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

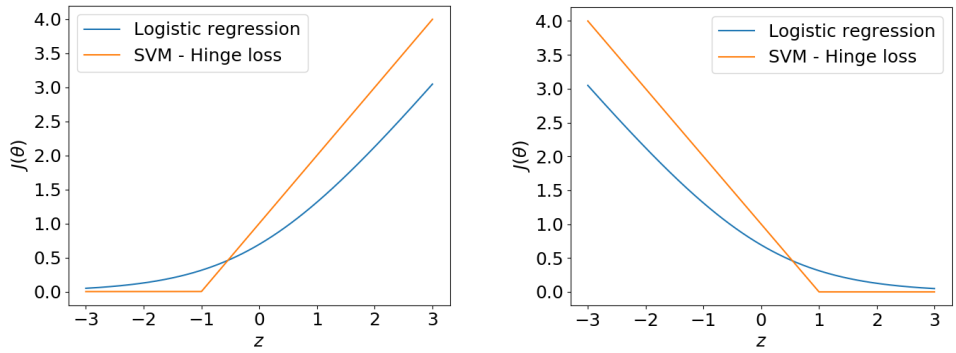

Figure 2.10: Illustrating the comparison of cost functions between SVM (green line) and logistic regression (blue line) for (left) $y = 0$ and (right) $y = 1$. Except for the gap $|z| \le \mu$, low-temperature SVM derives its origin from logistic regression.

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

Revisiting (2.111), the minimization objective can be rewritten as

$$\min_{\bar{\theta}', \xi^{(i)}} \left(\frac{1}{T^2} \sum_{i=1}^{M} \xi_+^{(i)} + \frac{1}{2} ||\bar{\theta}'||^2 \right) = \min_{\bar{\theta}', \xi^{(i)}} \left(\frac{1}{T} \sum_{i=1}^{M} \xi_+^{(i)} + \frac{T}{2} \sum_{j=1}^{N} \bar{\theta}_j^2 \right) \tag{2.118}$$

with hinge loss

$$\frac{\xi_+^{y=0}}{T} = \begin{cases} 0 & , \bar{\theta}^T x \leq -1 \\ \frac{1}{T}(1 + \bar{\theta}^T x) & , \text{otherwise} \end{cases} \quad ; \quad \frac{\xi_+^{y=1}}{T} = \begin{cases} 0 & , \bar{\theta}^T x \geq 1 \\ \frac{1}{T}(1 - \bar{\theta}^T x) & , \text{otherwise} \end{cases}$$

As $T \to 0$, we have

$$\frac{\xi_+^{y=0}}{T} = -\log \left(\frac{1}{1 + e^{\frac{1}{T}(\bar{\theta}^T x + 1)}} \right) \quad ; \quad \frac{\xi_+^{y=1}}{T} = -\log \left(\frac{1}{1 + e^{-\frac{1}{T}(\bar{\theta}^T x - 1)}} \right)$$

or

$$\frac{\xi_+}{T} = -(1 - y) \log \left(\frac{1}{1 + e^{\frac{1}{T}(\bar{\theta}^T x + 1)}} \right) - y \log \left(\frac{1}{1 + e^{-\frac{1}{T}(\bar{\theta}^T x - 1)}} \right) \tag{2.119}$$

Now, we define the soft-SVM hypothesis

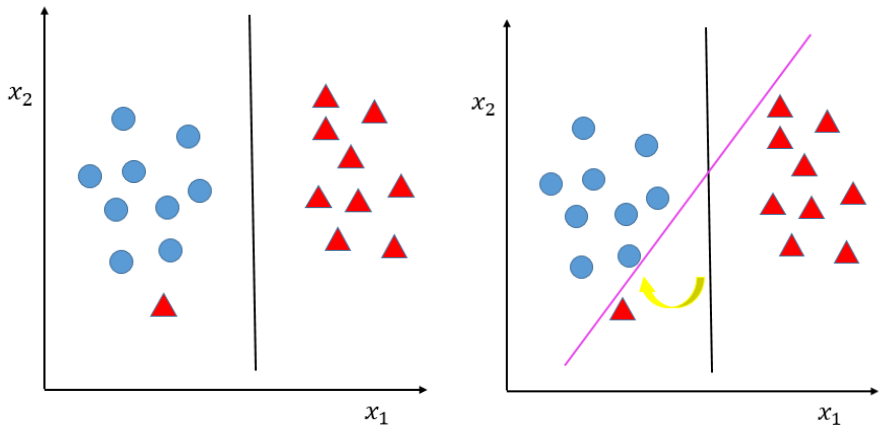

Figure 2.11: Illustrating the overfitting problem in low-temperature linear SVM. (Left) Assuming we have a new data Δ that falls in the wrong class. (Right) A rerun of the low-temperature SVM algorithm will almost certainly shift the black decision surface towards the purple. Hence, it is very sensitive to the outlier, which is highly undesirable. To reduce overfitting, simply perform SVM at a higher temperature T.

