i \left(y_i (w^T x_i + b) - 1 \right) \right) \right]$$

Note that if $\sum \alpha_i y_i \neq 0$, then $F(\alpha_1, \dots, \alpha_m) = -\infty$.

So, lets impose $\sum \alpha_i y_i = 0$

The infimum is attained at $w = \sum_{i=1}^{m} \alpha_i y_i x_i$

$$F(\alpha_1, \dots, \alpha_m) = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$
$$- \cancel{\sum_{i=1}^{m} \alpha_i y_i b}^{\,0} + \sum_{i=1}^{m} \alpha_i$$

$$F(\alpha_1, \dots, \alpha_m) = -\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_{i=1}^{m} \alpha_i$$

Dual optimization problem:-

$$(**) \begin{cases} \max\limits_{\alpha_1, \dots, \alpha_m} -\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_{i=1}^{m} \alpha_i \\ \text{s.t.} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \ i = 1 \dots m \end{cases}$$

$(**)$ is a constrained optimization problem, where the objective is quadratic and the constraints are linear.

The dimension of the dual problem is "m"

Can use QP solvers for $(**)$.

"Strong duality" holds. Solving (**) gives $\alpha_1, \dots, \alpha_m$, which can be used to determine the solution w, b of the primal problem, leading to the following SVM classifier:

$$h(x) = \operatorname{sign}(w^T x + b)$$

$$= \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i^T x) + b\right),$$

where $b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j^T x_i)$.

<u>Remark!</u> "Margin ρ of the SVM classifier".

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j^T x_i).$$

$$\sum_{i=1}^{m} \alpha_i y_i b = \sum_{i=1}^{m} \alpha_i y_i^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

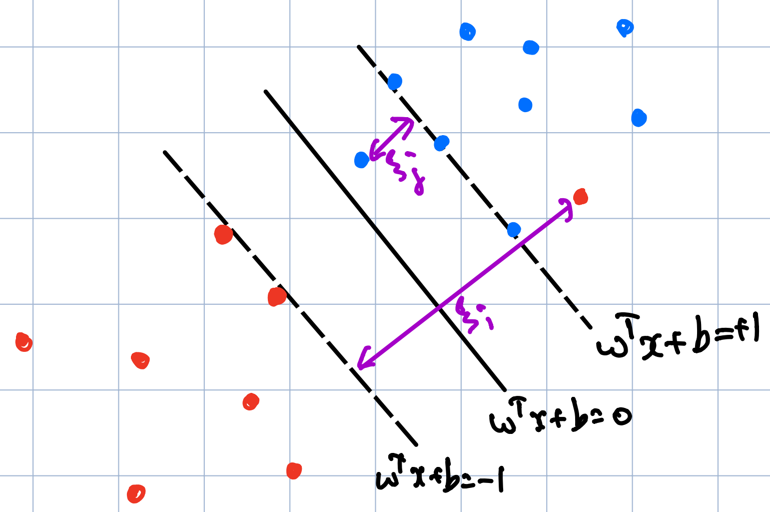
$$0 = \sum_{i=1}^{m} \alpha_i - \left(\sum_{i=1}^{m} \alpha_i y_i x_i\right)^T \left(\sum_{j=1}^{m} \alpha_j y_j x_j\right)$$

$$0 = \sum_{i=1}^{m} \alpha_i - \|w\|^2 \quad \text{— } (1)$$

used (1) & $\alpha_i \geq 0$

$$\rho^2 = \frac{1}{\|w\|^2} = \frac{1}{\sum_{i=1}^{m} \alpha_i}$$

SVM in the non-separable case (aka soft-margin SVM)



$S = \{x_1, \dots, x_m\}$

Training data is not linearly separable, which is equivalent to

$\forall$ hyperplanes $w^T x + b = 0$, $\exists\, x_i \in S$ s.t.

$$y_i (w^T x_i + b) \not\geq 1. \quad \text{——} (*)$$

A relaxed version of $(*)$ holds, i.e.,

$$\forall i = 1 \dots m, \ \exists\, \xi_i \geq 0 \quad \text{such that}$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i$$

$\xi_i \leftarrow$ slack variables & measure by how much a point x_i violates the separability constraint $y_i (w^T x + b) \geq 1$

If we ignore the outliers, then a margin of $\rho = \frac{1}{\|\omega\|}$ is achieved & hence, this margin is referred to as "soft margin".

