Representer Theorem (表示定理)

- By default, we use $\Phi(x) \in \mathbb{R}^{d'}$ to replace $x \in \mathbb{R}^d$
- e.g. $\Phi(x) = [x^T, 1]^T$, then $f(x) = w^T\Phi(x)$ can recover $w^Tx + b$, which is all linear models' uniform form

Uniform Form

- 我们之前学过的回归和SVM均可以写成统一的格式
  - Ridge (Linear) Regression

$$\underset{w}{min} \sum_{i=1}^{n}(w^T\Phi(x_i) - y_i)^2 + \lambda||w||^2$$

  - Ridge (Logistic) Regression

$$\underset{w}{min} \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T\Phi(x_i))) + \lambda||w||^2$$

  - Soft-margin SVM

$$\begin{aligned}&\underset{w}{min} \frac{1}{2}||w||^2 + c\sum_{i=1}^{n} \max\{0, 1 - y_i w^T\Phi(x_i)\}\\ =&\underset{w}{min} \sum_{i=1}^{n} \max\{0, 1 - y_i w^T\Phi(x_i)\} + \lambda||w||^2\end{aligned}$$

- 更进一步，我们可以把这种统一的格式改写成一致的核函数的形式
  - Linear Regression

$$\begin{aligned}&\frac{\partial L}{\partial w} = 2\sum_{i=1}^{n}(w^T\Phi(x_i) - y_i)\Phi(x_i) + 2\lambda w = 0\\ \Rightarrow &w = \frac{1}{\lambda}\sum_{i=1}^{n}(y_i - w^T\Phi(x_i))\Phi(x_i) := \sum_{i=1}^{n}\alpha_i\Phi(x_i)\\ \Rightarrow &f(x) = \sum_{i=1}^{n}\alpha_i\Phi^T(x_i)\Phi(x) := \sum_{i=1}^{n}\alpha_i K(x_i, x)\end{aligned}$$

  - Logistic Regression

$$\begin{aligned}&\frac{\partial L}{\partial w} = -\sum_{i=1}^{n}\frac{\exp(-y_i w^T\Phi(x_i))}{1 + \exp(-y_i w^T\Phi(x_i))}y_i\Phi(x_i) + 2\lambda w = 0\\ \Rightarrow &w = \sum_{i=1}^{n}\frac{1}{2\lambda}\sigma(-y_i w^T\Phi(x_i))y_i\Phi(x_i) := \sum_{i=1}^{n}\alpha_i\Phi(x_i)\\ \Rightarrow &f(x) = \sum_{i=1}^{n}\alpha_i\Phi^T(x_i)\Phi(x) := \sum_{i=1}^{n}\alpha_i K(x_i, x)\end{aligned}$$

  - Soft-margin SVM
    - introduce $\xi_i,\ \ \xi_i \geq 0,\ \ \xi_i \geq 1 - y_i w^T\Phi(x)$
    - Dual form

$$
\begin{aligned}
&\max_{\alpha,\beta\geq 0}\min_{w,\xi} \lambda w^T w+\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\alpha_i(1-y_i w^T\Phi(x_i)-\xi_i)-\sum_{i=1}^{n}\beta_i\xi_i \\
\Rightarrow &\begin{cases} w=\frac{1}{2\lambda}\sum_{i=1}^{n}\alpha_i y_i\Phi(x_i):=\sum_{i=1}^{n}\tilde{\alpha}_i\Phi(x_i) \\ \alpha_i+\beta_i=1\end{cases} \\
\Rightarrow & f(x)=\sum_{i=1}^{n}\tilde{\alpha}_i\Phi^T(x_i)\Phi(x):=\sum_{i=1}^{n}\tilde{\alpha}_i K(x_i,x)
\end{aligned}
$$

Reproducing Kernel Hilbert Space (再生核希尔伯特空间)

