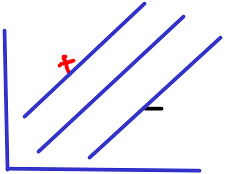
Lecture 26: Support Vector Classification, Unsupervised Learning

Instructor: Prof. Ganesh Ramakrishnan

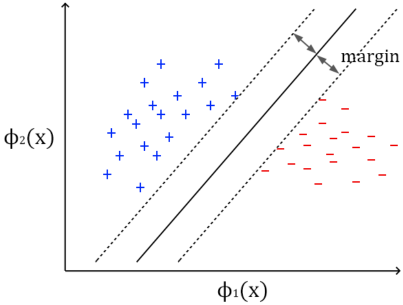
Support Vector Classification

- Perceptron does not find the *best* seperating hyperplane, it finds *any* seperating hyperplane.
- In case the initial $\mathbf{w}$ does not classify all the examples, the seperating hyperplane corresponding to the final $\mathbf{w}^*$ will often pass through an example.
- The seperating hyperplane does not provide enough breathing space – this is what SVMs address and we already saw that for regression!

- Perceptron does not find the *best* seperating hyperplane, it finds *any* seperating hyperplane.
- In case the initial $\mathbf{w}$ does not classify all the examples, the seperating hyperplane corresponding to the final $\mathbf{w}^*$ will often pass through an example.
- The seperating hyperplane does not provide enough breathing space – this is what SVMs address and we already saw that for regression!
  - **We now quickly do the same for classification** (is asymmetric in pts unlike regression)



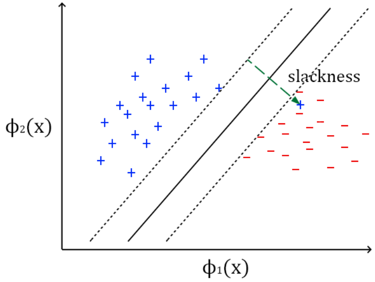
Support Vector Classification: Separable Case



$$\mathbf{w}^\top \phi(\mathbf{x}) + b \geq +1 \textit{ for } y = +1$$
$$\mathbf{w}^\top \phi(\mathbf{x}) + b \leq -1 \textit{ for } y = -1$$
$$\mathbf{w}, \phi \in \mathbb{R}^m$$

There is large margin to seperate the +ve and -ve examples

Support Vector Classification: Non-separable Case

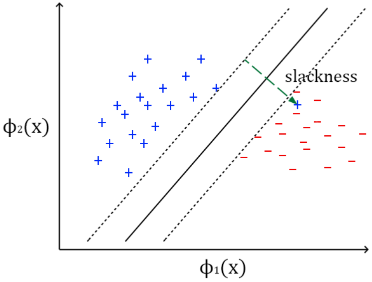


When the examples are not linearly seperable, we need to consider the slackness ξ_i (always $+$ve) of each example $\mathbf{x}^{(i)}$ (how far a misclassified point is from the seperating hyperplane):

$$w^T\phi(x^i) + b \geq 1 - \xi_i \quad \forall\; y^i = +1$$

$$w^T\phi(x^i) + b \leq -1 - \xi_i \quad \forall\; y^i = -1$$

Support Vector Classification: Non-separable Case



When the examples are not linearly seperable, we need to consider the slackness ξ_i (always +ve) of each example $\mathbf{x}^{(i)}$ (how far a misclassified point is from the seperating hyperplane):

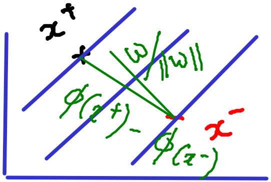
$$\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b \geq +1 - \xi_i \ (\textit{for } y^{(i)} = +1)$$
$$\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b \leq -1 + \xi_i \ (\textit{for } y^{(i)} = -1)$$

Multiplying $y^{(i)}$ on both sides, we get:

$$y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi_i, \ \forall i = 1, \ldots, n$$

Maximize the margin

Note: x^+ & x^- are imaginary



- We maximize the margin $(\phi(\mathbf{x}^+) - \phi(\mathbf{x}^-))^\top[\frac{\mathbf{w}}{\|\mathbf{w}\|}] = \frac{2}{\|w\|}$
- Here, $\mathbf{x}^+$ and $\mathbf{x}^-$ lie on boundaries of the margin.
- Recall that $\mathbf{w}$ is perpendicular to the separating surface
- We project the vectors $\phi(\mathbf{x}^+)$ and $\phi(\mathbf{x}^-)$ on $\mathbf{w}$, and normalize by $\mathbf{w}$ as we are only concerned with the direction of $\mathbf{w}$ and not its magnitude

