Machine Learning and Neural Networks (MATH3431) Epiphany term, 2025

Lecture notes 8: Support Vector Machines

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce the Support Vector Machines as a procedure. Motivation, set-up, description, computation, and implementation. We focus on the classical treatment.

Reading list & references:

(1) Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
- Ch. 15 (pp. 167-170, 171-172, 176-177) Support Vector Machine

(2) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
- Ch. 7.1 Sparse Kernel Machines/Maximum marginal classifiers

(3) Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.

(4) Boyd, S. P., & Vandenberghe, L. (2004). Convex optimization. Cambridge university press.
- Ch. 4, 5

1. Intro and motivation

Note 1. Support Vector Machines (SVM) is a ML procedure for learning linear predictors in high-dimensional feature spaces with regards the sample complexity challenges. Due to a duality property, SVM have sparse solutions, so that predictions for new inputs depend only on quantities evaluated at a subset of the whole training dataset.

Definition 2. Let $w \neq 0$. **Hyperplane** in space $\mathcal{X} \subseteq \mathbb{R}^d$ is called the sub-set

$$S = \left\{ x \in \mathbb{R}^d : \langle w, x \rangle + b = 0 \right\}. \tag{1.1}$$

Note 3. Hyperplane (1.1) separates $\mathcal{X}$ in two half-spaces

$$S_+ = \left\{ x \in \mathbb{R}^d : \langle w, x \rangle + b > 0 \right\}$$

and

$$S_- = \left\{ x \in \mathbb{R}^d : \langle w, x \rangle + b < 0 \right\}$$

Definition 4. Halfspace (hypothesis space) is hypotheses class $\mathcal{H}$ designed for binary classification problems, $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$ defined as

$$\mathcal{H} = \left\{ x \longmapsto \text{sign}\left(\langle w, x \rangle + b \right) : w \in \mathbb{R}^d, b \in \mathbb{R} \right\},$$

where b is called bias.

Definition 5. Each **halfspace hypothesis** $h \in \mathcal{H}$ has form

$$h_{w,b}(x) = \text{sign}(\langle w, x\rangle + b) . \tag{1.2}$$

It takes an input in $\mathcal{X} \subseteq \mathbb{R}^d$ and returns an output in $\mathcal{Y} = \{-1, +1\}$. We may refer to it as halfspace (w, b) as well.

Note 6. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be a training set of examples with $x_i \in \mathbb{R}^d$ the features and $y_i \in \{-1, +1\}$ the labels.

Note 7. Our goal is to train a halfspace hypothesis $h_{w,b}(x)$ in (1.2) against a training dataset S, with puppose to be able to classify a future feature x as $y = -1$ or $y = 1$.

Definition 8. The training set S is **linearly separable** if there exists a halfspace (w, b) such that for all $i = 1, ..., n$

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

or equivalently

$$y_i (\langle w, x_i\rangle + b) > 0 \tag{1.3}$$

Note 9. Halfspaces (w^*, b^*) satisfying the linearly separable condition (1.3) are ERM hypothesis under the $0-1$ loss function $\ell^{0-1}((w, b), z) = 1_{(y_i \neq \text{sign}(\langle w, x_i\rangle + b))}$ and Empirical Risk $R_S^{0-1}(w, b) = \frac{1}{m}\sum_{i=1}^m \ell((w, b), z_i) \geq 0$, i.e.

$$(w^*, b^*) = \underset{w,b}{\arg\min}\left(R_S^{0-1}(w, b)\right) = \underset{w,b}{\arg\min}\left(\frac{1}{m}\sum_{i=1}^m \ell^{0-1}((w, b), z_i)\right)$$

as $R_S^{0-1}(w^*, b^*) = 0$.

Definition 10. Margin of a hyperplane with respect to a training set is defined to be the minimal distance between a point in the training set and the hyperplane.

Note 11. There are several different halfspaces (w, b) satisfying (1.3) for the same linearly separable training dataset S; see Figure (1.1; left). In Figure (1.1; right) the margin γ is the distance from the hyperplane (solid line) to the closest points in either class (which touch the parallel dotted lines). Among the hyperplanes satisfying (1.3), a reasonable/desirable halfspace (w, b) is the one with the maximum margin γ; see Figure (1.1; right). The rational is that if a hyperplane has a large margin, then it will still separate the training set even if we slightly perturb each instance.