$$h_{\theta,T}^{[y=0]}(x) = \mathbb{P}(y=0\,|\,x\,;\,\theta,T) = \frac{1}{1+e^{\frac{1}{T}(\theta^T x+\mu)}} \tag{2.120}$$

$$h_{\theta,T}^{[y=1]}(x) = \mathbb{P}(y=1\,|\,x\,;\,\theta,T) = \frac{1}{1+e^{-\frac{1}{T}(\theta^T x-\mu)}} \tag{2.121}$$

Due to its close resemblance to statistical mechanics, we shall label T as temperature, and μ as chemical potential.

From 2.118, we now define the low-temperature SVM cost function

$$J_T(\theta) = -\frac{1}{M}\left[\sum_{i=1}^{M}\left\{y^{(i)}\log h_{\theta,T}^{[y^{(i)}=1]}(x) + (1-y^{(i)})\log h_{\theta,T}^{[y^{(i)}=0]}(x)\right\} - \frac{T/\mu^2}{2}\sum_{j=1}^{N}\theta_j^2\right]$$

with regularization parameter $\lambda = T/\mu^2$.[32] Here, except for the gap $|\theta^T x| \leq \mu$, we see that low-temperature SVM does in fact derive its origin from logistic regression.

Note that the cost function can also be calculated from the cross-entropy:

$$J_{T\to 0}(\theta) = -\sum_{k=0,1}\langle \pi_k \log \mathbb{P}_k\rangle \tag{2.122}$$

Last but not least, we shall define $\mathbb{P}(y=\Delta)$ for the sake of completeness:

$$\mathbb{P}(y=\Delta) + \mathbb{P}(y=0) + \mathbb{P}(y=1) = 1 \tag{2.123}$$

2.3.2 Nonlinear SVM

Transformation to linear SVM / Kernel trick

The idea is to invent a transformation $f : \mathbb{R}^{N+1} \to \mathbb{R}^{N'+1}$ where $N' > N$ such that $\{(f(x^{(i)}), y^{(i)})\}$ is linearly-separable in that transformed feature-space. In this case, we simply 1) replace $\theta^T x$ by $\Theta^T f = \Theta_0 + \Theta_1 f_1 + \cdots + \Theta_{N'} f_{N'}$, and 2) perform linear SVM to determine the optimal decision surface $\Theta^T f = 0$.

In Lagrangian dual form, in analogy to (2.114), this involves solving for

$$\alpha = \underset{0\leq\alpha_i\leq\frac{1}{T^2}}{\arg\max}\left(\sum_{i=1}^{M}\alpha_i - \frac{1}{2}\sum_{i_1,i_2=1}^{M}\alpha_{i_1}\alpha_{i_2}\,\bar{y}^{(i_1)}\bar{y}^{(i_2)}\left(f^{(i_1)\,T}f^{(i_2)}\right)\right)\ , \tag{2.124}$$

[32] At low temperature, we have a high variance SVM hypothesis which is susceptible to overfitting. In practice, low-temperature SVM is very sensitive to both outliers and noise, and they have therefore low forecastibility power. To reduce overfitting, we simply perform SVM at a higher temperature.

where the term $(x^{(i_1)T}x^{(i_2)})$ has been escalated to $(f^{(i_1)T}f^{(i_2)})$, thus increasing the computational complexity from $\mathcal{O}(M^2N^2)$ to $\mathcal{O}(M^2N'^2)$ which is clearly undesirable. To elevate this situation, we introduce kernel functions[33], i.e.

