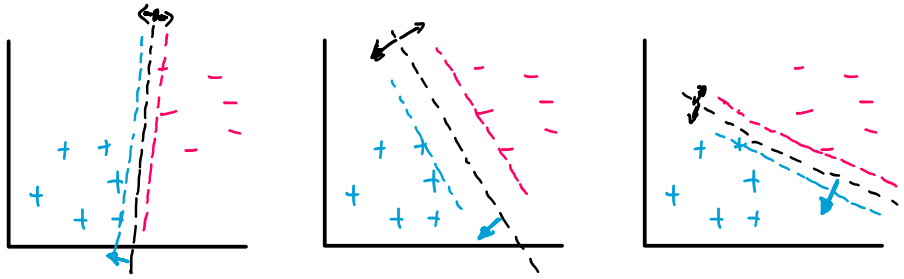
12-Nonlinear Models (SVMs)

Max-margin criteria

- Consider a binary, linearly separable classification task
  Non-binary, non-linear tasks => future development
- All the below classifiers give the same training accuracy (100%), but still one looks better than others.



- It looks the classifier with the maximum margin is the most stable one. (Max-margin classification has profound theoretical justifications, but we treat it as heuristics.)

Formulation

Input features: $\boldsymbol{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$

Target: $t \in \{+1, -1\}$

Note: There's no real difference whether we call the target {0, 1} or {+1, -1}

We would still like a linear (more rigorously, affine) model on $\boldsymbol{x}$, with parameters $\boldsymbol{w}, b$. Thus, the decision boundary is the hyperaplane $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$

For a data point $\left(\boldsymbol{x}^{(m)}, t^{(m)}\right)$, the "distance" between $\boldsymbol{x}^{(m)}$ to the hyperplane is $\frac{|\boldsymbol{w}^\top \boldsymbol{x}^{(m)} + b|}{\|\boldsymbol{w}\|}$

If $\boldsymbol{x}^{(m)}$ is in the half space that the normal vector points to, then $\boldsymbol{w}^\top \boldsymbol{x}^{(m)} + b > 0$, and vice versa.

Alternatively, we can express the distance as

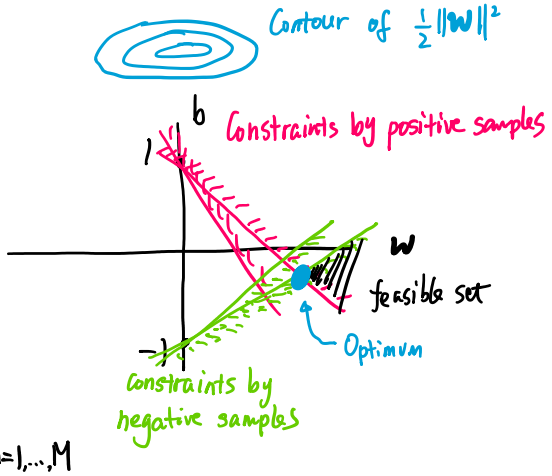
$$\frac{t^{(m)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(m)}+b\right)}{\|\boldsymbol{w}\|}$$

Thus, the training objective is to

$$\max_{\boldsymbol{w}} \min_{m} \frac{t^{(m)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(m)}+b\right)}{\|\boldsymbol{w}\|}$$

It is noted that we may scale $\boldsymbol{w}$ and b and the result is not changed. The trick is to impose a constraint

$$\min_{m} t^{(m)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(m)}+b\right) = 1$$



Then, the optimization is

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\text{maximize}} && \frac{1}{\|\boldsymbol{w}\|} \\ &\text{subject to} && \min_{m} t^{(m)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(m)}+b\right) = 1 \end{aligned}$$

and is equivalent to

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\text{minimize}} && \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} && t^{(m)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(m)}+b\right) \geq 1 \quad \text{for } m=1,\dots,M \end{aligned}$$

as the equality constraint $\min_{t} t^{(m)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(m)}+b\right)=1$ will be automatically satisfied.