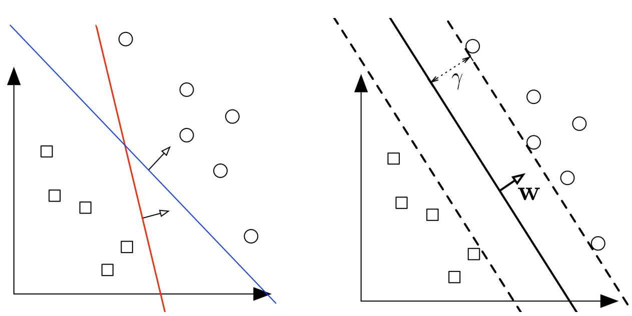


Figure 1.1

Note 12. Support Vector Machines (SVM) aims at learning the maximum margin separating hyperplane Figure (1.1; Right).

2. Hard Support Vector Machine

Assumption 13. *Assume the training sample $S = \{(x_i, y_i)\}_{i=1}^m$ is linearly separable.*

Definition 14. Hard Support Vector Machine (Hard-SVM) is the learning rule in which we return an ERM hyperplane that separates the training set with the largest possible margin given Assumption 13.

Problem 15. (Hard-SVM) Given a linearly separable training sample $S = \{(x_i, y_i)\}_{i=1}^m$ the Hard-SVM rule for the binary classification problem is the solution to the quadratic optimization problem:

$$\text{Solve} \quad \left(\tilde{w}, \tilde{b}\right) = \arg\min_{(w,b)} \frac{1}{2} \|w\|_2^2 \tag{2.1}$$

$$\text{subject to: } y_i \left(\langle w, x_i\rangle + b\right) \geq 1, \ \forall i = 1, ..., m \tag{2.2}$$

Scale

$$\hat{w} = \frac{\tilde{w}}{\|\tilde{w}\|}, \text{ and } \hat{b} = \frac{\tilde{b}}{\|\tilde{w}\|}$$

Note 16. Following we show why Problem 15 produces a Hard-SVM hyperplane stated in Note 14.

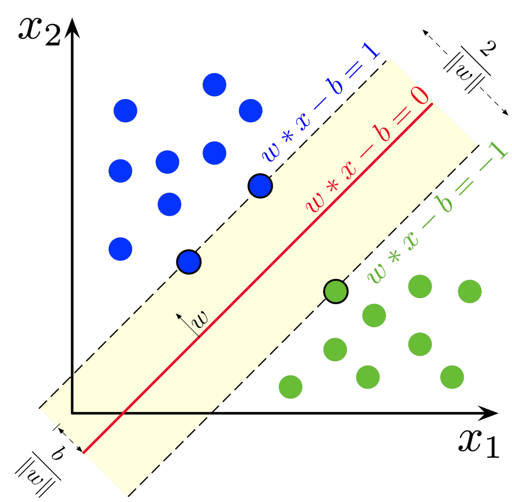
Lemma 17. *The distance between a point x and the hyperplane defined by (w, b) with is $|\langle w, x\rangle + b| / \|w\|$.*

Proof. We skip it. □

Note 18. On the right, see the geometry of Problem 15.

Note 19. Hard-SVM selects two parallel hyperplanes that separate the two classes of data so that the distance between them is as large as possible. The predictive hyperplane (rule) is the hyperplane that lies halfway between them.

Note 20. Hard-SVM in Problem 15 searches for the hyperplane with minimum norm w among all those that separate the data and have distance greater or equal to 1.



Proof. (Sketch of the proof of Problem 15)

(1) Based on Note 14, and Lemma 17, the closest point in the training set to the separating hyperplane has distance

$$\min_i \left(|\langle w, x_i\rangle + b| / \|w\|\right) \tag{2.3}$$

Without loss of generality, we make (2.3) identifiable by just picking those with $\|w\| = 1$. Hence, by Definition 14, the Hard-SVM hypothesis should be such as

$$(w^*, b^*) = \underset{(w,b):\|w\|=1}{\arg\max} \left(\min_i \left(|\langle w, x_i \rangle + b| \right) \right) \tag{2.4}$$

$$\text{subject to } \; y_i \left(\langle w, x_i \rangle + b \right) > 0, \; \forall i = 1, ..., m \tag{2.5}$$

(2) If there is a solution in (2.4) (namely, linearly separable dataset), then (2.4) is equivalent to (proof is omitted)

$$(w^*, b^*) = \underset{(w,b):\|w\|=1}{\arg\max} \left(\min_i \left(y_i \left(\langle w, x_i \rangle + b \right) \right) \right) \tag{2.6}$$

