2.3 Support Vector Machine (SVM)

2.3.1 Linear SVM

Revisiting binary classification / Hard margin

For (linear) logistic regression, binary classification takes a probabilistic view $\mathbb{P}_{\mp} = (1 + e^{\pm\theta^T x})^{-1}$,[28] thus resulting in a linear decision boundary $\theta^T x = 0$. Since $\mathbb{P}_{\mp} \to 1$ as $\theta^T x \to \mp\infty$, we expect that the classification would become increasingly accurate with x further away from the decision surface. There-

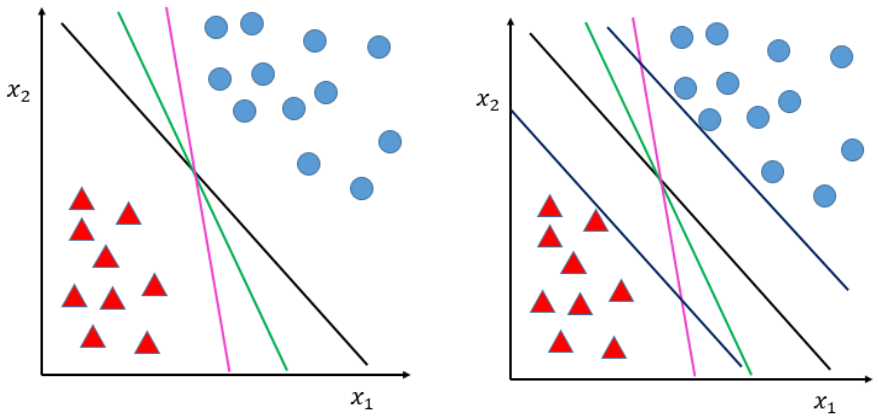

Figure 2.9: Illustrating linear SVM as a large margin classifier with a 2-attribute example. (Left) We see that the 2 classes ($\circ$ / Δ) are linearly separable via multiple decision surfaces, represented by the black, green, and pink lines. (Right) Here, the linear SVM algorithm chooses the black decision surface which gives the largest margin width, as indicated by the distance between the blue lines $2\mu/||\bar{\theta}'||$.

fore, for maximal classification accuracy, the optimal decision surface must be the one that would perpendicularly bisect the 2 classes at equidistance from one another.

In this section, we shall investigate how (linear) SVM derives its notion of perpendicular equidistance from the concept of margin. To start with, let us first assume that the 2 classes are linearly separable by some decision surface.

[28] Here, we denote $\mathbb{P}_- = \mathbb{P}_{k=0}$ and $\mathbb{P}_+ = \mathbb{P}_{k=1}$.

With $\mu^{(i)} = \theta^T x^{(i)}$,[29] we define

$$\begin{aligned} \mu_- &= \sup\{\, \mu^{(i)} \mid y^{(i)} = 0 \,\} \quad < 0 \\ \mu_+ &= \inf\{\, \mu^{(i)} \mid y^{(i)} = 1 \,\} \quad > 0 \end{aligned}$$

so that $\theta^T x^{(i)} \geq |\mu_+|$ if $y^{(i)} = 1$, and $\theta^T x^{(i)} \leq -|\mu_-|$ if $y^{(i)} = 0$. Next, we constrain θ_0 such that $|\mu_-| = |\mu_+| = \mu$ for some **functional margin** μ, thus leaving us with the remaining degrees of freedom $\theta' = (\theta_1, \theta_2 \cdots \theta_N)$ over which we attempt to maximize[30] the **geometric margin** $(\,\mu/||\theta'||\,)$, or the **margin width** $(\,2\mu/||\theta'||\,)$, i.e. (see figure 2.9)

$$\max_{\theta_1\,\theta_2 \cdots \theta_N} \frac{2\mu}{||\theta'||} \tag{2.96}$$

with constraints

$$(-1)^{1-y^{(i)}}\, \frac{\theta^T}{\mu} x^{(i)} \geq 1 \qquad , \quad \forall\, i = 1 \cdots M. \tag{2.97}$$

Hence, for the case of linearly-separable classes, they are separated by a hard margin with an optimal decision surface $\theta^T x = 0$, such that

$$\bar{\theta}' = \arg\min_{\bar{\theta}'} \frac{1}{2}||\bar{\theta}'||^2 \tag{2.98}$$

$$\bar{y}^{(i)} \bar{\theta}^{\,T} x^{(i)} \geq 1 \tag{2.99}$$

where $\bar{\theta} = \theta/\mu$ and $\bar{y} = (-1)^{1-y} = \mp 1$. Last but not least, **support vectors** $x^{(i)}$ are those which sit on the hard margin, i.e. $\bar{y}^{(i)} \bar{\theta}^{\,T} x^{(i)} = 1$.