Duality

A convex optimization problem has a standard form

$$\begin{aligned} &\text{minimize} && f_0(\boldsymbol{x}) \\ &\text{subject to} && f_i(\boldsymbol{x}) \leq 0, \quad i=1,\dots,m \\ &&& h_i(\boldsymbol{x}) = 0, \quad i=1,\dots,p \end{aligned}$$

f: convex
h: affine

Lagrangian is defined as

$$L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m}\lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p}\nu_i h_i(\boldsymbol{x})$$

Lagrange dual function

$$\begin{aligned} g(\boldsymbol{\lambda},\boldsymbol{\nu}) &= \inf_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu}) \\ &= \inf_{\boldsymbol{x}}\left(f_0(\boldsymbol{x}) + \sum_{i=1}^{m}\lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p}\nu_i h_i(\boldsymbol{x})\right) \end{aligned}$$

Dual problem

$$\begin{aligned} &\underset{\boldsymbol{\lambda},\boldsymbol{\nu}}{\text{maximize}} && g(\boldsymbol{\lambda},\boldsymbol{\nu}) \\ &\text{subject to} && \boldsymbol{\lambda} \succeq \boldsymbol{0} \end{aligned}$$

subject to $\boldsymbol{\lambda} \succeq \mathbf{0}$

Weak duality (always holds)

$$d^* \leq p^*$$

where d^* is the dual optimum and p^* is the primal optimum.

$$\begin{aligned} \forall \boldsymbol{\lambda}, \boldsymbol{\nu}, \quad g(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\boldsymbol{x}} \left(f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p} \nu_i h_i(\boldsymbol{x}) \right) \\ &\leq f_0(\boldsymbol{x}^*) + \sum_{i=1}^{m} \underbrace{\lambda_i}_{\geq 0} \underbrace{f_i(\boldsymbol{x}^*)}_{\leq 0} + \sum_{i=1}^{p} \nu_i \underbrace{h_i(\boldsymbol{x}^*)}_{=0} \quad \text{[because LHS is infemum]} \\ &\leq f_0(\boldsymbol{x}^*) \end{aligned}$$

Strong duality (i.e., $d^* = p^*$)

- Slater's condition (convex + strictly feasible) => strong duality
- (**KKT necessary conditions**) For a differentiable objective, if strong duality holds with $\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ being the primal and dual optimal points, then

$$\begin{aligned} f_i(\boldsymbol{x}^*) &\leq 0 \quad i = 1, \cdots m \\ h_i(\boldsymbol{x}^*) &= 0 \quad i = 1, \cdots, p \end{aligned} \Big] \text{ primal feasibility}$$

$$\lambda_i^* \geq 0 \quad i = 1, \cdots, m \quad \text{dual feasibility}$$

$$\lambda_i^* f_i(x^*) = 0 \quad i = 1, \cdots, m \quad \text{KKT complementary condition}$$

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)\big|_{\boldsymbol{x}=\boldsymbol{x}^*} = 0 \quad \text{Lagrangian gradient vanishing}$$

- (**KKT sufficiency for convex problems**) If the primal is convex, then KKT is suffient. That is, for any $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ satisfying KKT conditions, $\tilde{x}$ is the primal optimum; $\tilde{\lambda}$ and $\tilde{\nu}$ are dual optimum, and strong duality holds.

$$\begin{aligned} g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) &= L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \quad [\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \text{ vanishes at } \boldsymbol{x} = \tilde{\boldsymbol{x}}] \\ &= f_0(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{p} \tilde{\nu}_i h_i(\tilde{\boldsymbol{x}}) \\ &= f_0(\tilde{\boldsymbol{x}}) \end{aligned}$$

Solving the optimization problem for SVM

Primal: $\underset{\mathbf{w}, b}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2$

subject to $1 - t^{(m)}(\mathbf{w}^T \mathbf{x}^{(m)} + b) \leq 0 \quad \text{for } m = 1, \ldots, M$

Lagrangian: $L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{m=1}^{M} \alpha^{(m)} \left[1 - t^{(m)}(\mathbf{w}^T \mathbf{x}^{(m)} + b) \right]$

KKT gradient vanishing $\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) \overset{\text{set}}{=} \mathbf{0} \quad (1)$

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) \overset{\text{set}}{=} 0 \qquad (2)$$

(1) implies $\quad \mathbf{w} - \sum_{m=1}^{M} \alpha^{(m)} t^{(m)} \boldsymbol{x}^{(m)} = \mathbf{0}$

$$\mathbf{w} = \sum_{m=1}^{M} \alpha^{(m)} t^{(m)} \boldsymbol{x}^{(m)} \qquad (3)$$