(1) $w^T\phi(x^+) + b = 1$

(2) $w^T\phi(x^-) + b = -1$

$\Rightarrow$ (1) − (2)

$w^T(\phi(x^+) - \phi(x^-)) = 2$

$\Rightarrow (\phi(x^+) - \phi(x^-))^T \frac{w}{\|w\|} = \frac{2}{\|w\|}$

Simplifying the margin expression

- Maximize the margin $(\phi(\mathbf{x}^+) - \phi(\mathbf{x}^-))^\top [\frac{\mathbf{w}}{\|\mathbf{w}\|}]$
- At $\mathbf{x}^+$: $y^+ = 1$, $\xi^+ = 0$ hence, $(\mathbf{w}^\top \phi(\mathbf{x}^+) + b) = 1$ — (1)
  At $\mathbf{x}^-$: $y^- = 1$, $\xi^- = 0$ hence, $-(\mathbf{w}^\top \phi(\mathbf{x}^-) + b) = 1$ — (2)

Simplifying the margin expression

- Maximize the margin $(\phi(\mathbf{x}^+) - \phi(\mathbf{x}^-))^\top [\frac{\mathbf{w}}{\|\mathbf{w}\|}]$
- At $\mathbf{x}^+$: $y^+ = 1$, $\xi^+ = 0$ hence, $(\mathbf{w}^\top \phi(\mathbf{x}^+) + b) = 1$ — (1)
  At $\mathbf{x}^-$: $y^- = 1$, $\xi^- = 0$ hence, $-(\mathbf{w}^\top \phi(\mathbf{x}^-) + b) = 1$ — (2)
- Adding (2) to (1),
  $\mathbf{w}^\top(\phi(\mathbf{x}^+) - \phi(\mathbf{x}^-)) = 2$
- *Thus, the margin expression to maximize is:* $\frac{2}{\|\mathbf{w}\|}$

Formulating the objective

- Problem at hand: Find $\mathbf{w}^*, b^*$ that maximize the margin.
- $(\mathbf{w}^*, b^*) = \arg\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$
  s.t. $y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi_i$ and
  $\xi_i \geq 0, \forall i = 1, \ldots, n$
- However, as $\xi_i \to \infty$, $1 - \xi_i \to -\infty$

} Problem as $\xi_i \to \infty$ constraints easily satisfied even with $\|w\| \to 0$

And: $\|w\| \to 0$ will maximize the objective

Q: What are possible fixes?

(a) $\xi_i \leq \theta$

(b) Add $-C \sum \xi_i$ to objective

(c) Add $-C \sum \xi_i^2$ to objective & drop $\xi_i \geq 0$

Formulating the objective

- Problem at hand: Find $\mathbf{w}^*, b^*$ that maximize the margin.
- $(\mathbf{w}^*, b^*) = \arg\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$
  s.t. $y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi_i$ and
  $\xi_i \geq 0, \forall i = 1, \ldots, n$

  } & minimize $\Sigma \xi_i$
- However, as $\xi_i \to \infty$, $1 - \xi_i \to -\infty$
- Thus, with arbitrarily large values of ξ_i, the constraints become easily satisfiable for any $\mathbf{w}$, which defeats the purpose.
- *Hence, we also want to minimize the ξ_i's. E.g.*, minimize $\sum \xi_i$

Objective

- $(\mathbf{w}^*, b^*, \xi_i^*) = \arg\min_{\mathbf{w}, b, \xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$ (→ margin, → error)

  s.t. $y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi_i$ and

  $\xi_i \geq 0, \forall i = 1, \ldots, n$
- Instead of maximizing $\frac{2}{\|\mathbf{w}\|}$, minimize $\frac{1}{2}\|\mathbf{w}\|^2$

  ($\frac{1}{2}\|\mathbf{w}\|^2$ *is monotonically decreasing with respect to* $\frac{2}{\|\mathbf{w}\|}$)
- C determines the trade-off between the error $\sum \xi_i$ and the margin $\frac{2}{\|\mathbf{w}\|}$

Support Vector Machines

Dual Objective

2 Approaches to Showing Kernelized Form for Dual

Banking on $\|w\|_2^2 = \langle w, w \rangle$ [Not possible for L_1 norm $\|w\|_1$]