(3) Next we show that (2.6) is equivalent to the solution of Problem 15; i.e. $(w^*, b^*) = \left(\hat{w}, \hat{b} \right)$.
Let $\gamma^* := \min_i \left(|\langle w^*, x_i \rangle + b^*| \right)$ be the margin from (w^*, b^*). Firstly, because

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

$\left(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*} \right)$ satisfies conditions (2.1) and (2.2). Secondly, I have $\|\tilde{w}\| \leq \left\| \frac{w^*}{\gamma^*} \right\| = \frac{1}{\gamma^*}$ because of (2.1) and because of $\|w^*\| = 1$. Hence, for all $i = 1, ..., m$, it is

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

Hence $\left(\hat{w}, \hat{b} \right)$ is the optimal solution of (2.6).

□

Definition 21. Homogeneous halfspaces in SVM is the case where the halfspaces pass from the origin; that is when the bias term in (2.2) is zero $b = 0$.

3. Soft Support Vector Machine

Note 22. Hard-SVM assumes the strong Assumption 13 that the training set is linearly separable. This might not always the case, and hence there is need to derive a procedure that weakens this assumption.

Note 23. Soft Support Vector Machine (Soft-SVM) aims to relax the strong assumption of Hard-SVM that the training set is linearly separable (2.5) with purpose to extend the scope of application. Soft-SVM does not assume Assumption 13. Soft-SVM rule is given as the solution to the quadratic optimization problem 24.

Problem 24. (Soft-SVM) Given a training sample $S = \{(x_i, y_i)\}_{i=1}^m$ the Soft-SVM rule for the binary classification learning problem is solution to the quadratic optimization Problem

$$\text{Solve} \quad (w^*, b^*, \xi^*) = \underset{(w,b,\xi)}{\arg\min} \left(\frac{1}{2} \|w\|_2^2 + C \frac{1}{m} \sum_{i=1}^{m} \xi_i \right) \tag{3.1}$$

$$\text{subject to: } y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \ \forall i = 1, ..., m \tag{3.2}$$

$$\xi_i \geq 0, \ \forall i = 1, ..., m \tag{3.3}$$

Note 25. To relax the linearly separable training set Assumption 13, Soft-SVM relies on replacing the "harder" constraint (2.2) with the "softer" one in (3.2) through the introduction of non-negative unknown quantities (slack variables) $\{\xi_i\}_{i=1}^m$ controlling how much the separability assumption (2.2) is violated. Because any point that is misclassified has $\xi_i > 1$, it follows that $\sum_{i=1}^m \xi_i$ is an upper bound on the number of misclassified points.

Note 26. Parameter $C > 0$ controls the trade-off between the slack variable penalty and the margin. It controls the trade-off between minimizing training errors and controlling model complexity.

Note 27. Soft-SVM learns all (w, b, ξ) via the minimization part in (3.1) where the trade off between the two terms is controlled via the user specified parameter C.

Note 28. Proposition 29 shows that the Soft-SVM is a binary classification learning problem under the hinge loss function and with a regularization term biasing toward low norm separators.

Proposition 29. *The solution of Problem 24 is equivalent to the Ridge regularized loss minimization problem*

$$(w^*, b^*) = \underset{(w,b)}{\arg\min} \left(R_S^{hinge} + \lambda \|w\|_2^2 \right) \tag{3.4}$$

with the hinge loss function

$$\ell^{hinge}((w, b), z) = \max(0, 1 - y(\langle w, x \rangle + b))$$

Empirical Risk Function

$$R_S^{hinge}((w, b)) = \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(\langle w, x_i \rangle + b))$$

regularization parameter $\lambda = \frac{1}{2C}$ *and regularization term* $\|\cdot\|_2^2$.

