Kernel Methods & Regularization

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2014

Outline

1. **What Learning Theory Taught Us?**
   - Generalization Performance
   - Kernel Methods & Regularization
2. **Extended Linear Models**
   - Regularized Nonlinear Regression
   - Regularized Least Square Classification
3. **Kernels and RKHS**

Outline

1. **What Learning Theory Taught Us?**
   - Generalization Performance
   - Kernel Methods & Regularization
2. **Extended Linear Models**
   - Regularized Nonlinear Regression
   - Regularized Least Square Classification
3. **Kernels and RKHS**

Outline

1. **What Learning Theory Taught Us?**
   - Generalization Performance
   - Kernel Methods & Regularization
2. **Extended Linear Models**
   - Regularized Nonlinear Regression
   - Regularized Least Square Classification
3. **Kernels and RKHS**

Generalization Error

- Assuming a hypothesis class $\mathcal{H}$, let $h \in \mathcal{H}$ be the hypothesis trained from the dataset $\mathcal{X} = \{(\boldsymbol{x}^{(t)}, r^{(t)})\}_{t=1}^{N}$ by minimizing the empirical error: $R_{emp}[h] := \frac{1}{N}\sum_{t=1}^{N} l(h(\boldsymbol{x}^{(t)}), r^{(t)})$
- Generalization error of h:
  $R[h] := \int p(\boldsymbol{x}, r) l(h(\boldsymbol{x}), r) d(\boldsymbol{x}, r) = E_{\mathcal{I}\times\mathcal{L}}[l(h(\boldsymbol{x}), r)]$
- Let $h^* := \arg\inf_{g\in\mathcal{H}} R[g]$ and $R^* := \inf_{f:\mathcal{I}\to\mathcal{L}} R[f]$
- Instead, our ultimate goal is to have $R[h] \to R^*$!
- $R[h] - R^* = (R[h^*] - R^*) + (R[h] - R[h^*])$ Model太簡單，無法approximate
  - $R[h^*] - R^*$ is called the ***approximation error***: we need a complex model to reduce this
  - $R[h] - R[h^*]$ is called the ***estimation error***: we need a simple model

Model太複雜，會Overfit

Which Models to Assume?

- We can assume different models $\mathcal{H}_1$, $\mathcal{H}_2$, $\cdots$ and compares them in the model selection process
- But there are too many different models
- This lecture introduces methods that allow the complexity of $\mathcal{H}$ to be tuned ***after*** it is assumed
  - Simplifies the task of model assumption
  - Linear models suffice in most cases
- Does ***not*** mean $\mathcal{H}$ will have the right complexity
  - Model selection is still required

Outline

Kernel Methods

- To reduce approximation error, we need a complex model
- But complex model make the objective hard to solve
- One common approach, called ***kernel methods***, is to map/lift/kernelize examples $\mathcal{X} = \{\boldsymbol{x}^{(t)}, r^{(t)}\}_{t=1}^{N}$ from original ***input space*** to a higher dimensional ***feature space*** $\Phi(\mathcal{X}) = \{\Phi(\boldsymbol{x}^{(t)}), r^{(t)}\}_{t=1}^{N}$
  - $k(\boldsymbol{a}, \boldsymbol{b}) := \langle \Phi(\boldsymbol{a}), \Phi(\boldsymbol{b}) \rangle$ is called the ***kernel function***

不是讓hypothesis class更複雜
而是把data變複雜

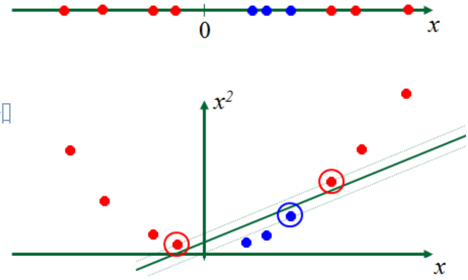
- E.g., suppose $x \in \mathbb{R}$, $\Phi(x) = [x, x^2]^\top \in \mathbb{R}^2$
- Why does it work?