1. **Approach 1:** The Reproducing Kernel Hilbert Space and Representer theorem (Generalized from derivation of Kernel Logistic Regression, Tutorial 7, Problem 3)
   See `http://qwone.com/~jason/writing/kernel.pdf` for list of kernelized objectives
2. **Approach 2:** Derive using First principles (provided for completeness in Tutorial 9)

Approach 1: Special case of Representer Theorem & Reproducing Kernel Hilbert Space (RKHS)

1. Generalized from derivation of Kernel Logistic Regression, Tutorial 7, Problem 3. See `http://qwone.com/~jason/writing/kernel.pdf` for list of kernelized objectives
2. Let $\mathcal{X}$ be the space of examples such that $\mathcal{D} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\right\} \subseteq \mathcal{X}$ and for any $\mathbf{x} \in \mathcal{X}$, $K(., \mathbf{x}) : \mathcal{X} \to \Re$ → Can also view it as $K_x(\cdot) : \mathcal{X} \to \mathbb{R}$
3. (Optional)[1] The solution $f^* \in \mathcal{H}$ (Hilbert space) to the following problem → Treat it as a space where $\|f\|^2 = \langle f, f \rangle$

$f \in \{w^T\phi(x) + b\}$

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{m} \mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \Omega(\|f\|_K)$$

error of (x^i, y^i) — Regularizer

eg: $\Omega = \frac{1}{2}(.)^2$

can be always written as $f^*(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$ provided $\Omega(\|f\|_K)$ is a monotonically increasing function of $\|f\|_K$. $\mathcal{H}$ is the Hilbert space and $K(., \mathbf{x}) : \mathcal{X} \to \Re$ is called the **Reproducing (RKHS) Kernel**

$\Omega(\|f\|)$ can be measured through Params such as w

[1] Proof provided in optional slide deck at the end

Intuition about "Reproducing" kernel

$$\phi_K(x) = [K(x, x^1), K(x, x^2) \, - - - \, K(x, x^m)]$$

$$\langle \phi_K(x), \phi_K(y) \rangle = K(x, y)$$

Approach 1: Special case of Representer Theorem & Reproducing Kernel Hilbert Space (RKHS)

1. (Optional) The solution $f^* \in \mathcal{H}$ (Hilbert space) to the following problem

Specific version

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{m} \mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \Omega(\|f\|_K)$$

can be always written as $f^*(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(\|f\|_K)$ is a ....

2. More specifically, if $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$ and $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$ then the solution $\mathbf{w}^* \in \Re^n$ to the following problem

$b \in \mathbb{R}$

Can I eliminate SVC constraints?

SVC: $\frac{1}{2}\|w\|_2^2$

$$(\mathbf{w}^*, b^*) = \underset{\mathbf{w}, b}{\operatorname{argmin}} \sum_{i=1}^{m} \mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \Omega(\|\mathbf{w}\|_2)$$

can be always written as $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(\|\mathbf{w}\|_2)$ is a monotonically increasing function of $\|\mathbf{w}\|_2$. $\Re^{n+1}$ is the Hilbert space and $K(., \mathbf{x}) : \mathcal{X} \to \Re$ is the **Reproducing (RKHS) Kernel** (b could be also pushed into w)

The Representer Theorem and SVC

1. The SVC Objective

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg\min_{\mathbf{w},b,\xi_i} C\sum_{i=1}^{m} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2$$

s.t. $y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi_i$ and
$\xi_i \geq 0, \forall i = 1, \ldots, m$

form so far → Can we get into $\sum_i E(f(x^i), y^i)$

2. Can be rewritten as

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg\min_{\mathbf{w},b,\xi_i} C\sum_{i=1}^{m} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2$$

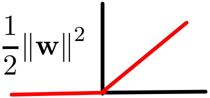
s.t. $\max\left(1 - y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b), 0\right) = \xi_i$

$\therefore -y^i(w^T\phi(x^i) + b) + 1 \leq \xi_i$
$0 \leq \xi_i$

3. That is,

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg\min_{\mathbf{w},b,\xi_i} C\sum_{i=1}^{m} \max\left(1 - y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b), 0\right) + \frac{1}{2}\|\mathbf{w}\|^2$$

Hinge loss



The Representer Theorem and SVC (contd.)

