Assume we have samples $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ as well as a new sample z_{n+1}. The classifier trained on the data $z_1, \ldots, z_n$ is $f_{z_1,\ldots,z_n}$.

The error of this classifier is

$$\text{Error}(z_1, \ldots, z_n) = \mathbb{E}_{z_{n+1}} I(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1}) = \mathbb{P}_{z_{n+1}}(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1})$$

and the *Average Generalization Error*

$$\text{A.G.E.} = \mathbb{E}\ \text{Error}(z_1, \ldots, z_n) = \mathbb{E}\mathbb{E}_{z_{n+1}} I(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1}).$$

Since $z_1, \ldots, z_n, z_{n+1}$ are i.i.d., in expectation training on $z_1, \ldots, z_i, \ldots, z_n$ and evaluating on z_{n+1} is the same as training on $z_1, \ldots, z_{n+1}, \ldots, z_n$ and evaluating on z_i. Hence, for any i,

$$\text{A.G.E.} = \mathbb{E}\mathbb{E}_{z_i} I(f_{z_1,\ldots,z_{n+1},\ldots,z_n}(x_i) \neq y_i)$$

and

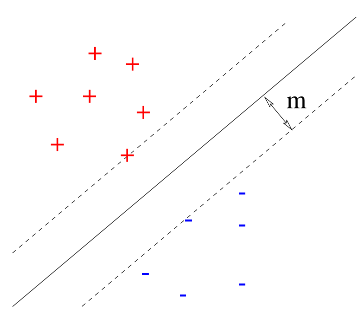
$$\text{A.G.E.} = \mathbb{E}\left[\underbrace{\frac{1}{n+1}\sum_{i=1}^{n+1} I(f_{z_1,\ldots,z_{n+1},\ldots,z_n}(x_i) \neq y_i)}_{\text{leave-one-out error}}\right].$$

Therefore, to obtain a bound on the generalization ability of an algorithm, it's enough to obtain a bound on its leave-one-out error. We now prove such a bound for SVMs. Recall that the solution of SVM is $\varphi = \sum_{i=1}^{n+1} \alpha_i^0 y_i x_i$.

Theorem 4.1.

$$L.O.O.E. \leq \frac{\min(\#\ \textit{support vect.}, D^2/m^2)}{n+1}$$

where D is the diameter of a ball containing all x_i, $i \leq n+1$ and m is the margin of an optimal hyperplane.



Remarks:

- dependence on sample size is $\frac{1}{n}$
- dependence on margin is $\frac{1}{m^2}$
- number of support vectors (sparse solution)

Lemma 4.1. *If x_i is a support vector and it is misclassified by leaving it out, then $\alpha_i^0 \geq \frac{1}{D^2}$.*

Given Lemma 4.1, we prove Theorem 4.1 as follows.

Proof. Clearly,

$$\text{L.O.O.E.} \leq \frac{\#\text{ support vect.}}{n+1}.$$

Indeed, if x_i is not a support vector, then removing it does not affect the solution. Using Lemma 4.1 above,

$$\sum_{i \in \text{supp.vect}} I(x_i \text{ is misclassified}) \leq \sum_{i \in \text{supp.vect}} \alpha_i^0 D^2 = D^2 \sum \alpha_i^0 = \frac{D^2}{m^2}.$$

In the last step we use the fact that $\sum \alpha_i^0 = \frac{1}{m^2}$. Indeed, since $|\varphi| = \frac{1}{m}$,

$$\begin{aligned}
\frac{1}{m^2} &= |\varphi|^2 = \varphi \cdot \varphi = \varphi \cdot \sum \alpha_i^0 y_i x_i \\
&= \sum \alpha_i^0 (y_i \varphi \cdot x_i) \\
&= \underbrace{\sum \alpha_i^0 (y_i(\varphi \cdot x_i + b) - 1)}_{0} + \sum \alpha_i^0 - b \underbrace{\sum \alpha_i^0 y_i}_{0} \\
&= \sum \alpha_i^0
\end{aligned}$$

□

We now prove Lemma 4.1. Let $u * v = K(u, v)$ be the dot product of u and v, and $\|u\| = (K(u,u))^{1/2}$ be the corresponding L_2 norm. Given $x_1, \cdots, x_{n+1} \in \mathbb{R}^d$ and $y_1, \cdots, y_{n+1} \in \{-1, +1\}$, recall that the primal problem of training a support vector classifier is $\text{argmin}_\psi \frac{1}{2}\|\psi\|^2$ subject to $y_i(\psi * x_i + b) \geq 1$. Its dual problem is $\text{argmax}_\alpha \sum \alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2$ subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$, and $\psi = \sum \alpha_i y_i x_i$. Since the Kuhn-Tucker condition can be satisfied, $\min_\psi \frac{1}{2}\psi * \psi = \max_\alpha \sum \alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2 = \frac{1}{2m^2}$, where m is the margin of an optimal hyperplane.