Primal optimization problem:

$$(**)\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^p \\ \text{s.t.} & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1 \dots m \\ & \xi_i \geq 0, \quad i = 1 \dots m \end{cases}$$

Objective here is $\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^p$,

$\min \|w\|^2$ $\Leftrightarrow$ maximizing the margin

minimize the slack. (slack is due to outliers)

where $p \geq 1$

$(**)$ is a convex optimization problem

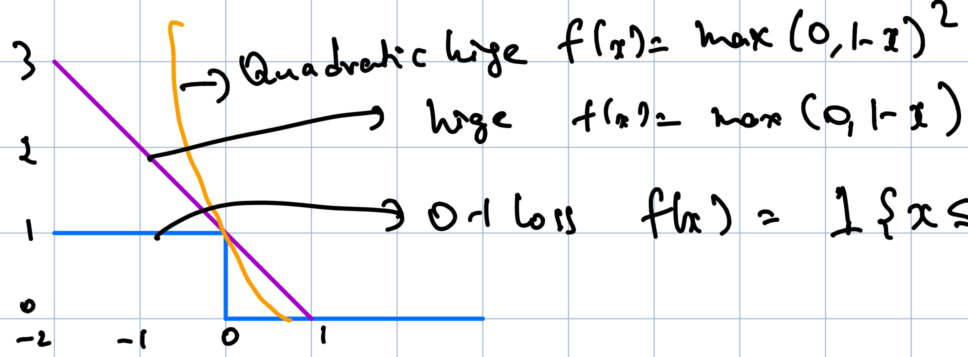
since $\xi \rightarrow \sum_{i=1}^{m} \xi_i^p = \|\xi\|_p^p$ $\quad \xi = (\xi_1, \dots, \xi_m)$

is a convex function

Choice of p:

$p = 1$ and $p = 2$

Hinge loss $\quad$ Quadratic hinge loss



Hinge loss is popular.

We focus on $p=1$.

allows trade off between max-margin & reducing slack

$$(\ast\ast) \rightarrow \begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} & \left.\begin{array}{l} y_i(w^T x_i + b) \ge 1 - \xi_i, \\ \xi_i \ge 0 \end{array}\right\} i = 1 \dots m \end{cases}$$

Recipe for solving $(\ast\ast)$ is

1. Apply KKT conditions
2. Work out dual function
3. formulate + Solve dual optimization problem
4. Use soln of dual problem to obtain the soft-margin SVM hyperplane classifier.

Lagrangian

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left(y_i(w^T x_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{m}\beta_i\xi_i$$

We apply KKT conditions

$$\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C$$

$$\forall i, \quad \alpha_i \left(y_i (w^T x_i + b) - 1 + \xi_i \right) = 0 \Rightarrow \alpha_i = 0 \text{ (or) } y_i (w^T x_i + b) = 1 - \xi_i$$

$$\forall i, \quad \beta_i \xi_i = 0 \Rightarrow \beta_i = 0 \text{ (or) } \xi_i = 0$$

Remarks:

① Optimal w has the same expression as before (= lin. sep. case)

② $w = \sum_{i=1}^{m} \alpha_i y_i x_i$

x_i appears only if $\alpha_i \neq 0$

Such x_i's are the support vectors.

For a support vector x_i, we have
$$y_i (w^T x_i + b) = 1 - \xi_i$$

If $\xi_i = 0$, then $y_i (w^T x_i + b) = 1$ & x_i is on one of the the marginal hyperplanes ($w^T x + b = \pm 1$)

If $\xi_i \neq 0$, then x_i is an outlier and $\beta_i = 0$ which implies $\alpha_i = C$.