1. If $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ and $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$ and given the SVC objective

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg\min_{\mathbf{w}, b, \xi_i} C\sum_{i=1}^{m} \max\left(1 - y^{(i)}(\mathbf{w}^\top\phi(\mathbf{x}^{(i)}) + b), 0\right) + \frac{1}{2}\|\mathbf{w}\|^2$$

2. setting $\mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) = C\max\left(1 - y^{(i)}(\mathbf{w}^\top\phi(\mathbf{x}^{(i)}) + b), 0\right)$ and $\Omega(\|\mathbf{w}\|_2) = \frac{1}{2}\|\mathbf{w}\|_2^2$, we can apply the Representer theorem to SVC, so that $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$

Approach 2: Derivation using First principles

Derivation similar to that for Support Vector Regression, and provided for completeness in extra slide deck as well as in Tutorial 9

- The dual optimization problem becomes:

$$\max_{\alpha} -\frac{1}{2}\sum_{i}\sum_{j}\alpha_i\alpha_j y^{(i)}y^{(j)}K\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)+\sum_{i}\alpha_i$$

s.t.
$\alpha_i \in [0, C]$, $\forall i$ and
$\sum_i \alpha_i y^{(i)} = 0$

SVR $\sum_i (\alpha_i - \alpha_i^*) y^i = 0$

Recall for SVR $(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)$ is now simply $\alpha_i\alpha_j$

For SVR was $\sum_i (\alpha_i - \alpha_i^*)$

Representer Theorem and RKHS

Dual Objective

(Optional)

The main idea

We first recap the main optimization problem

$$E(\mathbf{w}) = -\left[\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\mathbf{w}^T\phi(\mathbf{x}^{(i)}) - \log\left(1+\exp\left(\mathbf{w}^T\phi(\mathbf{x}^{(i)})\right)\right)\right)\right] + \frac{\lambda}{2m}||\mathbf{w}||^2 \tag{1}$$

and an expression for $\mathbf{w}$ at optimality

$$\mathbf{w} = \frac{1}{\lambda}\left[\sum_{i=1}^{m}\left(y^{(i)} - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\phi(\mathbf{x}^{(i)})\right] \tag{2}$$

To completely prove this specific case of KLR, let $\mathcal{X}$ be the space of examples such that $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\right\} \subseteq \mathcal{X}$ and for any $\mathbf{x} \in \mathcal{X}$, $K(., \mathbf{x}) : \mathcal{X} \to \Re$ be a function such that $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$. Recall that $\phi(\mathbf{x}) \in \Re^n$ and

$$f_{\mathbf{w}}(\mathbf{x}) = p(Y=1|\phi(\mathbf{x})) = \frac{1}{1+\exp\left(-\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x})\right)}$$

For the rest of the discussion, we are interested in viewing $-\mathbf{w}^T\phi(\mathbf{x})$ as a function $h(\mathbf{x})$

The Reproducing Kernel Hilbert Space (RKHS)

Consider the set of functions $\mathcal{K} = \left\{K(.,\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\right\}$ and let $\mathcal{H}$ be the set of all functions that are **finite** linear combinations of functions in $\mathcal{K}$. That is, any function $h \in \mathcal{H}$ can be written as $\mathbf{h}(.) = \sum_{t=1}^{T} \alpha_t K(.,\mathbf{x}_t)$ for some T and $\mathbf{x}_t \in \mathcal{X}, \alpha_t \in \Re$. One can easily verify that $\mathcal{H}$ is a vector space[2]

Note that, in the special case when $f(\mathbf{x}') = K(\mathbf{x}',\mathbf{x})$, then $T = m$ and

$$f(\mathbf{x}') = K(\mathbf{x}',\mathbf{x}) = \sum_{i=1}^{n} \phi_i(\mathbf{x}') K(\mathbf{e}_i,\mathbf{x})$$

where $\mathbf{e}_i$ is such that $\phi(\mathbf{e}_i) = \mathbf{u}_i \in \Re^n$, the unit vector along the i^{th} direction.
Also, by the same token, if $\mathbf{w} \in \Re^n$ is in the search space of the regularized cross-entropy loss function (**??**), then

$$\phi^{\mathbf{T}}(\mathbf{x}')\mathbf{w} = \sum_{i=1}^{n} w_i K(\mathbf{e}_i,\mathbf{x})$$

Thus, the solution to (**??**) is an $h \in \mathcal{H}$.

