Minimize $f_0(x)$ **Convex Optimization**

Subject to constraints (i.e. x should be such that the below conditions hold)

$$f_i(x) \leq 0 \qquad i = 1, \dots, m$$

$$h_i(x) = 0 \qquad i = 1, \dots, p$$

(*)

Consider the Lagrangian

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

This gives us the Lagrange dual function:

$$g(\lambda, \nu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \nu)$$

D: domain of $f_0(x)$ under constraints

which gives us a lower bound on the minimum of (*)

$$g(\lambda, \nu) \leq f_0(x^*)$$

whenever $\lambda \succeq 0$ (easy to prove)

So, we now replace our original difficult minimization problem (*) with an easier maximization of this lower bound to get as close as possible to the the minimum value $f_0(x^*)$. This max. problem is called the Lagrange dual problem:

$$\text{maximize } g(\lambda, \nu)$$
$$\text{subject to } \lambda \succeq 0$$

(**)

$\Longrightarrow$ This is a convex optimization problem!

($x \succeq 0$ simply means all components of vector x are ≥ 0)

The optimal solution (λ^*, ν^*) is <u>dual optimal</u>

Any pair (λ, ν) s.t. $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$ is <u>dual feasible</u>

~~As~~ we ~~say~~ stated above (easily proveable), <u>weak duality</u> always holds:

$$g(\lambda^*, \nu^*) \leq f_0(x^*)$$

But sometimes, <u>strong duality</u> holds!

$$g(\lambda^*, \nu^*) = f_0(x^*)$$

This holds whenever <u>constraint qualifications</u> are satisfied. One such example is:

sufficient, not necessary condition for strong duality

1. Primal problem is convex, i.e. $h(x) = Ax - b = 0$, (equality constraints is affine) $p = 1$
2. <u>Slater's condition</u> holds: there exists some (<u>strictly feasible</u>) point $\tilde{x}$ s.t. $f_i(\tilde{x}) < 0 \ \forall i$ and $A\tilde{x} = b$

If the objective f_0 and constraint f_i, h_i functions are differentiable, Slater's condition holds, and strong duality holds, then the <u>KKT conditions</u> are necessary and sufficient for global optimality

- $f_i(x) \leq 0$, $h_i(x) = 0$, $\lambda_i \geq 0$
- $\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$

(i.e. if you solve for x such that the KKT conditions hold, then you are at the global optimum)

- $\underbrace{\lambda_i f_i(x) = 0}$

↳ this condition is called <u>complimentary slackness</u> and it follows from strong duality:

$$f_0(x^*) = g(\lambda^*, \nu^*) = \inf_{x \in D} \left(f_0(x) + \sum \lambda_i^* f_i(x) + \sum \nu_i^* h_i(x) \right) \leq f_0(x^*) + \sum \lambda_i^* f_i(x^*) + \sum \nu_i^* h_i(x^*)$$

$$\therefore \sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0 \iff \begin{cases} \lambda_i^* > 0 \Rightarrow f_i(x^*) = 0 \\ f_i(x^*) < 0 \Rightarrow \lambda_i^* = 0 \end{cases}$$

~~Support Vector Machines~~

<u>Representer Theorem</u>: Suppose we have a set of data points $\{(x_i, y_i)\}_{i=1}^{N}$ and we want to find the function/ input-output mapping $f(\cdot)$ that minimizes the loss function:

$$f^* = \arg\min_{f \in \mathcal{H}} L_y(f(x_1), \dots, f(x_N)) + \Omega\left(\|f\|_{\mathcal{H}}^2\right)$$

where $\Omega(\cdot)$ is non-decreasing and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ parameterizes $L_y(\cdot)$. Note that L_y depends on x_i's only via $f(x_i)$. For example, in ridge regression $L_y(f(x_1), \dots, f(x_N)) = \frac{1}{2}\sum_{i=1}^{N}(y_i - f(x_i))^2$ and $\Omega(\|f\|_{\mathcal{H}}^2) = \lambda\|f\|_{\mathcal{H}}^2$. The theorem now tells us that a solution to this minimization takes the form:

$$f^* = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot)$$

if $\Omega(\cdot)$ is strictly increasing.

<u>Pf.</u> Let $f^* = f_s + f_\perp$, where f_s is the projection of f^* onto the subspace spanned by $\{K(x_i, \cdot)\}_{i=1}^{N}$ and $f_\perp$ is the orthogonal error relative to f.