Dual optimization problem:

$$F = \frac{1}{2}\left\|\sum_{i=1}^{m} \alpha_i y_i x_i\right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

$$- \cancel{\sum_{i=1}^{m} \alpha_i y_i b}^{\,0} + \sum_{i=1}^{m} \alpha_i$$

$$F = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

Dual function "is the same as" before (= lin. sep. case)

However, in addition to $\alpha_i \geq 0$ & $\sum \alpha_i y_i = 0$,

we require $\alpha_i \leq C$.

So, the dual optimization problem is

$$(***) \begin{cases} \max\limits_{\alpha_1 \dots \alpha_m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j) \\ \text{s.t.} \quad \alpha_i \geq 0, \ \alpha_i \leq C, \quad i = 1 \dots m \\ \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases}$$

$(***)$ is similar to $(**)$ (= dual problem in the separable case), with an additional constraint $\alpha_i \leq C$.

The complexity of solving (****) is comparable to that of (**), & one could use a QP solver in either case.

The solution $(\alpha_1, \ldots, \alpha_m)$ of (****) can be used to define the soft margin SVM hyperplane as follows:

$$h(x) = \text{sign}(w^T x + b)$$
$$= \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i^T x) + b \right),$$

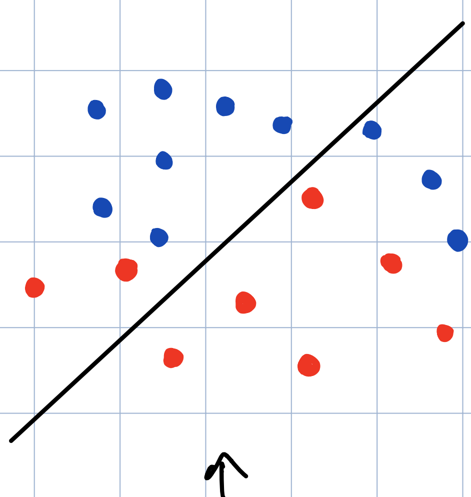
where

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j^T x_i),$$

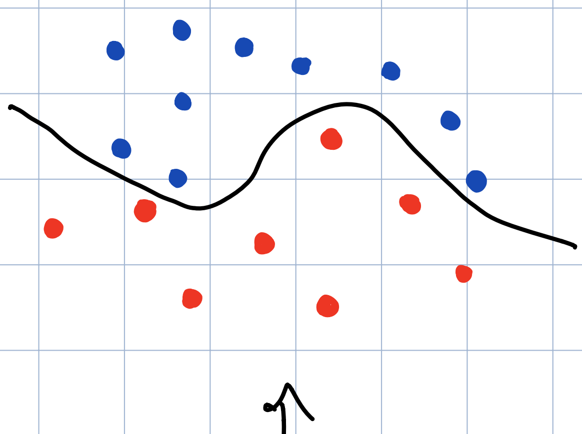
for any x_i with $0 < \alpha_i < C$.

Kernel methods

Ref: Chapter 6 of FoML book



"not" linearly separable



non-linear decision boundary

Idea: Take data (x_i)

$$\downarrow \Phi$$

High-dimensional space $\mathcal{H}$

& run SVMs in $\mathcal{H}$, since the data would be linearly separable in $\mathcal{H}$.

SVMs in high-dimensions:

$$h(x) = \operatorname{sign}(w^T x + b)$$

$$= \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \, \phi(x_i)^T \phi(x) + b \right),$$

is a high-dimensional feature mapping $\phi(x) \in \mathcal{H}$ → high-dimensional space.

where

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j \left(\phi(x_j)^T \phi(x_i) \right)$$

for any x_i with $0 < \alpha_i < C$.

Observe! feature x_i appear only through inner products. Using the kernel trick, the inner products can be computed efficiently even in high dimensions.