The dual formalism

Although the convex optimization (2.98)-(2.99) is solvable via generic quadratic programming, its dual form turns out to be computationally more efficient. To start with, we can define the corresponding Lagrangian

$$L(\bar{\theta}', \bar{\theta}_0, \alpha) = \frac{1}{2}||\bar{\theta}'||^2 - \sum_{i=1}^{M} \alpha_i \left(\bar{y}^{(i)} \bar{\theta}'^{\,T} x^{(i)} + \bar{y}^{(i)} \bar{\theta}_0 - 1 \right) \tag{2.100}$$

---

[29] Let $x_*^{(i)}$ be the orthogonal projection of $x^{(i)}$ on the decision surface, where $\theta^T x_*^{(i)} = 0$. Next, we define the distance vector from the decision surface, i.e. $d^{(i)} = x^{(i)} - x_*^{(i)}$, so that

$$\mu^{(i)} = \begin{cases} -||\theta||\,||d^{(i)}|| & ,\ \mathrm{y^{(i)}} = 0 \\ +||\theta||\,||d^{(i)}|| & ,\ \mathrm{y^{(i)}} = 1 \end{cases}$$

[30] In fact, SVM is known to be a **large margin classifier**.

with Lagrange multipliers $\alpha_i \geq 0$, such that

$$\max_{\alpha_i \geq 0} L(\bar{\theta}', \bar{\theta_0}, \alpha) = \frac{1}{2}||\bar{\theta}'||^2 \tag{2.101}$$

and thus the optimization objective

$$\max_{\alpha_i \geq 0} \min_{\bar{\theta}'} L(\bar{\theta}', \bar{\theta_0}, \alpha) \overset{\text{strong duality}}{=} \min_{\bar{\theta}'} \max_{\alpha_i \geq 0} L(\bar{\theta}', \bar{\theta_0}, \alpha). \tag{2.102}$$

whenever there exists some saddle point of $L(\bar{\theta}', \bar{\theta_0}, \alpha)$ which satisfies the Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{\partial L}{\partial \bar{\theta}'} = 0 \quad , \quad \frac{\partial L}{\partial \bar{\theta_0}} = 0 \tag{2.103}$$

$$\alpha_i \left(\bar{y}^{(i)} \bar{\theta}'^{\,T} x^{(i)} + \bar{y}^{(i)} \bar{\theta_0} - 1 \right) = 0 \quad , \; \forall \, i = 1 \cdots M \tag{2.104}$$

Therefore, only (a few) support vectors would have $\alpha_i > 0$, otherwise $\alpha_i = 0$. After trivial algebra, we arrive at

$$\bar{\theta}' = \sum_{i=1}^{M} \alpha_i \, \bar{y}^{(i)} x^{(i)} \quad , \quad \sum_{i=1}^{M} \alpha_i \, \bar{y}^{(i)} = 0 \tag{2.105}$$

and

$$\begin{aligned} L(\alpha) &= \min_{\bar{\theta}'} L(\bar{\theta}', \bar{\theta_0}, \alpha) \\ L(\alpha) &= \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i_1, i_2 = 1}^{M} \alpha_{i_1} \alpha_{i_2} \, \bar{y}^{(i_1)} \bar{y}^{(i_2)} \left(x^{(i_1)\,T} x^{(i_2)} \right) \; , \end{aligned} \tag{2.106}$$

thus the dual (quadratic) optimization problem:

$$\begin{aligned} \alpha &= \arg\max_{\alpha_i \geq 0} L(\alpha) \\ \alpha &= \arg\max_{\alpha_i \geq 0} \left(\sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i_1, i_2 = 1}^{M} \alpha_{i_1} \alpha_{i_2} \, \bar{y}^{(i_1)} \bar{y}^{(i_2)} \left(x^{(i_1)\,T} x^{(i_2)} \right) \right) \end{aligned} \tag{2.107}$$

Note that, via an SMO algorithm, α can be efficiently solved from (2.107).

Therefore, the (hard-SVM) decision surface is given by

$$\bar{\theta}' = \sum_{i \, \in \, \text{SV}} \alpha_i \, \bar{y}^{(i)} x^{(i)} \tag{2.108}$$

$$\bar{\theta_0} = \bar{y}^{(i)} \left(1 - \bar{y}^{(i)} \bar{\theta}'^{\,T} x^{(i)} \right) \; , \; \forall \, i \in \text{SV} \tag{2.109}$$

where SV is the set of all support vectors.

Finally, for some new data x, we would predict its classification as

$$\bar{y} = \text{sgn}\left[\bar{\theta}_0 + \sum_{i \in \text{SV}} \alpha_i \, \bar{y}^{(i)} \left(x^{(i)\,T} x\right)\right] \tag{2.110}$$

Noisy binary classification / Soft margin

In this section, we shall relax the linear-separability assumption between the 2 classes. We introduce the concept of **excitation**, where some data $(x^{(i)}, y^{(i)})$ get excited and move across the geometric margin away from its classification regime.