Proof. In Algorithm 24, we consider (3.1) as

$$\underset{(w,b)}{\arg\min} \left(\min_{\xi} \left(\lambda \|w\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right) \right) \tag{3.5}$$

with $\lambda = \frac{1}{2C}$. Consider (w, b) fixed and focus on the inside minimization. From (3.2), it is $\xi_i \geq 1 - y_i (\langle w^*, x_i \rangle + b^*)$, and from (3.3), it is $\xi_i \geq 0$. If $y_i (\langle w, x_i \rangle + b) \geq 1$, the best assignment in 3.5 is $\xi_i = 0$ because it is $\xi_i \geq 0$ from (3.3) and I need to minimize (3.5) wrt ξ_i's. If $y_i (\langle w, x_i \rangle + b) \leq 1$, the best assignment in (3.5) is $\xi_i = 1 - y_i (\langle w, x_i \rangle + b)$ because I need to minimize w.r.t ξ_i's. Hence $\xi_i = \max(0, 1 - y_i (\langle w, x_i \rangle + b))$. □

Example 30. Consider the Soft-SVM as a Ridge regularized loss minimization learning problem in (3.4) on a training dataset $S = \{(x_i, y_i)\}_{i=1}^{m}$ with $z_i = (x_i, y_i) \overset{\text{ind}}{\sim} g$ where g is a data generating process on $\mathcal{X} \times \{0,1\}$ with $\mathcal{X} = \{x : \|x\| \le \rho\}$. Let $\mathfrak{A}(S)$ be the solution of the Soft-SVM. Then:

- Because $\ell^{\text{hinge}}((w,b), z) = \max(0, 1 - y(\langle w, x\rangle + b))$ is $\|x\|$-Lipschitz, it is
$$\mathrm{E}_{S\sim g}\left(R_g^{\text{hinge}}(\mathfrak{A}(S))\right) \le R_g^{\text{hinge}}(u) + \lambda \|u\|^2 + \frac{2\rho^2}{\lambda m}$$
[Hint: Note 43 Handout 3: Learnability and stability in learning problems]
- If we assume a bounded learning problem, i.e. $\mathcal{H} = \{\|(w,b)\| \le B\}$ for $B > 0$ then
$$\mathrm{E}_{S\sim g}\left(R_g^{\text{hinge}}(\mathfrak{A}(S))\right) \le \min_{\|(w,b)\|\le B} R_g^{\text{hinge}}((w,b)) + \lambda B^2 + \frac{2\rho^2}{\lambda m}$$
- If we choose $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ then
$$\mathrm{E}_{S\sim g}\left(R_g^{\text{hinge}}(\mathfrak{A}(S))\right) \le \min_{\|(w,b)\|\le B} R_g^{\text{hinge}}((w,b)) + \sqrt{\frac{8\rho^2 B^2}{m}}$$
- Assume one was interested in minimizing a $0-1$ Risk $R_g^{0-1}(\cdot)$ under the $0-1$ loss $\ell^{0-1}((w,b), z) := 1_{(y\langle w,x\rangle \le 0)}$. But $\ell^{0-1}(\cdot, z)$ is non-convex for all z. It is $\ell^{0-1}((w,b), z) \le \ell^{\text{hinge}}((w,b), z)$ for all z. By using the convex hinge loss ℓ^{hinge} as a surrogate loss and implementing the Soft-SVM procedure, the error under $0-1$ Risk $R_g^{0-1}(\cdot)$ is
$$\mathrm{E}_{S\sim g}\left(R_g^{0-1}(\mathfrak{A}(S))\right) \le \mathrm{E}_{S\sim g}\left(R_g^{\text{hinge}}(\mathfrak{A}(S))\right) \le \min_{\|(w,b)\|\le B} R_g^{\text{hinge}}((w,b)) + \sqrt{\frac{8\rho^2 B^2}{m}}$$

Example 31. Given Proposition 29, Soft-SVM in Problem 24 can be learned via any variation of SGD, eg online SGD (batch size $m = 1$) with recursion
$$\varpi^{(t+1)} = \varpi^{(t)} - \eta_t v_t$$
$$\begin{bmatrix} w^{(t+1)} \\ b^{(t+1)} \end{bmatrix} = \begin{bmatrix} w^{(t)} \\ b^{(t)} \end{bmatrix} - \eta_t \begin{bmatrix} v_{w,t} \\ v_{b,t} \end{bmatrix}$$
where $v_{w,t} = \begin{cases} -y_i^{(t)} x_i^{(t)} & \text{if } y_i^{(t)}\left(\langle w^{(t)}, x_i^{(t)}\rangle + b^{(t)}\right) < 1 \\ 0 & \text{otherwise} \end{cases}$ and $v_{b,t} = \begin{cases} -y_i^{(t)} & \text{if } y_i^{(t)}\left(\langle w^{(t)}, x_i^{(t)}\rangle + b^{(t)}\right) < 1 \\ 0 & \text{otherwise} \end{cases}$.