Inner Product over RKHS $\mathcal{H}$

For any $g(.) = \sum_{t=1}^{S} \beta_s K(., \mathbf{x}'_s) \in \mathcal{H}$ and $h(.) = \sum_{t=1}^{T} \alpha_t K(., \mathbf{x}_t) \in \mathcal{H}$, define the inner product[3]

$$\langle h, g \rangle = \sum_{s=1}^{S} \beta_s \sum_{t=1}^{T} \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) \tag{4}$$

Further simplifying (4),

$$\langle h, g \rangle = \sum_{s=1}^{S} \beta_s \sum_{t=1}^{T} \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) = \sum_{s=1}^{S} \beta_s f(\mathbf{x}_s) \tag{5}$$

One immediately observes that in the special case that $g() = K(., \mathbf{x})$,

$$\langle h, K(., \mathbf{x}) \rangle = h(\mathbf{x}) \tag{6}$$

[3] Again, you can verify that $\langle f, g \rangle$ is indeed an inner product following properties such as symmetry, linearity in the first argument and positive-definiteness: `https://en.wikipedia.org/wiki/Inner_product_space`

Orthogonal Decomposition

Since $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\right\} \subseteq \mathcal{X}$ and $\mathcal{K} = \left\{K(., \mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\right\}$ with $\mathcal{H}$ being the set of all finite linear combinations of function in $\mathcal{K}$, we also have that

$$lin\_span\left\{K(., \mathbf{x}^{(1)}), K(.\mathbf{x}^{(2)}), \ldots, K(., \mathbf{x}^{(m)})\right\} \subseteq \mathcal{H}$$

Thus, we can use orthogonal projection to decompose any $h \in \mathcal{H}$ into a sum of two functions, one lying in $lin\_span\left\{K(., \mathbf{x}^{(1)}), K(.\mathbf{x}^{(2)}), \ldots, K(., \mathbf{x}^{(m)})\right\}$, and the other lying in the orthogonal complement:

$$h = h^{\|} + h^{\perp} = \sum_{i=1}^{m} \alpha_i K(., \mathbf{x}^{(i)}) + h^{\perp} \tag{7}$$

where $\langle K(., \mathbf{x}^{(i)}), h^{\perp}\rangle = 0$, for each $i = [1..m]$.
For a specific training point $\mathbf{x}^{(j)}$, substituting from (7) into (6) for any $h \in \mathcal{H}$, using the fact that $\langle K(., \mathbf{x}^{(i)}), h^{\perp}\rangle = 0$

$$h(\mathbf{x}^{(j)}) = \langle \sum^{m} \alpha_i K(., \mathbf{x}^{(i)}) + h^{\perp}, K(., \mathbf{x}^{(j)})\rangle = \sum^{m} \alpha_i \langle K(., \mathbf{x}^{(i)}), K(., \mathbf{x}^{(j)})\rangle = \sum^{m} \alpha_i K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$$

Analysis for the Empirical Risk

The Regularized Cross-Entropy Logistic Loss (1), has two parts (after ignoring the common $\frac{1}{m}$ factor), *viz.*, the **empirical risk**

$$-\left[\sum_{i=1}^{m}\left(y^{(i)}\mathbf{w}^{T}\phi(\mathbf{x}^{(i)}) - \log\left(1+\exp\left(\mathbf{w}^{T}\mathbf{x}^{(i)}\right)\right)\right)\right] \tag{9}$$

Since the **empirical risk** in (9) is only a function of $h(\mathbf{x}^{(i)}) = \mathbf{w}^{T}\phi(\mathbf{x}^{(i)})$ for $i = [1..m]$, based on (8) we note that the value of the **empirical risk** in (9) will therefore be independent of $h^{\perp}$ and therefore one only needs to equivalently solve the following **empirical risk** by substituting from (8) *i.e.*, $h(\mathbf{x}^{(j)}) = \sum_{i=1}^{m}\alpha_i K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$:

$$\left[\sum_{\mathbf{i=1}}^{\mathbf{m}}\left(\sum_{\mathbf{j=1}}^{\mathbf{m}} -\mathbf{y}^{(\mathbf{i})}\mathbf{K}\left(\mathbf{x}^{(\mathbf{i})}, \mathbf{x}^{(\mathbf{j})}\right)\alpha_{\mathbf{j}}\right) + \log\left(\mathbf{1} + \sum_{\mathbf{j=1}}^{\mathbf{m}}\alpha_{\mathbf{j}}\mathbf{K}\left(\mathbf{x}^{(\mathbf{i})}, \mathbf{x}^{(\mathbf{j})}\right)\right)\right]$$