For excited data, we define a slack variable $\xi_+^{(i)} = \xi^{(i)}$ such that $\bar{y}^{(i)}\,\bar{\theta}^T x^{(i)} = 1 - \xi^{(i)}$ with $\xi^{(i)} \geq 0$. Otherwise, for non-excited data, $\xi_+^{(i)} = 0$. Now, we penalize all $\xi_+^{(i)}$ to the first order, thus the following optimization objective:

$$\min_{\bar{\theta}',\, \xi^{(i)}} \left(\frac{1}{2}||\bar{\theta}'||^2 + \frac{1}{T^2} \sum_{i=1}^{M} \xi_+^{(i)}\right) \tag{2.111}$$

$$\begin{aligned} \bar{y}^{(i)}\,\bar{\theta}^T x^{(i)} &\geq 1 - \xi_+^{(i)} \\ \xi_+^{(i)} &\geq 0 \end{aligned}$$

for some parameter T.

To start with, we define the corresponding Lagrangian, i.e. $L(\bar{\theta}', \bar{\theta}_0, \xi, \alpha, \beta) =$

$$\left[\frac{1}{2}||\bar{\theta}'||^2 + \frac{1}{T^2} \sum_{i=1}^{M} \xi_+^{(i)}\right] - \sum_{i=1}^{M} \alpha_i \left(\bar{y}^{(i)} \bar{\theta}'^{\,T} x^{(i)} + \bar{y}^{(i)} \bar{\theta}_0 - 1 + \xi_+^{(i)}\right) - \sum_{i=1}^{M} \beta_i \, \xi_+^{(i)}$$

with $\alpha_i, \beta_i \geq 0$. Likewise, we impose the KKT conditions, i.e.

$$\frac{\partial L}{\partial \bar{\theta}'} = 0 \quad , \quad \frac{\partial L}{\partial \bar{\theta}_0} = 0 \quad , \quad \frac{\partial L}{\partial \xi_+^{(i)}} = 0 \quad , \tag{2.112}$$

and arrive at the same dual Lagrangian dual

$$L(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i_1, i_2 = 1}^{M} \alpha_{i_1} \alpha_{i_2} \, \bar{y}^{(i_1)} \bar{y}^{(i_2)} \left(x^{(i_1)\,T} x^{(i_2)}\right) \ , \tag{2.113}$$

which needs to be maximized such that $\alpha_i \in [0, \frac{1}{T^2}]$ and $\sum_{i=1}^{M} \alpha_i \bar{y}^{(i)} = 0$.

Via an SMO algorithm, say, α can be efficiently solved, i.e.

$$\alpha = \underset{0 \le \alpha_i \le \frac{1}{T^2}}{\arg\max} \left(\sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i_1, i_2 = 1}^{M} \alpha_{i_1} \alpha_{i_2} \, \bar{y}^{(i_1)} \bar{y}^{(i_2)} \left(x^{(i_1)\,T} x^{(i_2)} \right) \right) \tag{2.114}$$

Therefore, the (soft-SVM) decision surface is given by

$$\bar{\theta}' = \sum_{i \in \mathrm{SV}} \alpha_i \, \bar{y}^{(i)} x^{(i)} \tag{2.115}$$

$$\bar{\theta}_0 = \bar{y}^{(i)} \left(1 - \bar{y}^{(i)} \bar{\theta}'^{\,T} x^{(i)} \right) \;, \; \forall \, i \in \mathrm{SV} \tag{2.116}$$

where SV is the set of all support vectors.[31]

Finally, for some new data x, we would predict its classification as

$$\bar{y} = \mathrm{sgn} \left[\bar{\theta}_0 + \sum_{i \in \mathrm{SV}} \alpha_i \, \bar{y}^{(i)} \left(x^{(i)\,T} x \right) \right] \tag{2.117}$$

Statistical mechanics

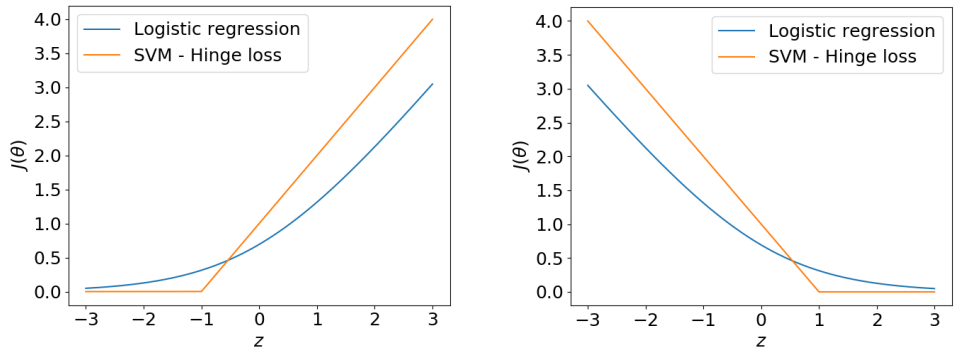

Figure 2.10: Illustrating the comparison of cost functions between SVM (green line) and logistic regression (blue line) for (left) $y = 0$ and (right) $y = 1$. Except for the gap $|z| \le \mu$, low-temperature SVM derives its origin from logistic regression.

[31] Recall that $\alpha_i > 0$ for all support vectors. Otherwise, $\alpha_i = 0$.