First note that

$$L_y(f(x_1), \ldots, f(x_N))$$

$$= L_y(\langle f, K(x_1, \cdot)\rangle, \ldots, \langle f, K(x_N, \cdot)\rangle)$$

$$= L_y(\langle f_s + f_\perp, K(x_1, \cdot)\rangle, \ldots, \langle f_s + f_\perp, K(x_N, \cdot)\rangle)$$

$$= L_y(\langle f_s, K(x_1, \cdot)\rangle, \ldots)$$

$$= L_y(f_s(x_1), \ldots, f_s(x_N))$$

So, minimizing L_y w.r.t. f_s is the same as minimizing w.r.t. f: we can forget $f_\perp$ without losing anything.

Now note that f_s is in fact the minimum of $\Omega(\|f\|_{\mathcal{H}}^2)$ if it is non-decreasing, since in this case,

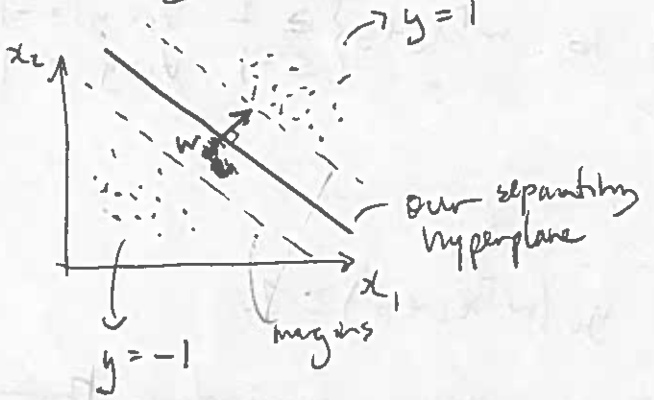
$$\Omega(\|f\|_{\mathcal{H}}^2) = \Omega(\|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2) \geq \Omega(\|f_s\|_{\mathcal{H}}^2)$$

Thus, this component is minimized when $\|f_\perp\|_{\mathcal{H}}^2 = 0$, leaving the unique (only unique whenever $\Omega(\cdot)$ is strictly increasing) solution

$$\underline{f = f_s} = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot)$$

Support Vector Classification

The problem is to find a hyperplane that separates the data correctly according to some classification criteria. Formally, what we want is a hyperplane such that the scalar projection $w^T x_i$ of all data points onto the direction w perpendicular to it gives us the correct classification:



$$y_i = \text{sign}(w^T x_i + b)$$

We can find the best such hyperplane by maximizing the minimum distance b/w it and each class ($y=+1, y=-1$), i.e. maximizing the margin. We can compute this by considering a pair of points of different classes x^+, x^- lying on each margin: these will be at the minimum distance from the hyperplane, which

~~we can impose that this minimum distance be 1, measured by the scalar projection onto w:~~

$$\cancel{w^T x^+ + b = \min(w^T x_i + b) = 1 \quad \forall x_i : y_i = +1}$$

$$\cancel{w^T x^- + b = \max(w^T x_i + b) = -1 \quad \forall}$$

can be computed via $\dfrac{x^{+T} w}{\|w\|}, \dfrac{x^{-T} w}{\|w\|}$

Since x^+ is of class $y=+1$ and x^- of class $y=-1$, we know that

$w^T x^+ + b \geq 0, \quad w^T x^- + b < 0.$

In fact, to ensure accuracy of our classifier we are going to enforce choice of w, b such that

$w^T x_i + b \geq 1 \quad \forall i : y_i = 1$ and $w^T x_i + b \leq -1 \quad \forall i : y_i = -1$

~~where the inequalities~~ become equalities ~~for~~ points on the margins ~~that are closest~~ to the hyperplane.

Maximizing the minimum distance b/w classes and the hyperplane thus entails maximizing the distance b/w margins which we now have is

$$\frac{x^{+T} w}{\|w\|} - \frac{x_i^{-T} w}{\|w\|} = \frac{(1-b) - (-1-b)}{\|w\|} = \frac{2}{\|w\|}$$

~~Our~~ job has thus become solving the following optimization problem:

$$\text{maximize} \quad \frac{2}{\|w\|} \quad \text{subject to} \quad w^T x_i + b \begin{cases} \geq 1 & \forall i : y_i = +1 \\ \leq -1 & \forall i : y_i = -1 \end{cases}$$

which can be rewritten as

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i (w^T x_i + b) \geq 1$$

However, it will rarely be possible to find a hyperplane that perfectly separates the ~~two~~ classes, so we soften the ~~[illegible]~~ constraint and modify our objective to include a trade-off (controlled by C) with errors (i.e. data points within the margins or on the wrong side of the hyperplane):