<u>Def (Kernel)</u>

input space (say e.g. some subset of $\mathbb{R}^d$)

A function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$\forall x, x' \in \mathcal{X}, \quad K(x, x') = \langle \phi(x), \phi(x') \rangle$$

for some function $\phi: \mathcal{X} \to \mathcal{H}$.

$H \to$ Hilbert space

(Hilbert space H is a vector space that is equipped with an inner product & is complete (= all Cauchy sequences converge).

The norm induced by the inner product is

$$\|x\|_H = \sqrt{\langle x, x\rangle} \quad \forall x \in H \quad)$$

Note: There are kernels, where $K(x, x')$ can be computed in $O(d)$ (d = input feature dimension), while computing $\langle \phi(x), \phi(x')\rangle$ is

$$O(\dim(H)), \text{ with } \dim(H) \gg d$$

Question: Given a function K, can we infer that K is a kernel, (or) equivalently there is a Hilbert space H & feature mapping ϕ s.t.

$$(*) \longrightarrow K(x, x') = \langle \phi(x), \phi(x')\rangle \, ?$$

Yes, if K is "positive definite symmetric" (PDS)

If K is PDS, then $\exists$ $\mathcal{H}$ & ϕ s.t. (*) holds.

So, one need not "define" or "compute" ϕ.

K being PDS ensures existence of a ϕ.

Definition (PDS)

A kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is PDS if for any $\{x_1, \ldots x_m\} \subseteq \mathcal{X}$, the matrix

$$\mathbf{K} = [[K(x_i, x_j)]]_{i,j=1\sim m}$$ is Symmetric positive semi-definite (SPSD)

Kernel matrix

$\mathbf{K}$ is SPSD if one of the following conditions holds:

1. Eigenvalues of $\mathbf{K}$ are non-negative
2. for any vector $c = (c_1, \ldots c_m)^T$,

$$c^T \mathbf{K} c = \sum_{i,j} c_i c_j K(x_i, x_j) \geq 0$$

Examples of PDS Kernels

① Polynomial Kernel

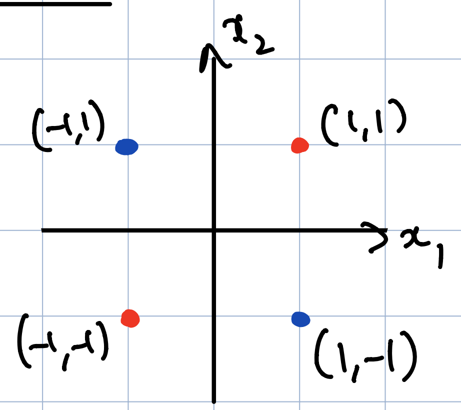
Fix a constant $c > 0$. A polynomial kernel of degree β is given by

$$K(x, x') = (x^T x' + c)^{\beta}, \quad \forall x, x' \in \mathbb{R}^d$$

Special case with $\beta = 2$, $d = 2$, $\forall\, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, x' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$

$$\begin{aligned}
K(x, x') &= (x^T x' + c)^2 \\
&= (x_1 x_1' + x_2 x_2' + c)^2 \\
&= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{bmatrix}^T \begin{bmatrix} x_1'^2 \\ x_2'^2 \\ \sqrt{2}\, x_1' x_2' \\ \sqrt{2c}\, x_1' \\ \sqrt{2c}\, x_2' \\ c \end{bmatrix} \\
&= \phi(x)^T \phi(x')
\end{aligned}$$

Achieving linear separation using a polynomial Kernel:



XOR problem
"not" linearly separable

Use a polynomial kernel with $d=2$, $c=1$

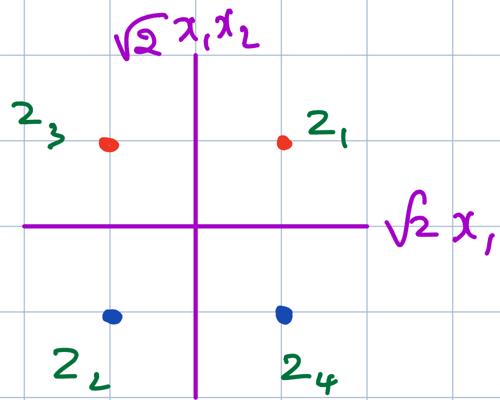
$(1,1) \to (1,1,\sqrt{2}, \sqrt{2}, \sqrt{2}, 1)\ z_1$