(2) implies $\quad \sum_{m=1}^{M} \alpha^{(m)} t^{(m)} = 0 \qquad (4)$

Put (3) and (4) to the Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \underbrace{\|\mathbf{w}\|^2}_{\mathbf{w}^T\mathbf{w}} + \sum_{m=1}^{M} \alpha^{(m)} \left[1 - t^{(m)} (\mathbf{w}^T \boldsymbol{x}^{(m)} + b)\right]$$

$$= \frac{1}{2} \sum_{n=1}^{M} \sum_{m=1}^{M} \alpha^{(n)} \alpha^{(m)} t^{(n)} t^{(m)} \boldsymbol{x}^{(n)T} \boldsymbol{x}^{(m)}$$

$$+ \sum_{m=1}^{M} \alpha^{(m)} - \sum_{m=1}^{M} \alpha^{(m)} t^{(m)} \sum_{n=1}^{M} \alpha^{(n)} t^{(n)} \left[\boldsymbol{x}^{(n)}\right]^T \boldsymbol{x}^{(m)} - b \cdot \underbrace{\sum_{m=1}^{M} \alpha^{(m)} t^{(m)}}_{=0 \text{ by } (4)}$$

$$= \sum_{m=1}^{M} \alpha^{(m)} - \sum_{n=1}^{M} \sum_{m=1}^{M} \alpha^{(n)} \alpha^{(m)} t^{(n)} t^{(m)} \left[\boldsymbol{x}^{(n)}\right]^T \boldsymbol{x}^{(m)}$$

Dual:

$$\text{maximize} \quad \sum_{m=1}^{M} \alpha^{(m)} - \frac{1}{2} \sum_{n=1}^{M} \sum_{m=1}^{M} \alpha^{(n)} \alpha^{(m)} t^{(n)} t^{(m)} \left[x^{(n)}\right]^T x^{(m)}$$

$$\text{subject to} \quad \alpha_i \geq 0$$

$$\sum_{m=1}^{M} \alpha^{(m)} t^{(m)} = 0 \qquad \left[\text{Due to } \nabla_b L = 0 \quad \text{Otherwise } \min_{w,b} L(w, b, \lambda) = -\infty\right]$$

Sequential minimal optimization (SMO)

Loop until converging:

- Pick $\alpha^{(i)}, \alpha^{(j)}$
- Represent $\alpha^{(i)} = -t^{(i)} \sum_{m \neq i} \alpha^{(m)} t^{(m)}$
- Optimize $\alpha^{(j)}$ in the dual
- Project back to feasible set

Primal optimal point: If $\alpha_*^{(m)}$ is solved

$$\mathbf{w}_* = \sum_{m=1}^{M} \alpha_*^{(m)} t^{(m)} \boldsymbol{x}^{(m)}$$

$$b_* = \frac{1}{2} \left[\min_{i: t^{(i)}=1} \mathbf{w}_*^T \boldsymbol{x}^{(i)} + \max_{i: t^{(i)}=-1} \mathbf{w}_*^T \boldsymbol{x}^{(i)}\right]$$

Decision boundary

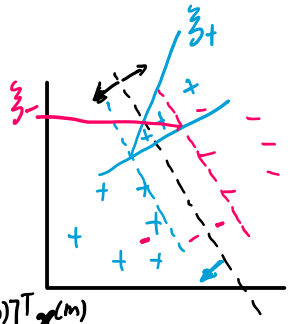
$$\mathbf{w}_*^T \boldsymbol{x} + b_* = 0$$

$$\sum_{m=1}^{M} \alpha_*^{(m)} t^{(m)} \left[\boldsymbol{x}^{(m)}\right]^T \boldsymbol{x} + b_* = 0$$

SVM with slack variables

- The above SVM only works for linearly separable problems
- Insensitive to data far away from the margin (dual sparsity)
- Sensitive to data on the margin (support vectors)

$$\underset{\boldsymbol{\xi}, \mathbf{w}, b}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{m=1}^{M}\xi^{(m)}$$
$$\text{subject to} \quad t^{(m)}(\mathbf{w}^T\boldsymbol{\phi}^{(m)} + b) \geq 1 - \xi^{(m)}$$
$$\xi^{(m)} \geq 0$$



Dual (C-SVM) [HW]

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \sum_{m=1}^{M}\alpha^{(m)} - \frac{1}{2}\sum_{n=1}^{M}\sum_{m=1}^{M}\alpha^{(n)}\alpha^{(m)}t^{(n)}t^{(m)}[\boldsymbol{\phi}^{(n)}]^T\boldsymbol{\phi}^{(m)}$$
$$\text{subject to} \quad 0 \leq \alpha^{(m)} \leq C$$
$$\sum_{m=1}^{M}\alpha^{(m)}t^{(m)} = 0$$

Alternative formulation (ν-SVM)