如果objective有做內積的可能
那麼feature space也要可以做內積

Kernel Methods

- To reduce approximation error, we need a complex model
- But complex model make the objective hard to solve
- One common approach, called ***kernel methods***, is to map/lift/kernelize examples $\mathcal{X} = \{\boldsymbol{x}^{(t)}, r^{(t)}\}_{t=1}^{N}$ from original ***input space*** to a higher dimensional ***feature space*** $\Phi(\mathcal{X}) = \{\Phi(\boldsymbol{x}^{(t)}), r^{(t)}\}_{t=1}^{N}$
  - $k(\boldsymbol{a}, \boldsymbol{b}) := \langle \Phi(\boldsymbol{a}), \Phi(\boldsymbol{b}) \rangle$ is called the ***kernel function***

- E.g., suppose $x \in \mathbb{R}$, $\Phi(x) = [x, x^2]^\top \in \mathbb{R}^2$
- Why does it work? 看右邊就知
- Note a linear h in the feature space is ***not linear*** in the input space anymore



Regularization

- To reduce estimation error, we need a simple model
- Note that in the same $\mathcal{H}$, there are still hypotheses that are more complex than the others
- We can add a term, called ***regularization term***, in our objective that penalizes complex hypotheses in $\mathcal{H}$:

$$\arg\min_{g \in \mathcal{H}} \sum_{t=1}^{N} \underbrace{l(g(\boldsymbol{x}^{(t)}), r^{(t)})}_{\text{data term}} + \underbrace{\lambda \Omega(g)^2}_{\text{smoothness term}},$$

如果g太複雜 就提高「人造的」誤差

where

  - Ω is a smoothness measure
  - λ is a ***hyperparameter*** (fixed during the training process), which controls the trade-off between a) minimizing the empirical error; and b) maximizing function smoothness
- Why does it work? A "smoother" h allows unseen instances to learn values from those of nearby examples

Outline

Outline

Linear Regression

- Let $\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \ddots & \vdots \\ x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$, $\boldsymbol{w} = [w_1, \cdots, w_d]^\top$, and $\boldsymbol{r} = [r^{(1)}, r^{(2)}, \cdots, r^{(N)}]^\top$
- The linear regression problem:

$$\arg\min_{\boldsymbol{w}, b} \|\boldsymbol{r} - \begin{bmatrix} \boldsymbol{1} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix} \|^2$$

  - b is called the bias term
- Solution: $\begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}^* = (\begin{bmatrix} \boldsymbol{1} & \boldsymbol{X} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{1} & \boldsymbol{X} \end{bmatrix})^{-1} \begin{bmatrix} \boldsymbol{1} & \boldsymbol{X} \end{bmatrix}^\top \boldsymbol{r}$ if $\begin{bmatrix} \boldsymbol{1} & \boldsymbol{X} \end{bmatrix}$ has full column rank

Closed form solution $(X^T X)^{-1} X^T r$

Kernelization

- Objective:
$$\arg\min_{\boldsymbol{w},b} \|\boldsymbol{r} - \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}\|^2,$$
where $\boldsymbol{X} = \begin{bmatrix} \Phi(\boldsymbol{x}^{(1)})^\top \\ \vdots \\ \Phi(\boldsymbol{x}^{(N)})^\top \end{bmatrix}$
- Note that the number of variables to solve now becomes (dimension of feature space + 1)

Regularization

- The ***regularized least square problem***:

$$\arg\min_{\boldsymbol{w},b} \|\boldsymbol{r} - \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix} \|^2 + \lambda \|\boldsymbol{w}\|^2,$$