$(-1,1) \to (1,1,\sqrt{2}\ -\sqrt{2}, -\sqrt{2}, 1)\ z_2$

$(-1,-1) \to (1,1,-\sqrt{2}\ -\sqrt{2}, \sqrt{2}, 1)\ z_3$

$(1,-1) \to (1,1,-\sqrt{2}, \sqrt{2}, -\sqrt{2}, 1)\ z_4$

Linear separation using a polynomial kernel $\longrightarrow$



Note: Polynomial kernel is PDS since we wrote as a inner product with an explicit ϕ.

Example 2: Gaussian Kernels

$$\forall x, x' \in \mathbb{R}^d, \quad K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$$

Gaussian Kernel is PDS. (Can be shown using the normalization property of PDS Kernels)

Example 3: Sigmoid kernels

$$\forall x, x' \in \mathbb{R}^d, \quad K(x,x') = \tanh(a(x^T x') + b)$$

note: sigmoid kernels + SVM = simple neural network

Claim without proof:

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel.
Then, there exists a Hilbert space H and a mapping $\phi : \mathcal{X} \to H$ s.t.

$$\forall x, x' \in \mathcal{X}, \quad K(x,x') = \langle \phi(x), \phi(x') \rangle.$$

Verifying that the Gaussian kernel is PDS.

Normalized kernel K' associated with a PDS kernel K is

$$\forall x, x' \in \mathcal{X}, \quad K'(x,x') = \begin{cases} 0 & \text{if } K(x,x)=0 \text{ (or) } K(x',x')=0 \\ \dfrac{K(x,x')}{\sqrt{K(x,x)\,K(x',x')}} & \text{else} \end{cases}$$

By definition, $K'(x,x) = 1 \quad \forall x \in \mathcal{X}$.

A Gaussian Kernel can be seen as the normalized kernel of the kernel $K'(x,x') = \exp\left(\frac{x^T x'}{\sigma^2}\right)$

This can be argued as follows:

$$\frac{K'(x,x')}{\sqrt{K'(x,x)\,K'(x',x')}} = \frac{\exp\left(\frac{x^T x'}{\sigma^2}\right)}{\exp\left(\frac{\|x\|^2}{2\sigma^2}\right)\exp\left(\frac{\|x'\|^2}{2\sigma^2}\right)}$$

$$= \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$$

is the Gaussian Kernel (obtained by normalizing K')

Claim without proof: If K' is PDS, then its normalized variant is PDS as well.

To see that $K'(x,x') = \exp\left(\frac{x^T x'}{\sigma^2}\right)$ is PDS,

$$K'(x,x') = \sum_{j=0}^{\infty} \frac{(x^T x')^j}{\sigma^{2j}\, j!}$$

polynomial kernel

K' is a positive linear combination of polynomial kernels.

Claim: If K, K' are PDS kernels, then $K+K'$ is a PDS kernel.

Why: $c \in \mathbb{R}^{m \times 1}$

$$c^T K c \geq 0, \quad c^T K' c \geq 0 \quad \text{(since } K, K' \text{ are PDS)}$$

$$\Rightarrow \quad c^T (K + K') c \geq 0$$

Claim: PDS kernels closed under pointwise limit & composition with $f: x \to \sum_{j=0}^{\infty} a_j x^j, \; a_j \geq 0.$ ⟵ check this

Hence, Gaussian kernel is PDS.