4. Duality and sparsity

Recal Lagrange mukltipliers and duality from previous years

4.1. Lagrangian duality (recall from previous years).

Notation 32. Let $f : \mathbb{R}^q \to \mathbb{R}$ be an objective function, let $\{g_i : \mathbb{R}^q \to \mathbb{R}\}_{i=1}^{m}$ inequality constraint convex functions, and let $\{h_j : \mathbb{R}^q \to \mathbb{R}\}_{j=1}^{n}$ equality constraint functions.

Definition 33. (Primal problem) Consider we have the following minimization problem that we will call Primal problem

$$p^* = \min_x (f(x)) \quad \text{s.t.}\, g_i(x) \le 0,\ \forall i = 1, ..., m \quad h_j(x) = 0,\ \forall j = 1, ..., n \tag{4.1}$$

Definition 34. (Lagrangian function) To the Primal problem 33, we associate the Lagrangian function $L : \mathbb{R}^q \times \mathbb{R}^m \times \mathbb{R}^n$ with

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{n} \beta_j h_j(x) \tag{4.2}$$

Note 35. Note that the Primal problem is equivalent to

$$p^* = \min_x \left(\max_{\alpha \ge 0, \beta} (L(x, \alpha, \beta)) \right)$$

For instance, if I ignore the $\{h_j\}$ terms which is zero for a suitable solution for simplicity, I observe that

$$\max_{\alpha \ge 0, \beta} (L(x, \alpha, \beta)) = \begin{cases} f(x) & g_i(x) \le 0, \forall i \\ \infty, & g_i(x) > 0, \exists i \end{cases}$$

Definition 36. (Lagrangian dual problem) The associated Lagrangian dual problem is

$$d^* = \max_{\alpha, \beta} \left(\min_x (L(x, \alpha, \beta)) \right) \quad \text{s.t.}\, \alpha_i \ge 0 \quad \beta_j \in \mathbb{R}$$

Definition 37. (Dual function) To the Dual problem, we associate a function $\tilde{L} : \mathbb{R}^m \times \mathbb{R}^n$ with

$$\tilde{L}(\alpha, \beta) := \min_x (L(x, \alpha, \beta))$$

called dual function.

Proposition 38. *In general it is $p^* \ge d^*$; i.e.*

$$\mathit{min}_x \max_{\alpha \ge 0, \beta} (L(x, \alpha, \beta)) \ge \mathit{max}_{\alpha \ge 0, \beta} \min_x (L(x, \alpha, \beta))$$

Definition 39. We call the general case $p^* \ge d^*$ weak duality.

Definition 40. When $p^* = d^*$ we say we have a strong duality.

Proposition 41. *(Strong duality via Slater condition) If the primal problem (4.1) is convex, and satisfies the weak Slater's condition, i.e.*

$$(\exists x_0 \in \mathcal{D}) \ : \ (g_i(x_0) < 0,\ \forall i = 1, ..., n) \ \mathit{and}\ (h_i(x_0) = 0,\ \forall i = 1, ..., n)$$

then strong duality holds, that is: $p^* = d^*$. It other words

$$\min_{x} \max_{\alpha\geq 0,\beta} \left(L\left(x,\alpha,\beta\right)\right) = \max_{\alpha\geq 0,\beta} \min_{x} \left(L\left(x,\alpha,\beta\right)\right)$$

Proposition 42. *When f, $\{h_j\}$, $\{g_i\}$ are convex and the Slater condition holds, Karush–Kuhn–Tucker (KKT) conditions are necessary and sufficient conditions for x^* to be local minimum*

$$\begin{aligned}
0 =& \nabla f\left(x^*\right) + \sum_{j=1}^{n} \beta_j \nabla h_j\left(x^*\right) + \sum_{i=1}^{m} \alpha_i \nabla g_i\left(x^*\right) && \textit{Stationarity} \\
& g_i\left(x^*\right) \leq 0,\ \ \forall i = 1,...,m && \textit{Primal feasibility} \\
& h_j\left(x^*\right) = 0,\ \ \forall j = 1,...,n && \\
& \alpha_i \geq 0\ \ \forall i = 1,...,m && \textit{Dual feasibility} \\
& \alpha_i g_i\left(x^*\right) = 0,\ \ \forall i = 1,...,m && \textit{Complementary slackness}
\end{aligned} \tag{4.3}$$

4.2. **Implementation in the Hand SVM.**

Note 43. Consider the Primal problem in Hard-SVM Algorithm 15 (2.2) and (2.2).