Proof. Define $w(\alpha) = \sum_i \alpha_i - \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2$. Let $\alpha^0 = \text{argmax}_\alpha w(\alpha)$ subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$. Let $\alpha' = \text{argmax}_\alpha w(\alpha)$ subject to $\alpha_p = 0$, $\alpha_i \geq 0$ for $i \neq p$ and $\sum \alpha_i y_i = 0$. In other words, α^0 corresponds to the support vector classifier trained from $\{(x_i, y_i) : i = 1, \cdots, n+1\}$ and α' corresponds to the support vector classifier trained from $\{(x_i, y_i) : i = 1, \cdots, p-1, p+1, \cdots, n+1\}$. Let $\gamma = \left(\overset{1}{\underset{0}{\downarrow}}, \cdots, \overset{p-1}{\underset{0}{\downarrow}}, \overset{p}{\underset{1}{\downarrow}}, \overset{p+1}{\underset{0}{\downarrow}}, \cdots, \overset{n+1}{\underset{0}{\downarrow}} \right)$. It follows that $w(\alpha^0 - \alpha_p^0 \cdot \gamma) \leq w(\alpha') \leq w(\alpha^0)$. (For the dual problem, α' maximizes $w(\alpha)$ with a constraint that $\alpha_p = 0$, thus $w(\alpha')$ is no less than $w(\alpha^0 - \alpha_p^0 \cdot \gamma)$, which is a special case that satisfies the constraints, including $\alpha_p = 0$. α^0 maxmizes $w(\alpha)$ with a constraint $\alpha_p \geq 0$, which raises the constraint $\alpha_p = 0$, thus $w(\alpha') \leq w(\alpha^0)$. For the primal problem, the training problem corresponding to α' has less samples (x_i, y_i), where $i \neq p$, to separate with maximum margin, thus its margin $m(\alpha')$ is no less than the margin $m(\alpha^0)$,

and $w(\alpha') \le w(\alpha^0)$. On the other hand, the hyperplane determined by $\alpha^0 - \alpha_p^0 \cdot \gamma$ might not separate (x_i, y_i) for $i \neq p$ and corresponds to a equivalent or larger "margin" $1/\|\psi(\alpha^0 - \alpha_p^0 \cdot \gamma)\|$ than $m(\alpha')$).
Let us consider the inequality

$$\max_t w(\alpha' + t \cdot \gamma) - w(\alpha') \le w(\alpha^0) - w(\alpha') \le w(\alpha^0) - w(\alpha^0 - \alpha_p^0 \cdot \gamma).$$

For the left hand side, we have

$$\begin{aligned} w(\alpha' + t\gamma) &= \sum \alpha_i' + t - \frac{1}{2} \left\| \sum \alpha_i' y_i x_i + t \cdot y_p x_p \right\|^2 \\ &= \sum \alpha_i' + t - \frac{1}{2} \left\| \sum \alpha_i' y_i x_i \right\|^2 - t \left(\sum \alpha_i' y_i x_i \right) * (y_p x_p) - \frac{t^2}{2} \|y_p x_p\|^2 \\ &= w(\alpha') + t \cdot (1 - y_p \cdot \underbrace{(\sum \alpha_i' y_i x_i)}_{\psi'} * x_p) - \frac{t^2}{2} \|x_p\|^2 \end{aligned}$$

and $w(\alpha' + t\gamma) - w(\alpha') = t \cdot (1 - y_p \cdot \psi' * x_p) - \frac{t^2}{2} \|x_p\|^2$. Maximizing the expression over t, we find $t = (1 - y_p \cdot \psi' * x_p)/\|x_p\|^2$, and

$$\max_t w(\alpha' + t\gamma) - w(\alpha') = \frac{1}{2} \frac{(1 - y_p \cdot \psi' * x_p)^2}{\|x_p\|^2}.$$

For the right hand side,

$$\begin{aligned} w(\alpha^0 - \alpha_p^0 \cdot \gamma) &= \sum \alpha_i^0 - \alpha_p^0 - \frac{1}{2} \| \underbrace{\sum \alpha_i^0 y_i x_i}_{\psi_0} - \alpha_p^0 y_p x_p \|^2 \\ &= \sum \alpha_i^0 - \alpha_p^0 - \frac{1}{2} \|\psi_0\|^2 + \alpha_p^0 y_p \psi_0 * x_p - \frac{1}{2} \left(\alpha_p^0\right)^2 \|x_p\|^2 \\ &= w(\alpha_0) - \alpha_p^0 (1 - y_p \cdot \psi_0 * x_p) - \frac{1}{2} \left(\alpha_p^0\right)^2 \|x_p\|^2 \\ &= w(\alpha_0) - \frac{1}{2} \left(\alpha_p^0\right)^2 \|x_p\|^2. \end{aligned}$$

The last step above is due to the fact that (x_p, y_p) is a support vector, and $y_p \cdot \psi_0 * x_p = 1$. Thus $w(\alpha^0) - w(\alpha^0 - \alpha_p^0 \cdot \gamma) = \frac{1}{2} \left(\alpha_p^0\right)^2 \|x_p\|^2$ and $\frac{1}{2} \frac{\left(1 - y_p \cdot \psi' * x_p\right)^2}{\|x_p\|^2} \le \frac{1}{2} \left(\alpha_p^0\right)^2 \|x_p\|^2$. Thus

$$\begin{aligned} \alpha_p^0 &\ge \frac{|1 - y_p \cdot \psi' * x_p|}{\|x_p\|^2} \\ &\ge \frac{1}{D^2}. \end{aligned}$$

The last step above is due to the fact that the support vector classifier associated with ψ' misclassifies (x_p, y_p) according to assumption, and $y_p \cdot \psi' * x_p \le 0$, and the fact that $\|x_p\| \le D$. □