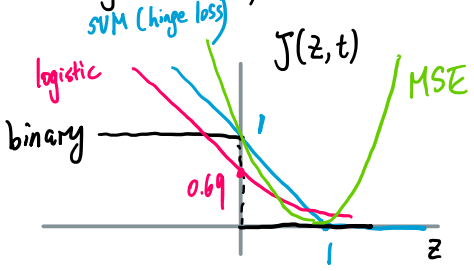
$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad -\frac{1}{2}\sum_{n=1}^{M}\sum_{m=1}^{M}\alpha^{(n)}\alpha^{(m)}t^{(n)}t^{(m)}[\boldsymbol{\phi}^{(n)}]^T\boldsymbol{\phi}^{(m)}$$
$$\text{subject to} \quad 0 \leq \alpha^{(m)} \leq \frac{1}{M}$$
$$\sum_{m=1}^{M}a^{(m)}t^{(m)} = 0$$
$$\sum_{m=1}^{M}a^{(m)} \geq \nu$$

$\nu \in (0,1]$: upper-bound of fraction of support vectors

Ref: https://www.csie.ntu.edu.tw/~cjlin/papers/nusvmtutorial.pdf

Comparing SVM loss, logistic regression loss, and binary risk

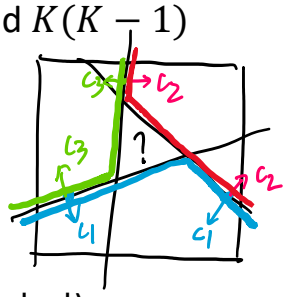
Assume target $t = 1$, $z = \mathbf{w}^T\boldsymbol{\phi} + b$



Multi-class SVM

- SVM intrinsically works for binary classification
- Multi-class SVM also works well in practice, but usually with inelegant treatment

- One-vs-one
  - For K-category classification, build $K(K-1)$ classifiers, and do the voting. The result may be ambiguous.
- One-vs-the-rest (one-vs-all)
  - Train K classifiers and pick the one category with the largest margin (not normalized by $|\boldsymbol{w}|$).



- More structured training

Constraints: For any data point $x^{(m)}, t^{(m)}$

$$t^{(m)} = i \in \{1, \cdots, K\}$$

$$\boldsymbol{w}_i^T \boldsymbol{x}^{(m)} + b_i - (\boldsymbol{w}_j^T \boldsymbol{x}^{(m)} + b_j) \geq 1 - \xi^{(m)} \quad \forall j \neq i,\ j \in \{1, \cdots, K\}$$

Kernel tricks

- Idea: Both training and prediction in SVM only depend on inner-product. We may specify the inner-product in some non-linearly transformed space without explicit representation.

Hilbert Space

- **Vector space** V

  Commutativity $\forall \boldsymbol{u}, \boldsymbol{v} \in V, \boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$
  Associativity $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V,\ (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} = \boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w})$
  Additive identity $\exists \boldsymbol{0} \in V, \forall \boldsymbol{v} \in V,\ \boldsymbol{v} + \boldsymbol{0} = \boldsymbol{v}$
  Additive inverse $\forall \boldsymbol{v} \in V, \exists \boldsymbol{w} \in V, \boldsymbol{v} + \boldsymbol{w} = \boldsymbol{0}$
  Multiplicative identity $\forall \boldsymbol{v} \in V, 1\boldsymbol{v} = \boldsymbol{v}$
  Distributive properties $\forall a, b \in F, \boldsymbol{u}, \boldsymbol{v} \in V$
  $a(\boldsymbol{u} + \boldsymbol{v}) = a\boldsymbol{u} + a\boldsymbol{v}$ and $(a + b)\boldsymbol{u} = a\boldsymbol{u} + b\boldsymbol{u}$

- **Inner-product space** is a vector space, with inner product
  $\forall \boldsymbol{u}, \boldsymbol{v} \in V, \quad \langle \boldsymbol{u}, \boldsymbol{v} \rangle \in F$
  satisfying the following properties

  Positivity $\forall \boldsymbol{v} \in V, \langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0$
  Definiteness $\langle \boldsymbol{v}, \boldsymbol{v} \rangle = 0$ iff $\boldsymbol{v} = \boldsymbol{0}$
  Additivity in first slot $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$
  $\langle \boldsymbol{u} + \boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{w} \rangle + \langle \boldsymbol{v}, \boldsymbol{w} \rangle$
  [also holds for the second slot]
  Homogeneity in first slot $\forall a \in F, \forall\, \boldsymbol{v}, \boldsymbol{w} \in V$
  $\langle a\boldsymbol{v}, \boldsymbol{w} \rangle = a\langle \boldsymbol{v}, \boldsymbol{w} \rangle$
  [second slot needs conjugate]
  Conjugate symmetry $\forall \boldsymbol{v}, \boldsymbol{w} \in W,\ \langle \boldsymbol{v}, \boldsymbol{w} \rangle = \overline{\langle \boldsymbol{w}, \boldsymbol{v} \rangle}$

  The norm $\|\cdot\| : V \to \mathbb{R}$ is defined as

$$\|v\| = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}$$

- **Hilbert space** is a complete inner-produce space

  E.g. A finite-dimensional vector space is a Hilbert space.