SVMs with PDS Kernels:

Idea: Replace $x^T x'$ with $K(x, x')$

Dual optimization problem after incorporating a kernel is

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1 \ldots m$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

Solving the optimization problem would lead to the following classifier:

$$h(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b \right),$$

where $b = y_i - \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i)$

for x_i s.t. $0 < \alpha_i < C$

Kernel Ridge Regression

[Sec 11.3 of FOML book]

Input space $\mathcal{X} \subseteq \mathbb{R}^d$

Feature mapping $\phi : \mathcal{X} \to \mathbb{R}^{d'}$

Linear hypothesis set:

$$\{ h \mid h(x) = w^T \phi(x),\ w \in \mathbb{R}^{d'} \}$$

<u>Linear regression (recall)</u> $S = \{ (x_i, y_i),\ i = 1 \dots m \}$

$$\min_w \frac{1}{m} \sum_{i=1}^{m} (w^T \phi(x_i) - y_i)^2$$

$$= \min_w \left\{ J(w) := \frac{1}{m} \| A w - Y \|^2 \right\},$$

where A is the feature matrix with rows $\phi(x_i)^T$,

Y is a vector with components y_i

$J(w)$ is minimized by the solution to

$$A^T A W = A^T Y$$

<u>Ridge regression & its connection to Kernels:-</u>

$$\min_{w} \sum_{i=1}^{m} (w^T \phi(x_i) - y_i)^2 + \lambda \|w\|^2$$

$$= \min_{w} \underbrace{\|A w - Y\|^2 + \lambda \|w\|^2}_{J(w)}$$

$$\nabla J(w) = 0$$

$$\Rightarrow (A^T A + \lambda I) W = A^T Y$$

(or) $W = (A^T A + \lambda I)^{-1} A^T Y$ — (1)

$(A^T A + \lambda I)$ is invertible because $A^T A$ is positive semidefinite & $\lambda > 0$

An equivalent formulation for ridge regression:

$$\left.\begin{aligned} &\min_{w} \sum_{i=1}^{m} (w^T \phi(x_i) - y_i)^2 \\ &\text{subject to } \|w\|^2 \leq \lambda^2 \end{aligned}\right\} \rightarrow (01)$$

(O1) is a constrained optimization problem with convex objective aswell as convex constraints (why?)

(O1) can be re-written as

$$\left.\begin{aligned}&\min_{w} \sum_{i=1}^{m} \xi_i^2 \\ &\text{subject to} \quad \|w\|^2 \leq \Lambda^2, \text{ and} \\ &\qquad \xi_i = y_i - w^T \phi(x_i), \quad i = 1 \ldots m\end{aligned}\right\} - (O2)$$

(O2) → convex optimization problem

Lagrangian

$$\boxed{\begin{aligned}\xi &= (\xi_1 \ldots \xi_m)^T \\ \alpha' &= (\alpha_1', \ldots \alpha_m')^T\end{aligned}}$$

$L(\xi, w, \alpha', \lambda)$

$$= \sum_{i=1}^{m} \xi_i^2 + \sum_{i=1}^{m} \alpha_i' \left(y_i - \xi_i - w^T \phi(x_i)\right) + \lambda\left(\|w\|^2 - \Lambda^2\right)$$

Lagrange multipliers

Applying KKT conditions, we obtain

$$\nabla_w L = -\sum_{i=1}^{m} \alpha_i' \phi(x_i) + 2\lambda w = 0 \quad \Rightarrow \quad w = \frac{1}{2\lambda} \sum_{i=1}^{m} \alpha_i' \phi(x_i)$$

$$\nabla_{\xi_i} L = 2\xi_i - \alpha_i' = 0 \implies \xi_i = \frac{\alpha_i'}{2}$$

$$\alpha_i' \left(y_i - \xi_i - w^T \phi(x_i) \right) = 0, \; i = 1 \dots m$$

$$\lambda \left(\|w\|^2 - \Lambda^2 \right) = 0$$

Plugging in expressions for w & ξ_i from KKT conditions into the Lagrangian

$$\sum_{i=1}^m \frac{\alpha_i'^2}{4} + \sum_{i=1}^m \alpha_i' y_i - \sum_{i=1}^m \frac{\alpha_i'^2}{2} - \frac{1}{2\lambda} \sum_{i,j=1} \alpha_i' \alpha_j' \phi(x_i)^T \phi(x_j)$$
$$+ \lambda \left(\frac{1}{4\lambda^2} \left\| \sum_{i=1}^m \alpha_i' \phi(x_i) \right\|^2 - \Lambda^2 \right)$$