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i \right) \quad \text{subject to} \quad y_i (w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

This gives us the following Lagrangian:

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

Noting that each of our constraints $f_i(x) = 1 - \xi_i - (w^T x_i + b) y_i \leq 0$, $g_i(\xi_i) = -\xi_i \leq 0$ are convex, and that there always exists ~~some~~ x, ξ that satisfies them (i.e. Slater's condition holds), we ~~[illegible]~~ have that strong duality holds. Therefore, we need only solve for the KKT conditions to get the global optimum.

1) $\lambda_i \geq 0, \quad \alpha_i \leq 0$

2) $\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \iff \underline{\underline{w = \sum_{i=1}^{N} \alpha_i y_i x_i}}$

$\frac{\partial \mathcal{L}}{\partial b} = \underline{\underline{\sum_{i=1}^{N} \alpha_i y_i = 0}}$

$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \iff \underline{\underline{\alpha_i = C - \lambda_i}}$

$\implies \underline{\underline{\alpha_i \leq C}}$ since $\lambda_i \geq 0$

3) (complementary slackness)

For $\alpha_i = C \ (\neq 0)$,

$\lambda_i = 0 \implies \xi_i \geq 0$ → x_i lies inside (~~[illegible]~~ the margins)

$1 - (w^T x_i + b) y_i - \xi_i = 0 \iff y_i(w^T x_i + b) = 1 - \xi_i$

For $0 < \alpha_i < C$,

$\lambda_i > 0 \implies \xi_i = 0$ → (x_i lies on margin)

as in first case, $y_i(w^T x_i + b) = 1 - \xi_i = 1$

For $\alpha_i = 0$

$\lambda_i > 0 \implies \xi_i = 0$ →

$y_i(w^T x_i + b) \geq 1$ (x_i's correctly outside the margins)

In other words, we find that our solution for α is such that

- it is sparse: only points on the margin or w/ ~~error~~ large error (i.e. inside the margins) have $\alpha_i \geq 0$
- only those points contribute to the ~~support vectors~~ $w = \sum_i \alpha_i y_i x_i$
- the contribution of ~~error~~ large error x_i's is bounded by C

Thus, these are called the support vectors

We ~~[illegible]~~ can now solve for the support vector by ~~[illegible]~~ maximizing the dual $g(\alpha)$ with respect to α. We first express the full dual ~~[illegible]~~ $g(\alpha, \lambda)$ in terms of just α, which we can do given our KKT conditions we derived above:

$$g(\alpha, \lambda) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i + \sum_i \alpha_i (1 - y_i(w^T x_i + b) - \xi_i) + \sum_i \lambda_i(-\xi_i)$$

$$= \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j + C\sum_i \xi_i + \sum_i \alpha_i - \sum_i \alpha_i \xi_i$$

$$- \sum_i \alpha_i y_i x_i \sum_j \alpha_j y_j x_j^T - b\underbrace{\sum_i \alpha_i y_i}_{=0} - \sum_i (C - \alpha_i)\xi_i$$

$$= -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i + \cancel{C\sum_i \xi_i} - \cancel{\sum_i (C - \alpha_i + \alpha_i)\xi_i}$$

$$= g(\alpha)$$

We now simply minimize $g(\alpha)$ subject to the constraints ~~[illegible]~~

$$0 \le \alpha_i \le C$$

$$\sum_i \alpha_i y_i = 0$$

which is a quadratic program. The resulting solution then gives us the support vector w by our equation derived above. We get b by solving the equation $y_i(w^T x_i + b) = 1$ for an x_i on the margin or by averaging the solutions for all x_i on the margins.

ν – SVM

We can also give an alternative formulation of the problem that yields more interpretable parameters (as opposed to C, which is rather opaque). The following formulation is called ν – SVM:

$$\min_{w, \rho, \xi} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \xi_i \right) \quad \text{subject to} \quad \begin{array}{l} \rho \ge 0 \\ \xi_i \ge 0 \\ y_i(w^T x_i) \ge \rho - \xi_i \end{array}$$

where we have dropped the offset b purely for simplicity.