  - The bias term b is not regularized
- Why minimizing $\|\boldsymbol{w}\|^2$? A flat h learns from all examples (by the average of their label values)
- Can be expressed as an ordinary least squares:
  $\arg\min_{\boldsymbol{w},b} \|\widetilde{\boldsymbol{r}} - \widetilde{\boldsymbol{X}} \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix} \|^2$, where
  $\widetilde{\boldsymbol{X}} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \\ 0 & \mathbf{0} \\ \mathbf{0} & \sqrt{\lambda}\boldsymbol{I}_d \end{bmatrix} \in \mathbb{R}^{(N+d+1)\times(d+1)}$ and $\widetilde{\boldsymbol{r}} = [\boldsymbol{r}, 0]^\top \in \mathbb{R}^{N+d+1}$
  [Proof]
- Solution: $\begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}^* = (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{r}} =$
  $(\begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix}^\top \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} + \lambda \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \sqrt{\lambda}\boldsymbol{I}_d \end{bmatrix})^{-1} \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix}^\top \boldsymbol{r}$
  - $\widetilde{\boldsymbol{X}}$ must be full column rank

The Bias Term

- With some particular kernel functions, we can simply set $b = 0$
- Simplified objective:

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2$$

- Solution: $\boldsymbol{w}^* = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top\boldsymbol{r} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_N)^{-1}\boldsymbol{r}$
- In a very high (or infinite) dimensional feature space, this usually makes little difference in performance

  在越高維度下，越不需要bias term也可以解釋資料，
  無限維下更是如此，因為可以透過各維度的線性組合，
  產生與bias term等效的結果，因此便少了一項

Outline

Regularized Least Square Classification

- $\boldsymbol{r} \in \{1, -1\}^N$
- Recall that in linear case, we cannot directly apply linear regression to the classification problem
  - Why? The linear model is obviously "too simple" such that the SSE will be large for all lines (large bias)
  - We therefore assumed the "logistic" model
- Can we directly apply regularized nonlinear regression to the classification problem?
  - Yes, as given a sufficiently high dimensional feature space, the nonlinear model will always be complex enough to produce low SSE
- This is called ***regularized least square classification***

Questions?

- To maximize the flexibility of model complexity, we want a Φ that maps $\boldsymbol{x}$'s to a feature space of dimension as high as possible
  - Ideally, to an infinite dimensional feature space
  - Q1: How to obtain such an Φ?
- Meanwhile, the number of variables to solve (in $\boldsymbol{w}$) increases as the dimension of feature space becomes higher
  - Q2: How to solve $\boldsymbol{w}$ in an infinite dimensional feature space?

Outline

Common Kernel Functions

Φ(a) = 1, a, a² Φ(b) = 1, b, b²
〈Φ(a), Φ(b)〉= (係數未標準化)
1, a, a², b, ab, b² ...

- Kernel function: $k(\boldsymbol{a},\boldsymbol{b}) := \langle \Phi(\boldsymbol{a}), \Phi(\boldsymbol{b}) \rangle$
- Polynomial kernel: $k(\boldsymbol{a},\boldsymbol{b}) = (\boldsymbol{a}^\top \boldsymbol{b}/\alpha + \beta)^\gamma$
  - E.g., let $\alpha = 1$, $\beta = 1$, $\gamma = 2$ and $\boldsymbol{a} \in \mathbb{R}^2$, then $\Phi(\boldsymbol{a}) = [1, \sqrt{2}a_1, \sqrt{2}a_2, a_1^2, a_2^2, \sqrt{2}a_1 a_2]^\top \in \mathbb{R}^6$