- Vector Space(Euclidean Space $\mathbb{R}^d$) $\rightarrow$ possibly infinite dimensional Hilbert space $\mathcal{H}$(vector space inner product)
  - function $f\in\mathcal{H}$ can be understood as infinite dimensional vector $[f(x_1),\dots,f(x_\infty)]$
- 该空间满足以下性质:
  - $<f,g>_\mathcal{H}=<g,f>_\mathcal{H}$
  - $<a_1f_1+a_2f_2,g>_\mathcal{H}=a_1<f_1,g>_\mathcal{H}+a_2<f_2,g>_\mathcal{H}$
  - $<f,f>_\mathcal{H}\geq 0,\quad <f,f>_\mathcal{H}=0\Leftrightarrow f=0$
- *RKHS*
  - $f\in\mathcal{H}$, $\mathcal{H}$ is a Hilbert space of real-valued functions $f:X\rightarrow\mathbb{R}$
  - $\mathcal{H}$ is associated with a kernel $k(\cdot,\cdot),\quad s.t.\ \ f(x)=<f,k(\cdot,\cdot)>_\mathcal{H}$
  - $k(\cdot,\cdot)$ is called the **reproducing kernel** of RKHS
  - k maps every $x\in X$ to a point in $\mathcal{H}$, k produces $g\in\mathcal{H}$ that $g(z)=k(z,x)$
  - $\mathcal{H}:=\text{Completion of}\{k(\cdot,x)|x\in X\}$
- $<k(z,\cdot),k(\cdot,x)>_\mathcal{H}=k(z,x)$
- Example 1
  - $f(x)=w^Tx$ defines $k(z,x)=z^Tx$ is the reproducing kernel of $\mathcal{H}$
  - $f(x)=<k(w,\cdot),k(\cdot,x)>_\mathcal{H}=k(w,x)=w^Tx$
- Example 2
  - $f(x)=w^T\Phi(x)$ defines $k(z,x)=\Phi^T(z)\Phi(x)$
  - $f(x)=<k(\Phi^{-1}(w),\cdot),k(\cdot,x)>_\mathcal{H}=w^T\Phi(x)$
  - Φ includes all polynomials that can proximate any functions
- *Most function spaces are RKHS*
- Typically, a function $f:X\rightarrow\mathbb{R},\ f\in\mathcal{H}$ can be represented as

$$
f(\cdot)=\sum_{i=1}^{\infty}\alpha_i k(x_i,\cdot)
$$

  - $k(x_i,\cdot)$ are basis of $\mathcal{H}$
  - then we have

$$
\begin{aligned}
<f,k(\cdot,x)>_\mathcal{H}&=<\sum_{i=1}^{\infty}\alpha_i k(x_i,\cdot),k(\cdot,x)>_\mathcal{H}\\
&=\sum_{i=1}^{\infty}\alpha_i k(x_i,x)=f(x)
\end{aligned}
$$

- Norm of f
  - $||f||^2_\mathcal{H}=<f,f>_\mathcal{H}$
  - if $f(x)=w^T\Phi(x)$, then

$$
\begin{aligned}
<f,f>_\mathcal{H}&=<k(\Phi^{-1}(w),\cdot),k(\cdot,\Phi^{-1}(w))>_\mathcal{H}\\
&=k(\Phi^{-1}(w),\Phi^{-1}(w))\\
&=w^Tw
\end{aligned}
$$

Formal Form

- Consider a RKHS $\mathcal{H}$ with representing kernel $k:X\times X\rightarrow\mathbb{R}$. Given training data $\{(x_1,y_1),\dots,(x_n,y_n)\}\in X\times\mathbb{R}$, a strictly increasing regularization function $R:[0,+\infty)\rightarrow\mathbb{R}$ and a loss function $L:\mathbb{R}\times\mathbb{R}\rightarrow\mathbb{R}$, then any $f\in\mathcal{H}$ that minimizes

$\sum_{i=1}^n L(f(x_i), y_i) + R(\|f\|)$ can be presented as $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$

- *Proof:*

  Decompose f into two functions, one lying in span $\{k(x_1,\cdot),\dots,k(x_n,)\}$, the other component orthogonal to it.
  $f = \sum_{i=1}^n \alpha_i k(x_i,\cdot) + \mu \quad \text{s.t.} \; <\mu, k(x_i,\cdot)>=0$
  Applying f to any training example x_j, we have

$$
\begin{aligned}
f(x_j) &= \sum_{i=1}^n \alpha_i k(x_i, x_j) + \mu(x_j) \\
&= \sum_{i=1}^n \alpha_i k(x_i, x_j) + <\mu, k(x_j,\cdot)>
\end{aligned}
$$