Note 44. According to Definition 34 the Lagrangian function is

$$L\left(w,b,\alpha\right) = \frac{1}{2}\left\|w\right\|_2^2 - \sum_{i=1}^{m} \alpha_i\left(y_i\left(\langle w, x_i\rangle + b\right) - 1\right) \tag{4.4}$$

where $\{\alpha_i\}_{i=1}^{m}$ are non-negative Lagrange multipliers.

Insight with regards to Note 35

Note 45. (Insight with regards to Note 35) The Hard-SVM Problem 15 (2.2) and (2.2), can be rewritten in the form

$$\min_{w}\left(\frac{1}{2}\left\|w\right\|_2^2 + g\left(w\right)\right) \tag{4.5}$$

with

$$g\left(w\right) = \max_{\alpha\geq 0} \sum_{i=1}^{m} \alpha_i\left(1 - y_i\left(\langle w, x_i\rangle + b\right)\right) = \begin{cases} 0 & \text{if } \ y_i\left(\langle w, x_i\rangle + b\right) \geq 1 \\ \infty & \text{else} \end{cases}$$

Hence we get the weak duality

$$\begin{aligned}
&\min_{w,b} \max_{\alpha\geq 0}\left(\frac{1}{2}\left\|w\right\|_2^2 + \sum_{i=1}^{m} \alpha_i\left(1 - y_i\left(\langle w, x_i\rangle + b\right)\right)\right) \\
&\geq \max_{\alpha\geq 0} \min_{w,b}\left(\underbrace{\frac{1}{2}\left\|w\right\|_2^2 + \sum_{i=1}^{m} \alpha_i\left(1 - y_i\left(\langle w, x_i\rangle + b\right)\right)}_{=L(w,b;\alpha)}\right)
\end{aligned} \tag{4.6}$$

Note 46. Strong duality (equality) holds because of the convexity. Here, keeping α fixed, $L(w, \alpha, b)$ is minimized when

$$0 = \nabla_w L(w, \alpha, b)|_{(w^*, b^*)} \implies w^* = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{4.7}$$

$$0 = \nabla_b L(w, \alpha, b)|_{(w^*, b^*)} \implies 0 = \sum_{i=1}^{m} \alpha_i y_i \tag{4.8}$$

The dual Lagrangian function is

$$\begin{aligned}\tilde{L}(\alpha) = \min_{w,b}(L(w, \alpha, b)) &= \frac{1}{2}\left\|\sum_{i=1}^{m} \alpha_i y_i x_i\right\|_2^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(\langle \sum_{j=1}^{m} \alpha_j y_j x_j, x_i \rangle + b\right) - 1\right) \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_j x_i \rangle\end{aligned}$$

Then dual problem is

$$\max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_j x_i \rangle\right)$$

Note 47. The Dual problem of the Hard-SVM (primal) problem 15 is

$$\alpha^* = \arg \max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle\right) \tag{4.9}$$

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

Note 48. Once the optimal α^* is computed from (4.9), the optimal weights w^* can be computed (from (4.7)) as

$$w^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i \tag{4.10}$$

Note 49. If $\alpha_i^* = 0$, the example (x_i, y_i) does not contribute to (4.10). If $\alpha_i^* \neq 0$, (x_i, y_i) contributes to (4.10). Moreover if $\alpha_i^* \neq 0$ the KKT condition (4.3)

$$\alpha_i^* (y_i (\langle w^*, x_i \rangle + b^*) - 1) = 0$$

implies that $y_i (\langle w^*, x_i \rangle + b^*) = 1$ meaning that x_i is on the boundary of the margin. Features x_i associated to $\alpha_i^* \neq 0$ which contribute to the evaluation of the optimal weights w^* and hence the evaluation of the separating predictive rule $h_{w,b}$ are called **supporting vectors** (see Figure 4.1).