常用的
無限維
Kernel

- Gaussian Radial Basis Function (RBF)[1] kernel: $k(\boldsymbol{a},\boldsymbol{b}) = \exp(-\frac{\|\boldsymbol{a}-\boldsymbol{b}\|_2^2}{2\sigma^2})$ or $\exp(-\gamma\|\boldsymbol{a}-\boldsymbol{b}\|_2^2)$, $\gamma \geqslant 0$
  - $k(\boldsymbol{a},\boldsymbol{b}) = \exp(-\gamma\|\boldsymbol{a}-\boldsymbol{b}\|_2^2) = \exp(-\gamma\|\boldsymbol{a}\|^2 + 2\gamma\boldsymbol{a}^\top\boldsymbol{b} - \gamma\|\boldsymbol{b}\|^2) = \exp(-\gamma\|\boldsymbol{a}\|^2 - \gamma\|\boldsymbol{b}\|^2)(1 + \frac{2\gamma\boldsymbol{a}^\top\boldsymbol{b}}{1!} + \frac{(2\gamma\boldsymbol{a}^\top\boldsymbol{b})^2}{2!} + \cdots)$
  - Let $\boldsymbol{a} \in \mathbb{R}^2$, then $\Phi(\boldsymbol{a}) = \exp(-\gamma\|\boldsymbol{a}\|^2)[1, \sqrt{\frac{2\gamma}{1!}}a_1, \sqrt{\frac{2\gamma}{1!}}a_2, \sqrt{\frac{2\gamma}{2!}}a_1^2, \sqrt{\frac{2\gamma}{2!}}a_2^2, 2\sqrt{\frac{\gamma}{2!}}a_1 a_2, \cdots]^\top \in \mathbb{R}^\infty$
- α, β, γ, and σ are hyperparameters in the objective

[1] A radial basis function of $\boldsymbol{a}$ and $\boldsymbol{b}$ is a real-valued function whose values depend only on $\|\boldsymbol{a}-\boldsymbol{b}\|$

Questions Revisited

- Q1: How to obtain a feature mapping Φ whose range is infinite dimensional?
- Q2: How to solve $\boldsymbol{w}$ in an infinite dimensional feature space?
- Our previous definition of Φ over Gaussian RBF kernel answers Q1, but not Q2
- To answer Q2, we need a new perspective on Φ

Why Another Perspective?

- Recall that in regularized linear regression ($\arg\min_{\boldsymbol{w}} \|\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w}\|^2+\lambda\|\boldsymbol{w}\|^2$ with the bias term b omitted), we have $\boldsymbol{w}^*=(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top\boldsymbol{r}$
- Indeed, it can be shown that $\boldsymbol{w}^*=(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top\boldsymbol{r}=\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top+\lambda\boldsymbol{I}_N)^{-1}\boldsymbol{r}$ [Homework]
- Letting $\boldsymbol{c}=(\boldsymbol{X}\boldsymbol{X}^\top+\lambda\boldsymbol{I}_N)^{-1}\boldsymbol{r}$ we see that $\boldsymbol{w}^*=\sum_{t=1}^N c_t\boldsymbol{x}^{(t)}$ is a linear combination of the examples

使empirical error最小的w*
是來自於data的線性組合

c_t 是d×1

- Given any lifting Φ, if $\boldsymbol{w}$ always admit the form $\boldsymbol{w}=\sum_{t=1}^N c_t\Phi(\boldsymbol{x}^{(t)})$ for some $\boldsymbol{c}$, we can instead solve:

看似要解無限維的w
但是w必然是c的線性組合
所以只要解N維的c

$$\arg\min_{\boldsymbol{c}} \|\boldsymbol{r}-\boldsymbol{K}\boldsymbol{c}\|^2+\lambda\boldsymbol{c}^\top\boldsymbol{K}\boldsymbol{c}$$

  - $\|\boldsymbol{w}\|^2=\boldsymbol{w}^\top\boldsymbol{w}=\sum_{i=1}^N\sum_{j=1}^N c^{(i)}\boldsymbol{x}^{(i)\top}\boldsymbol{x}^{(j)}c^{(j)}=\boldsymbol{c}^\top\boldsymbol{K}\boldsymbol{c}$
- The number of variables to solve (in $\boldsymbol{c}\in\mathbb{R}^N$) now becomes independent with the dimension of feature space!