  So the loss $\sum_{i=1}^n L(f(x_i), y_i)$ is independent of μ, then:

$$
\begin{aligned}
R(\|f\|) &= R(\|\sum_{i=1}^n \alpha_i k(x_i,\cdot) + \mu\|) \\
\|f_0 + \mu\| &= \sqrt{<f_0+\mu, f_0+\mu>_{\mathcal{H}}} \\
&= \sqrt{<f_0, f_0>_{\mathcal{H}} + 2<f_0,\mu>_{\mathcal{H}} + <\mu,\mu>_{\mathcal{H}}} \\
&= \sqrt{\|f_0\|^2 + \|\mu\|^2} \\
{}_{i=1}^{n} R(\|f\|) &= R(\sqrt{\|f_0\|^2 + \|\mu\|^2}) \geq R(\|f_0\|) \quad \text{取等当且仅当}\mu = 0
\end{aligned}
$$

  Regularization minimizes at $\mu = 0 \Rightarrow f = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$

- Significance
  - Turns a potentially infinite dimension optimization of f into a search of $\alpha_1,\dots,\alpha_n$
  - Shows that a wide range of learning algorithms have solutions expressed as **weight sum of kernel functions** on finite training data $f = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$
- Ridge Linear Regression:

$$
\min_w J(w) = \frac{1}{2}\sum_{i=1}^n (w^T\Phi(x_i) - y_i)^2 + \frac{1}{2}\lambda w^T w
$$

According to Representer Theorem:

$$
\begin{aligned}
f(x) = \sum_{i=1}^n \alpha_i k(x_i, x) &= \sum_{i=1}^n \alpha_i \Phi(x_i)^T \Phi(x_i) \\
\Rightarrow w &= \sum_{i=1}^n \alpha_i \Phi(x_i)
\end{aligned}
$$

Let

$$
\Phi = \begin{pmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{pmatrix} \in \mathbb{R}^{n\times d'} \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \quad \Rightarrow \quad w = \Phi^T \alpha
$$

So we will have

$$
\begin{aligned}
J(w) &= \frac{1}{2}(\Phi w - y)^T(\Phi w - y) + \frac{\lambda}{2} w^T w \\
&= \frac{1}{2}(w^T\Phi^T\Phi w + y^T y - 2w^T\Phi^T y) + \frac{\lambda}{2} w^T w \\
&= \frac{1}{2}\alpha^T\Phi\Phi^T\Phi\Phi^T\alpha + \frac{1}{2}y^T y - \alpha^T\Phi\Phi^T y + \frac{\lambda}{2}\alpha^T\Phi\Phi^T\alpha \\
\Phi\Phi^T &= \begin{pmatrix} \phi(x_1)^T\phi(x_1) & \dots & \phi(x_1)^T\phi(x_n) \\ \vdots & \ddots & \vdots \\ \phi(x_n)^T\phi(x_1) & \dots & \phi(x_n)^T\phi(x_n) \end{pmatrix} = \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix} = K \quad \text{(Gram Matrix)}
\end{aligned}
$$

K是实对称矩阵，半正定矩阵

$$
\begin{aligned}
J(w) &= \frac{1}{2}\alpha^T KK\alpha + \frac{1}{2}y^T y - \alpha^T Ky + \frac{\lambda}{2}\alpha^T K\alpha := J(\alpha) \\
\frac{\partial J(\alpha)}{\partial \alpha} &= KK\alpha - Ky + \lambda K\alpha = 0 \\
\text{if} \quad & K \succ 0 \quad (\text{K正定且对角线占优}) \\
&\Rightarrow (K + \lambda I)\alpha = y \\
&\Rightarrow \alpha = (K + \lambda I)^{-1} y
\end{aligned}
$$

Define

$$
k(x) \in \mathbb{R}^n, \quad k(x) = \begin{pmatrix} k(x_1, x) \\ \vdots \\ k(x_n, x) \end{pmatrix}
$$

So

$$
\begin{aligned}
f(x) &= \sum_{i=1}^{n} \alpha_i k(x_i, x) \\
&= k(x)^T \alpha \\
&= k(x)^T (K + \lambda I)^{-1} y
\end{aligned}
$$

$$
\lambda \to \infty \Rightarrow K + \lambda I \to \lambda I \Rightarrow f(x) \approx \frac{1}{\lambda} k(x)^T y = \frac{1}{\lambda} \sum_{i=1}^{n} k(x_i, x) y_i
$$

$$
f(x) = w^T \phi(x) = \phi(x)^T w = \phi(x)^T (\Phi\Phi^T + \lambda I)^{-1} \Phi^T y = k(x)^T (K + \lambda I)^{-1} y
$$