$$K(x^{(i_1)}, x^{(i_2)}) = f(x^{(i_1)})^T f(x^{(i_2)}) \tag{2.126}$$

[33]In machine learning, a symmetric function $K(x^{(i_1)}, x^{(i_2)})$, which defines some pairwise similarity measure over $\{(x^{(i)}, y^{(i)}) \,|\, i = 1 \cdots M\}$, is called a kernel (function), if and only if, there exists some mapping $f : \mathbb{R}^{N+1} \to \mathcal{H}$ into a separable Hilbert space $\mathcal{H}$ such that $K(x^{(i_1)}, x^{(i_2)}) \equiv \langle\, f(x^{(i_1)}), f(x^{(i_2)})\,\rangle_{\mathcal{H}}$. Furthermore, the **Mercer's theorem** states that $K(x^{(i_1)}, x^{(i_2)})$ will satisfy the Mercer's condition, i.e.

$$\int_{\mathbb{R}^{N+1}} \int_{\mathbb{R}^{N+1}} K(x^{(i_1)}, x^{(i_2)})\, g(x^{(i_1)})\, g(x^{(i_2)})\, dx^{(i_1)}\, dx^{(i_2)} \geq 0 \tag{2.125}$$

for all square-integrable functions $g \in L^2(\mathbb{R}^{N+1})$. For convenience, we shall now relabel $u \equiv x^{(i_1)}$ and $v \equiv x^{(i_2)}$.

First, we see that the linear kernel $K(u, v) = u^T v$ satisfies the Mercer's condition, i.e. $\int\int u^T v\, g(u)\, g(v)\, du\, dv = |\int u\, g(u)\, du|^2 \geq 0$. Second, note that indeed the polynomial kernel $K(u, v) = (u^T v)^p$, where $p \in \mathbb{N}$, also satisfies the Mercer's condition, i.e.

$$\begin{aligned}
& \int\int (u^T v)^p\, g(u)\, g(v)\, du\, dv \\
= & \sum_{r_0 \cdots r_N \geq 0}^{r_0 + \cdots + r_N = p} \frac{p!}{r_0! \cdots r_N!} \int\int (u_0\, v_0)^{r_0} \cdots (u_N\, v_N)^{r_N} g(u)\, g(v)\, du\, dv \\
= & \sum_{r_0 \cdots r_N \geq 0}^{r_0 + \cdots + r_N = p} \frac{p!}{r_0! \cdots r_N!} \left| \int u_0^{r_0} \cdots u_N^{r_N}\, g(u)\, du \right|^2 \geq 0 \ .
\end{aligned}$$

Third, the RBF kernel / Gaussian kernel $K(u, v) = \exp\left(-\frac{|u-v|^2}{2\sigma^2}\right)$ does likewise satisfies the Mercer's condition, i.e.

$$\begin{aligned}
& \int\int e^{-\frac{|u-v|^2}{2\sigma^2}}\, g(u)\, g(v)\, du\, dv \\
= & \int\int e^{\frac{u^T v}{\sigma^2}} \left(e^{-\frac{|u|^2}{2\sigma^2}} g(u)\right) \left(e^{-\frac{|v|^2}{2\sigma^2}} g(v)\right) du\, dv \\
= & \sum_{n=0}^{\infty} \frac{1}{n!\, \sigma^{2n}} \int\int \left(u^T v\right)^n \left(e^{-\frac{|u|^2}{2\sigma^2}} g(u)\right) \left(e^{-\frac{|v|^2}{2\sigma^2}} g(v)\right) du\, dv \geq 0
\end{aligned}$$

which follows from the consequence of polynomial kernels. (RBF = Radial Basis Function)

Other popular kernels include 1) sigmoid kernel, 2) χ^2 kernel, 3) string kernel, and 4) histogram intersection kernel.

Last but not least, similarity measures are not unique. They are usually in some sense the inverse of some distance metric. Similarity measures play a crucial role in the field of information retrieval.

such that the solving for

$$\alpha = \underset{0\leq\alpha_i\leq\frac{1}{T^2}}{\arg\max} \left(\sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i_1,i_2=1}^{M} \alpha_{i_1}\alpha_{i_2}\, \bar{y}^{(i_1)}\bar{y}^{(i_2)} K(x^{(i_1)}, x^{(i_2)}) \right) \tag{2.127}$$

remains its computational complexity at $\mathcal{O}(M^2N^2)$. Therefore, the kernel trick elegantly avoids the curse of dimensionality, without imposing additional computational complexity.