Kernels

Definition

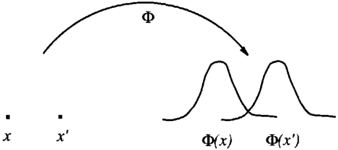
Given a function $k : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$, let $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ be the kernel matrix where $K_{i,j} := k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \langle \Phi(\boldsymbol{x}^{(i)}), \Phi(\boldsymbol{x}^{(j)}) \rangle$. Then k is a called a ***kernel function*** iff $\boldsymbol{K}$ is positive semidefinite.

正半定才有Convex optimization可以用

- E.g., $k(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^\top \boldsymbol{b}$,
  - As $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$ and for any $\boldsymbol{v}$, we have $\boldsymbol{v}^\top \boldsymbol{K} \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} = \|\boldsymbol{X}^\top \boldsymbol{v}\|^2 \geqslant 0$
- E.g., polynomial and Gaussian RBF [Proof]

function space：代一個數字，得到一個function

- We can define a lifting as $\Phi(\boldsymbol{x}) = k(\boldsymbol{x}, \cdot)$, where k is a kernel function
  lifting function是x的function，
  定義為k代入x所得到的function
- And define a complete Hilbert space as the collection of all $\sum_{i=1}^{n} \alpha^{(i)} k(\boldsymbol{x}^{(i)}, \cdot)$
  - n, $c^{(i)}$, and $\boldsymbol{x}^{(i)}$ are all arbitrary任意的
  - Infinite dimensional



- $\Phi(\boldsymbol{x})$ is a ***function***

- With the inner product:[2] $\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha^{(i)} \beta^{(j)} k(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ for any $f = \sum_{i=1}^{n} \alpha^{(i)} k(\boldsymbol{x}^{(i)}, \cdot)$ and $g = \sum_{j=1}^{m} \beta^{(j)} k(\boldsymbol{y}^{(j)}, \cdot)$
  - Well-defined [Homework] k(y) 其實是基底
- ***Reproducing properties***: $\langle f, k(\boldsymbol{y}, \cdot) \rangle = f(\boldsymbol{y})$ and $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle = \langle k(\boldsymbol{x}, \cdot), k(\boldsymbol{y}, \cdot) \rangle = k(\boldsymbol{x}, \boldsymbol{y})$ [Proof]

內積有良好定義的空間，
可以稱為Hilbert space

[2] A vector space is called a Hilbert space if it is endowed with an inner product, and is complete if every Cauchy sequence (a sequence of points with decreasing distances) in it converges.

Representer Theorem

Theorem

Let $L : \mathcal{X} \times \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ *be an arbitrary loss function over* $\mathcal{X} = \{\boldsymbol{x}^{(t)}, r^{(t)}\}_{t=1}^N$, $C_k : \mathcal{X} \times \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$, $0 \leqslant i \leqslant K$, *constraint functions, and* $\Omega : [0, \infty) \to \mathbb{R}$ *a strictly increasing function. Then each minimizer* $h \in \mathcal{H}$ *of the regularized risk functional:*

$$\arg\min_{g \in \mathcal{H}} L((\boldsymbol{x}^{(1)}, r^{(1)}, g(\boldsymbol{x}^{(1)})), \cdots, (\boldsymbol{x}^{(N)}, r^{(N)}, g(\boldsymbol{x}^{(N)}))) + \Omega(\|g\|_{\mathcal{RKHS}})$$
$$\textit{subject to } C_k((\boldsymbol{x}^{(1)}, r^{(1)}, g(\boldsymbol{x}^{(1)})), \cdots, (\boldsymbol{x}^{(N)}, r^{(N)}, g(\boldsymbol{x}^{(N)}))) \leqslant 0, \forall k$$

任何這種form

admits the form $h(\boldsymbol{x}) = \sum_{t=1}^N c_t k(\boldsymbol{x}^{(t)}, \boldsymbol{x})$. 都可以換成這種form