$$= -\frac{1}{4} \sum_{i=1}^m \alpha_i'^2 + \sum_{i=1}^m \alpha_i' y_i - \frac{1}{4\lambda} \sum_{i,j=1}^m \alpha_i' \alpha_j' \phi(x_i)^T \phi(x_j)$$
$$- \lambda \Lambda^2$$

Make the substitution $\alpha_i' = 2\lambda \alpha_i$,

$$= -\lambda^2 \sum_{i=1}^m \alpha_i^2 + 2\lambda \sum_{i=1}^m \alpha_i y_i - \lambda \sum_{i,j=1}^m \alpha_i \alpha_j \phi(x_i)^T \phi(x_j)$$
$$- \lambda \Lambda^2$$

The dual optimization problem is

$$\underset{\alpha \in \mathbb{R}^m}{\arg\max} \left(-\lambda^2 \sum_{i=1}^{m} \alpha_i^2 + 2\lambda \sum_{i=1}^{m} \alpha_i y_i - \lambda \sum_{i,j=1}^{m} \alpha_i \alpha_j \phi(x_i)^T \phi(x_j) - \lambda \Lambda^2 \right)$$

$$= \underset{\alpha \in \mathbb{R}^m}{\arg\max} \underbrace{\left(-\lambda \sum_{i=1}^{m} \alpha_i^2 + 2 \sum \alpha_i y_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j \phi(x_i)^T \phi(x_j) \right)}_{= G(\alpha)}$$

$$\max_{\alpha} \quad G(\alpha)$$

$$= \max_{\alpha} \quad -\lambda \alpha^T \alpha + 2 \alpha^T Y - \alpha^T (A A^T) \alpha$$

$$= \max_{\alpha} \quad -\alpha^T (K + \lambda I) \alpha + 2 \alpha^T Y$$

$K = AA^T$ is the Kernel matrix

Optimizing the dual

$$\nabla G(\alpha) = 0$$

$$\iff \quad 2 (K + \lambda I) \alpha = 2Y$$

$$\alpha = (K + \lambda I)^{-1} Y \quad \text{— (2)}$$

is invertible (why?)

Using KKT condition,

$$w = \sum_{i=1}^{m} \alpha_i \phi(x_i) = A^T\alpha = A^T(K+\lambda I)^{-1} Y$$

Linear hypothesis

$$h(x) = w^T \phi(x)$$

$$h(x) = \sum_{i=1}^{m} \alpha_i K(x_i, x)$$

Any PDS Kernel can be used to arrive at this predictor.

	Solving primal i.e., (1)	Solving dual, i.e., (2)
Computational cost	A^TA: $O(md'^2)$ $(A^TA+\lambda I)^{-1}$: $O(d'^3)$ multiply with A^T: $O(d')$	Let κ be the max cost of computing $K(x,x')$ $\forall x,x'$ Kernel matrix is computed in $O(\kappa m^2)$ Inverting $(K+\lambda I)$: $O(m^3)$ multiplication with Y: $O(m^2)$
Total cost:	$O(md'^2 + d'^3)$	Total cost: $O(\kappa m^2 + m^3)$

$$O((d')^3) \quad \text{vs} \quad O(m^3)$$

If ϕ is a mapping onto a high dimensional feature space & if the # of training samples is moderate, $d' >> m$, then solving the dual is computationally advantageous.

Prediction cost: $w^T \phi(x)$ computed in $O(d')$ for the primal

In case of the dual,

computing $(K(x_1, x), - - - -, K(x_m, x)) = \psi$

for a given x is $O(km)$

& $\psi^T \alpha$ is $O(m)$

So, total prediction cost $O(km)$