Analysis with Regularizer

Consider the regularizer function $||\mathbf{w}||_2^2$ which is a strictly monotonically increasing function of $||\mathbf{w}||$. Substituting $\mathbf{w} = \frac{1}{\lambda}\left[\sum_{i=1}^{m}\left(y^{(i)} - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\phi(\mathbf{x}^{(i)})\right]$ from (**??**), one can view $\Omega(||h||)$ as a strictly monotonic function of $||h||$.

$$\Omega(||h||) = \Omega\left(||\sum_{i=1}^{m}\alpha_i K(.,\mathbf{x}^{(i)}) + h^{\perp}||\right) = \Omega\left(\sqrt{||\sum_{i=1}^{m}\alpha_i K(.,\mathbf{x}^{(i)})||^2 + ||h^{\perp}||^2}\right)$$

and therefore,

$$\Omega(||h||) = \Omega\left(\sqrt{||\sum_{i=1}^{m}\alpha_i K(.,\mathbf{x}^{(i)})||^2 + ||h^{\perp}||^2}\right) \geq \Omega\left(\sqrt{||\sum_{i=1}^{m}\alpha_i K(.,\mathbf{x}^{(i)})||^2}\right)$$

That is, setting $h^{\perp} = 0$ does not affect the first term of (1) while strictly increasing the second term. That is, any minimizer must have optimal $h^*(.)$ with $h^{\perp} = 0$. That is,

Derivation of SVM Dual using First Principles (also included in Tutorial 9)

Dual Objective

Dual function

- Let $L^*(\alpha, \mu) = \min_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, \mu)$
- By weak duality theorem, we have:
  $L^*(\alpha, \mu) \leq \min_{\mathbf{w}, b, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$
  s.t. $y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b) \geq 1 - \xi_i$, and
  $\xi_i \geq 0, \forall i = 1, \ldots, n$
- The above is true for any $\alpha_i \geq 0$ and $\mu_i \geq 0$
- Thus,

$$\max_{\alpha, \mu} L^*(\alpha, \mu) \leq \min_{\mathbf{w}, b, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

Dual objective

- In case of SVM, we have a strictly convex objective and linear constraints – therefore, strong duality holds:

$$\max_{\alpha,\mu} L^*(\alpha, \mu) = \min_{w,b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

- This value is precisely obtained at the $(\mathbf{w}^*, b^*, \xi^*, \alpha^*, \mu^*)$ that satisfies the necessary (and sufficient) optimality conditions
- Assuming that the necessary and sufficient conditions (KKT or Karush–Kuhn–Tucker conditions) hold, our objective becomes:

$$\max_{\alpha,\mu} L^*(\alpha, \mu)$$

- $L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y^{(i)}(\mathbf{w}^\top \phi(\mathbf{x}^{(i)}) + b)) - \sum_{i=1}^{n} \mu_i \xi_i$
- We obtain $\mathbf{w}$, b, ξ in terms of α and μ by setting $\nabla_{w,b,\xi} L = 0$:
  - **w.r.t. $\mathbf{w}$:** $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \phi(\mathbf{x}^{(i)})$
  - **w.r.t. b:** $-b \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$
  - **w.r.t. ξ_i:** $\alpha_i + \mu_i = C$
- Thus, we get:
  $L(w,b,\xi,\alpha,\mu)$
  $= \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} \phi^\top(\mathbf{x}^{(i)})\phi(\mathbf{x}^{(j)}) + C\sum_i \xi_i + \sum_i \alpha_i - \sum_i \alpha_i \xi_i -$
  $\sum_i \alpha_i y^{(i)} \sum_j \alpha_j y^{(j)} \phi^\top(\mathbf{x}^{(j)})\phi(\mathbf{x}^{(i)}) - b\sum_i \alpha_i y^{(i)} - \sum_i \mu_i \xi_i$
  $= -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} \phi^\top(\mathbf{x}^{(i)})\phi(\mathbf{x}^{(j)}) + \sum_i \alpha_i$

- The dual optimization problem becomes:

$$\max_{\alpha} -\frac{1}{2}\sum_{i}\sum_{j}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi^\top(\mathbf{x}^{(i)})\phi(\mathbf{x}^{(j)}) + \sum_{i}\alpha_i$$

  s.t.
  $\alpha_i \in [0, C], \forall i$ and
  $\sum_i \alpha_i y^{(i)} = 0$
- Deriving this did not require the complementary slackness conditions
- Conveniently, we also end up getting rid of μ