- L is more general than l seen previously as the former allows coupling between samples
- $\Omega(\|g\|_{\mathcal{RKHS}})$ can be written as $\widetilde{\Omega}(\|g\|^2_{\mathcal{RKHS}})$ without loss of generality
  - As the quadratic function is strictly increasing on $[0, \infty)$, hence Ω is strictly increasing iff $\widetilde{\Omega}$ is so
  - In particular, $\widetilde{\Omega}(\|g\|^2_{\mathcal{RKHS}})$ can be $\lambda\|g\|^2_{\mathcal{RKHS}}$ for some $\lambda > 0$

Proof

任何向量都可以分別投影到兩個 orthogonal space, 表示成它們的和

- We consider $\widetilde{\Omega}(\|g\|^2_{\mathcal{RKHS}})$ for convenience
- By the fundamental theorem of linear algebra, we can decompose any $g \in \mathcal{H}$ into two vectors parallel and orthogonal to span of $k(\boldsymbol{x}^{(1)},\cdot),\cdots,k(\boldsymbol{x}^{(N)},\cdot)$ respectively; i.e., $g(\boldsymbol{x}) = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x}) + g_\perp(\boldsymbol{x})$ 反證:假設不能寫成那個form，即g⊥≠0
- Since $\langle g_\perp, k(\boldsymbol{x}^{(i)},\cdot)\rangle = 0$ for all $1 \leqslant i \leqslant N$, we have g⊥和任何k(x⁽ⁱ⁾,·)都垂直，因此內積為0 $g(\boldsymbol{x}^{(i)}) = \langle g, k(\boldsymbol{x}^{(i)},\cdot)\rangle = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x}^{(i)}) + \langle g_\perp, k(\boldsymbol{x}^{(i)},\cdot)\rangle = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x}^{(i)})$
- Now suppose the minimizer h has the form $h(\boldsymbol{x}) = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x}) + h_\perp(\boldsymbol{x})$, next we show that $h - h_\perp$ is always a better solution, which contradicts our assumption

$\boxed{h - h_\perp = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x})}$

一個更小的h也滿足，所以是更好的h，與假設矛盾

- First, $h - h_\perp$ satisfies all constrains C_k's, as $(h-h_\perp)(\boldsymbol{x}^{(i)}) = h(\boldsymbol{x}^{(i)})$ for all $1 \leqslant i \leqslant N$ 垂直的部分，內積=0
- Due to the same reason, $h - h_\perp$ has the same loss score from L as h
- Furthermore, $\widetilde{\Omega}(\|h-h_\perp\|^2) = \widetilde{\Omega}(\|\sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x})\|^2_{\mathcal{RKHS}}) \leqslant \widetilde{\Omega}(\|\sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)},\boldsymbol{x})\|^2_{\mathcal{RKHS}} + \|h_\perp\|^2_{\mathcal{RKHS}}) = \widetilde{\Omega}(\|h\|^2_{\mathcal{RKHS}})$

Kernel Machines

- The minimizers of the problems with the form

$$\arg\min_{g\in\mathcal{H}} L((\boldsymbol{x}^{(1)},r^{(1)},g(\boldsymbol{x}^{(1)})),\cdots,(\boldsymbol{x}^{(N)},r^{(N)},g(\boldsymbol{x}^{(N)})))+\Omega(\|g\|_{\mathcal{RKHS}})$$
$$\text{subject to } C_k((\boldsymbol{x}^{(1)},r^{(1)},g(\boldsymbol{x}^{(1)})),\cdots,(\boldsymbol{x}^{(N)},r^{(N)},g(\boldsymbol{x}^{(N)}))) \leqslant 0,\ \forall k$$

  are called ***kernel machines***
- The representer theorem tells us that although the RKHS is infinite dimensional, the solution will always be in a subspace spanned by $k(\boldsymbol{x}^{(1)},\cdot),\cdots,k(\boldsymbol{x}^{(N)},\cdot)$
  - We don't need to search for the entire RKHS to obtain a solution
- For ***any kernel***, we only need to solve N variables, ***independent*** with the dimension of feature space