  E.g. $P_m(F)$, mth degree polynomial functions with coeffients in F, equipped with the following inner-produce

  $$\langle p, q \rangle = \int_0^1 p(x)\overline{q(x)}\, dx$$

  forms a Hilbert space.

Mercer Kernel: $K: X \times X \to R$

satisfying $\iint f(u)K(u,v)f(v)du\, dv \geq 0$

for all square-integrable function f, i.e., $\int |f(x)|^2 dx < +\infty$

Discrete analog

$\boldsymbol{u}^\top K \boldsymbol{u} \succcurlyeq 0$ [PSD]

Example. $K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^\top \boldsymbol{z}$

Then, $\phi(\boldsymbol{x}) = \boldsymbol{x}$

Example (Polynomial kernel). $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^\top \boldsymbol{z} + 1)^Q$

$K(\boldsymbol{x}, \boldsymbol{z}) = (x_1 z_1 + x_2 z_2 + \cdots + x_d z_d + 1)^Q$

Then, ϕ: all polynomial features up to the degree of Q with some coinefficients

Example. Gaussian kernel, Radius Basis Function (RBF) kernel

$K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma ||\boldsymbol{x} - \boldsymbol{z}||^2)$

Consider a simple case $x, z \in \mathbb{R}$, $\gamma = 1$

$$K(x,z) = \exp(-(x-z)^2)$$
$$= \exp(-x^2)\exp(-z^2)\exp(2xz)$$
$$= \exp(-x^2)\exp(-z^2)\sum_{k=0}^{\infty}\frac{(2xz)^k}{k!}$$
$$= \exp(-x^2)\exp(-z^2)\sum_{k=0}^{\infty}\frac{2^k x^k z^k}{k!}$$
$$= \sum_{k=0}^{\infty}\left(\sqrt{\frac{2^k}{k!}}\exp(-x^2)x^k\right)\left(\sqrt{\frac{2^k}{k!}}\exp(-z^2)z^k\right)$$

Thus, ϕ could be

$$x \mapsto \left(\sqrt{\frac{2}{1!}}\exp(-x^2)\cdot x,\ \sqrt{\frac{2^2}{2!}}\exp(-x^2)\cdot x^2,\ \cdots\right)$$

Note: the feature mapping is not unique.

Reproducing kernel Hilbert space is a set of $f: \mathcal{X} \to \mathbb{R}$

$\mathcal{H} = \text{span}\{K(x,\cdot) \mid x \in \mathcal{X}\}$

Reproducing property:

$\forall f \in \mathcal{H}$, $\langle f(\cdot), K(\cdot, x)\rangle_{\mathcal{H}} = f(x)$

$\langle f, \phi_x\rangle_{\mathcal{H}}$

In particular $\langle K(\cdot,x), K(\cdot,y)\rangle_{\mathcal{H}} = K(x,y)$

Operations preserving the kernel

$k(x,z) = c k_1(x,z)$	$\forall$ kernel k_1, $\forall c>0$
$k(x,z) = f(x)\, k_1(x,z)\, f(z)$	$\forall f$
$k(x,z) = q(k_1(x,z))$	$\forall$ polynomials of nonnegtive coefficients
$k(x,z) = \exp(k_1(x,z))$	
$k(x,z) = k_1(x,z) + k_2(x,z)$	$\forall$ kernel k_2

...

Summary

- Support vectors machines originally work for binary, linearly separable classification.
- Soft SVM is still a linear classifier, but deals with non-linearly separable tasks by slack variables.
- Kernel SVM could be a non-linear classifier for input, which accomplishes linear classification in an implicit Hilbert space.
    - The linear classification is oftentimes not representable by $\boldsymbol{w}, b$ in the Hilbert space, but can be fully specified by support vectors, which could in turn be thought of as model parameters. Thus, the number of parameters of kernel SVM grows with data samples. This is known as a non-parametric model.
    - The model capacity shall not be viewed as the dimension of the Hilbert space.
      A better measure of SVM's model capacity is the number of support vectors, which is usually small.
      [Recall the ν-SVM]
      ...

- This gives a clever tradeoff between overfitting (too powerful hypothesis class) and underfitting (too restricted hypothesis class).