The resulting Lagrangian is:

$$\mathcal{L}(w, \nu, \rho, \{\xi_i\}, \alpha, \lambda, \gamma)$$

$$= \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i(\rho - y_i w^T x_i - \xi_i) + \sum_{i=1}^{N} \lambda_i(-\xi_i) + \overset{-\gamma\rho}{\downarrow}$$

where we can interpret the new parameter ρ as the margin width we want to optimize, along with the support vector w and the errors ξ_i. We now follow the same exercise as above, first writing out the KKT conditions after noting that again strong duality holds and then writing out the dual function:

1) $\alpha_i \geq 0, \ \lambda_i \geq 0, \ \gamma \geq 0$

2) $\dfrac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \iff \underline{\underline{w = \sum_{i=1}^{N} \alpha_i x_i y_i}}$

$\dfrac{\partial \mathcal{L}}{\partial \xi_i} = \dfrac{1}{N} - \alpha_i - \lambda_i = 0 \iff \underline{\underline{\alpha_i + \lambda_i = \dfrac{1}{N}}}$

$\dfrac{\partial \mathcal{L}}{\partial \rho} = -\nu + \sum_{i=1}^{N} \alpha_i - \gamma = 0 \iff \underline{\underline{\nu = \sum_{i=1}^{N} \alpha_i - \gamma}}$

3) Complementary slackness. ~~Lets~~ assume that $\rho > 0$ to consider only non-trivial cases w.r.t. our new parameter ν. By complementary slackness, this implies that $\gamma = 0$, which implies that $\nu = \sum_{i=1}^{N} \alpha_i$. Now we consider two cases for ξ_i:

For $\xi_i > 0$:

$\Rightarrow \lambda_i = 0 \iff \alpha_i = \dfrac{1}{N}$

then, for all such points $N(\alpha)$

$$\sum_{i \in N(\alpha)} \alpha_i = \frac{|N(\alpha)|}{N} \leq \sum_{i=1}^{N} \alpha_i = \nu$$

Noting that $N(\alpha)$ is the set of all points that fall inside the margins, we can interpret ν as an upper bound on the number of such 'errors'.

For $\xi_i = 0$,

$\lambda_i > 0 \iff \alpha_i < \frac{1}{N}$ $\quad \sum_{i \in N(\alpha)} \alpha_i + \sum_{i \in M(\alpha)} \alpha_i < \frac{|N(\alpha)| + |M(\alpha)|}{N} \leq \nu$

$\Rightarrow$ ν is upper bound on total # of support vectors w/ non-zero weight

Let $M(\alpha)$ be the set of points such that $0 < \alpha_i < \frac{1}{N}$, i.e. the points with $\xi_i = 0$ that still contribute to w (i.e. $\alpha_i \neq 0$)

The dual function is then:

$$g(\alpha) = \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j - \nu\rho + \frac{1}{N}\sum_i \xi_i + \sum_i \alpha_i \rho - \sum_i \alpha_i \xi_i - \sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$- \sum_i \left(\frac{1}{N} - \alpha_i\right)\xi_i + \left(\nu - \sum_i \alpha_i\right)\rho = -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

So we now maximize $g(\alpha)$ subject to: $\sum_{i=1}^{N} \alpha_i \geq \nu$

$$0 \leq \alpha_i \leq \frac{1}{N}$$

Kernelized SVM

We can easily accomodate a kernelized solution to the problem by recognizing the form* of the objective function being minimized and invoking the representer theorem, telling us that $w = \sum_{i=1}^{N} \beta_i K(x_i, \cdot)$. We can thus interpret the minimization of $\|w\|_{\mathcal{H}}^2$ (i.e. the maximization of the margin) as enforcing smoothness of the function $w \in \mathcal{H}$.

Our objective function in terms of ξ_i thus becomes (again dropping b for simplicity)

$$\min_{\beta,\xi}\left(\frac{1}{2}\beta^T K \beta + C\sum_{i=1}^{N}\xi_i\right) \quad \text{subject to} \quad \xi_i \geq 0$$

$$y_i \sum_{j=1}^{N} \beta_j K(x_j, x_i) \geq 1 - \xi_i$$

Since K is positive definite, this objective is convex and strong duality holds, giving the dual function

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

which we maximize subject to $0 \leq \alpha_i \leq C$.

* In fact, to see this we need to put our objective in the form

$\frac{1}{2}\|w\|_{\mathcal{H}}^2 + C\sum_{i=1}^{N}[1 - y_i\langle w, K(x_i,\cdot)\rangle_{\mathcal{H}}]_+$ ← rectification to invoke the representer theorem.

This is equivalent to the form in terms of ξ_i, just harder to minimize b/c of the non-linearity