Nonlinear SVM in practice via Gaussian kernel

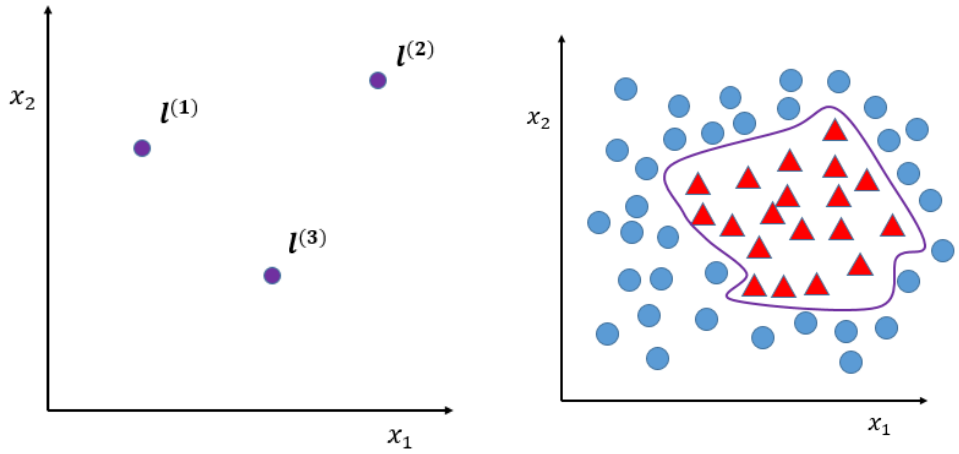

Figure 2.12: A naive nonlinear SVM approach. (Left) A set of M landmarks are chosen to coincide with the examples, i.e. $\{\, l_i \,|\, l_i = x^{(i)}\,;\, i = 1 \cdots M\}$. (Right) By performing low-temperature linear SVM on $\{(f(x^{(i)}), y^{(i)})\}$, we can derive Θ and thus the nonlinear decision surface as illustrated as the magenta curve. Nonetheless, this naive approach becomes computationally more expensive with increasing M. In practice, unless M is sufficiently small, say $M < 10^4$, we would stay away from this approach.

Intuitively, we may define 1) a set of M landmarks $\{\, l_i \,|\, l_i = x^{(i)}\,;\, i = 1 \cdots M\}$ from the examples, and 2) a mapping $f : \mathbb{R}^{N+1} \to \mathbb{R}^{M+1} \ni$

$$f_{j'}(x) = \exp\left(-\frac{|x - l_{j'}|^2}{2\sigma^2} \right) \quad ,\ j' = 1 \cdots M \tag{2.128}$$

and $f_0(x) = 1$. Hence, we have $\{(x^{(i)}, y^{(i)})\} \xrightarrow{f} \{(f(x^{(i)}), y^{(i)})\}$.

Now, we might be very tempted to perform low-temperature linear SVM:

$$J_T(\Theta) = -\frac{1}{M}\left[\sum_{i=1}^{M}\left\{y^{(i)}\log h_{\Theta,T}^{[y^{(i)}=1]}(f^{(i)}) + (1-y^{(i)})\log h_{\Theta,T}^{[y^{(i)}=0]}(f^{(i)})\right\} - \frac{T/\mu^2}{2}\sum_{j=1}^{M}\Theta_j^2\right]$$

with

$$h_{\Theta,T}^{[y=0]}(f) = \mathbb{P}(y=0 \mid f\,;\,\Theta,T) = \frac{1}{1+e^{\frac{1}{T}(\Theta^T f+\mu)}} \tag{2.129}$$

$$h_{\Theta,T}^{[y=1]}(f) = \mathbb{P}(y=1 \mid f\,;\,\Theta,T) = \frac{1}{1+e^{-\frac{1}{T}(\Theta^T f-\mu)}} \tag{2.130}$$

as conceptually illustrated in figure (2.12).

However, this naive approach suffers from the curse of dimensionality with increasing M. To reduce computational complexity, we have to make use of the Gaussian kernel, i.e.

$$K(x^{(i_1)}, x^{(i_2)}) = \exp\left(-\frac{|x^{(i_1)}-x^{(i_2)}|^2}{2\sigma^2}\right) = f(x^{(i_1)})^T\, J\, f(x^{(i_2)}) \tag{2.131}$$

for some Jacobian J, whose existence is guaranteed by the Mercer's theorem.

In practice, we usually start straight-away from the Lagrangian dual (2.127) with the Gaussian kernel (2.131), and derive the support vectors for future classification.

Last but not least, the Gaussian spread σ is one extra regularization parameter for us to tune our nonlinear decision surface curvature.

Multiclass classification

By design, SVM is only for binary classification. For multiclass classification, one would need to use the OneVsAll approach, and selects the most confident class with maximum $|\theta^T x|$.