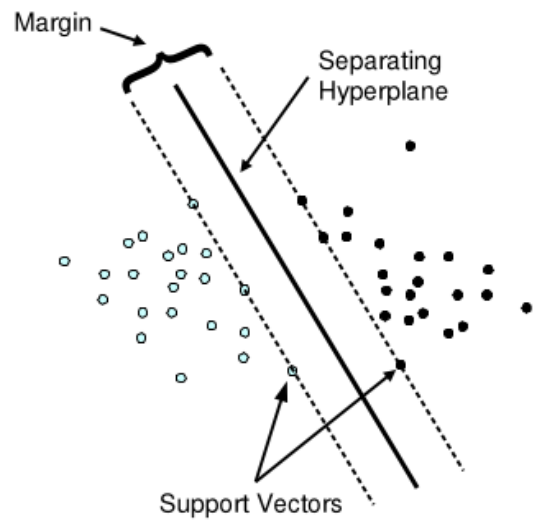


FIGURE 4.1

Note 50. Given optimal α^* (4.10) becomes

$$w^* = \sum_{i \in \mathcal{I}} \alpha_i^* y_i x_i \tag{4.11}$$

where $\mathcal{I} = \{i : y_i (\langle w^*, x_i \rangle + b^*) - 1 = 0\}$.

Note 51. The bias b parameter can be computed by KKT condition (4.3), it is

$$\alpha_i^* (y_i (\langle w^*, x_i \rangle + b^*) - 1) = 0$$

Consider $\mathcal{I} = \{i : y_i (\langle w^*, x_i \rangle + b^*) - 1 = 0\}$ then, for $i \in \mathcal{I}$ it is

$$y_i^2 (\langle w^*, x_i \rangle + b^*) - y_i = 0 \overset{y_i^2=1}{\Longrightarrow} \langle w^*, x_i \rangle + b^* - y_i = 0$$

and summing up all $i \in \mathcal{I}$, I get

$$\sum_{i \in \mathcal{I}} (\langle w^*, x_i \rangle + b^* - y_i) = 0 \implies b^* = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (y_i - \langle w^*, x_i \rangle) \tag{4.12}$$

Note 52. Therefore the predictive halfspace hypothesis is given as

$$\begin{aligned} h_{w,b}(x) &= \text{sign}(\langle w^*, x \rangle + b^*) \\ &= \text{sign}\left(\sum_{i \in \mathcal{I}} \alpha_i^* y_i \langle x_i, x \rangle + b^* \right) \end{aligned} \tag{4.13}$$

with $\mathcal{I} = \{i : y_i (\langle w^*, x_i \rangle + b^*) - 1 = 0\}$.

Note 53. Compared to the Primal problem (Problem 15), the dual problem (Note 47) is computationally desirable in cases that the training data size m is smaller than the number of the dimensions d of the feature space; i.e. $d \gg m$. The solution of the Primal quadratic programming problem in $d+1$ and complexity $O(d+1)$, while the dual quadratic programming problem has m variables and complexity $O(m)$.

Note 54. Compared to the Primal problem, the dual problem (Note 47), can provide sparse solutions relying on a small number of support vectors (see Eq. 4.11, 4.12, and 4.13).

Note 55. (Snapshot of the next concept "The kernel trick") The importance of the existence of the dual problem is that eg (4.9) and (4.13) involves the inner products between instances and does not require the direct access to specific elements of features within instance. For instance, if we consider the hypothesis (1.2) as an expansion of bases $h_{w,b}(x) = \text{sign}(\langle w, \phi(x)\rangle + b)$ with $\phi(x) = (\phi_1(x), ..., \phi_d(x))$ with a large d such as $d \gg m$ then, based on (4.13), the separation rule is

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

and 4.9 becomes

$$\alpha^* = \arg \max_{\alpha \in \mathbb{R}^m : \alpha \geq 0} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j)\right)$$

with $k(x', x) = \langle \phi(x'), \phi(x)\rangle$. As we will see in the "kernel methods" lecture, in specific case, one can specify the "kernel" function $k(\cdot, \cdot)$ (having with specific desirable properties) avoiding the direct specification of a possibly high-dimensional dictionary of features $\{\phi_j(\cdot)\}$.

Note 56. In the Soft-SVM case, the solution is the same as in the Hard-SVM (4.10) and (4.13), the only difference is that in the quadratic programming problem (4.9) it is subject to $\alpha_j \in [0, C]$ where $C = 1/2\lambda$, it is called "cost" and controls the cost of having the constraint violation by adding the $\xi_i' s$. (See Exercise 16 from the Exercise sheet).

Note 57. Sparsity and duality in Soft SVM can be seen in Exercise 16 in the Exercise sheet.

Note 58. Relevance Vector Machine (a version of Bayesian SVM) can be seen in Exercise 17 in the Exercise sheet.