Example: Regularized Nonlinear Regression

- Let $\Phi(\boldsymbol{x}) = k(\boldsymbol{x}, \cdot)$, we can write the objective $(b = 0)$ as

$$\arg \min_{g : g(\Phi(\boldsymbol{x})) = \boldsymbol{w}^\top \Phi(\boldsymbol{x})} \sum_{t=1}^{N} (r^{(t)} - g(\Phi(\boldsymbol{x}^{(t)})))^2 + \lambda \|g\|^2_{\mathcal{RKHS}}$$

- Then for any kernel, we can always solve an alternative objective:

$$\arg\min_{\boldsymbol{c}} \|\boldsymbol{r} - \boldsymbol{K}\boldsymbol{c}\|^2 + \lambda \boldsymbol{c}^\top \boldsymbol{K}\boldsymbol{c}$$

  - $\|g\|^2_{\mathcal{RKHS}} = \langle g, g \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} c^{(i)} c^{(j)} k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \sum_{i=1}^{N} \sum_{j=1}^{N} c^{(i)} \boldsymbol{x}^{(i)\top} \boldsymbol{x}^{(j)} c^{(j)} = \boldsymbol{c}^\top \boldsymbol{K}\boldsymbol{c}$
- $\boldsymbol{c} \in \mathbb{R}^N$, so the number of variables to solve is independent with the dimension of feature space
- In the linear case, we have $\Phi(\boldsymbol{x}) = k(\boldsymbol{x}, \cdot) = \boldsymbol{x}^\top [\cdot]$
  - The representer theorem coincides with our previous observation in linear regression that $\boldsymbol{w}^* = \sum_{t=1}^{N} c_t \boldsymbol{x}^{(t)}$
  - Same kernel function $k(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}^\top \boldsymbol{b}$, different feature spaces ($\Phi(\boldsymbol{x}) = \boldsymbol{x}$ vs. $\Phi(\boldsymbol{x}) = \boldsymbol{x}^\top [\cdot]$)

Semiparametric Representer Theorem

Theorem

Following the previous theorem, let $\widetilde{g} := g + b\psi$*, where* $g \in \mathcal{H}$*,* $b \in \mathbb{R}$*, and* $\psi : \mathcal{I} \to \mathbb{R}$*. Then each minimizer* $\widetilde{h}$ *of the regularized risk functional:*

$$\arg\min_{\widetilde{g}} L((\boldsymbol{x}^{(1)}, r^{(1)}, \widetilde{g}(\boldsymbol{x}^{(1)})), \cdots, (\boldsymbol{x}^{(N)}, r^{(N)}, \widetilde{g}(\boldsymbol{x}^{(N)}))) + \Omega(\|g\|_{\mathcal{RKHS}})$$
$$\text{subject to } C_k((\boldsymbol{x}^{(1)}, r^{(1)}, \widetilde{g}(\boldsymbol{x}^{(1)})), \cdots, (\boldsymbol{x}^{(N)}, r^{(N)}, \widetilde{g}(\boldsymbol{x}^{(N)}))) \leqslant 0, \forall k$$

admits the form $\widetilde{h}(\boldsymbol{x}) = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)}, \boldsymbol{x}) + b\psi(\boldsymbol{x})$*. [Proof]*

- When $\psi(\boldsymbol{x}) = 1$, we have $\widetilde{h}(\boldsymbol{x}) = \sum_{t=1}^{N} c_t k(\boldsymbol{x}^{(t)}, \boldsymbol{x}) + b$
- Applicable to the kernel machines with a unregularized bias term (i.e., $b \neq 0$)
  - What is the objective of regularized nonlinear regression with $b \neq 0$ after applying this theorem? [Homework]