Reproducing kernel Hilbert spaces and regularization

Nuno Vasconcelos

ECE Department, UCSD
Classification

- A classification problem has **two types of variables**
  - $X$ - vector of **observations** (features) in the world
  - $Y$ - **state** (class) of the world

**Perceptron**: classifier implements the **linear decision rule** with appropriate when the classes are linearly separable.

$$h(x) = \text{sgn}[g(x)] \quad \text{with} \quad g(x) = w^T x + b$$

- to deal with **non-linear separability**, we introduce a kernel
Types of kernels

these three are equivalent

- dot product kernel
- positive definite kernel
- Mercer kernel
Dot-product kernels

Definition: a mapping

\[ k: \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{H} \]
\[ (x,y) \rightarrow k(x,y) \]

is a dot-product kernel if and only if

\[ k(x,y) = \langle \Phi(x), \Phi(y) \rangle \]

where

\[ \Phi: \mathcal{X} \rightarrow \mathcal{H}, \]
\[ \mathcal{H} \text{ is a vector space}, \]
\[ \langle ., . \rangle \text{ is a dot-product in } \mathcal{H} \]
positive definite and Mercer kernels

**Definition:** $k(x,y)$ is a positive definite kernel on $X \times X$ if \( \forall l \) and \( \forall \{x_1, ..., x_l\}, x_i \in X \), the Gram matrix

$$[K]_{ij} = k(x_i, x_j)$$

is positive definite.

**Definition:** a symmetric mapping $k: X \times X \rightarrow \mathbb{R}$ such that

$$\int\int k(x, y)f(x)f(y)dx\,dy \geq 0,$$

$$\forall f(x) \text{ s.t. } \int f(x)^2 \, dx < \infty$$

is a Mercer kernel.
Two different pictures

- different definitions lead to different interpretations of what the kernel does

**Reproducing kernel map**

$$\mathcal{H}_K = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \right\}$$

$$\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m^l} \alpha_i \beta_j k(x_i, x'_j)$$

$$\Phi : x \rightarrow k(., x)$$

**Mercer kernel map**

$$\mathcal{H}_M = l_2^d = \left\{ x \mid \sum_{i=1}^{d} x_i^2 < \infty \right\}$$

$$\langle f, g \rangle_* = f^T g$$

$$\Phi : x \rightarrow \left( \sqrt{\lambda_1} \phi_1(x), \sqrt{\lambda_2} \phi_2(x), \ldots \right)^T$$

where $\lambda_i, \phi_i$ are the eigenvalues and eigenfunctions of $k(x,y)$

$$\lambda_i > 0$$
The dot-product picture

when we use the Gaussian kernel $K(x, x_i) = e^{-\frac{\|x-x_i\|^2}{\sigma^2}}$

- the point $x_i \in \mathcal{X}$ is mapped into the Gaussian $G(x, x_i, \sigma)$
- $\mathcal{H}$ is the space of all functions that are linear combinations of Gaussians
- the kernel is a dot product in $\mathcal{H}$, and a non-linear similarity on $\mathcal{X}$
- reproducing property on $\mathcal{H}$: analogy to linear systems
The Mercer picture

▸ this is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^d$

▸ much more like to a multi-layer Perceptron than before

▸ the kernelized Perceptron as a neural net
The reproducing property

with this definition of $\mathcal{H}_K$ and $\langle \cdot, \cdot \rangle_*$

$$\forall f \in \mathcal{H}_K, \quad \langle k(\cdot, x), f(\cdot) \rangle_* = f(x)$$

this is called the reproducing property

an analogy is to think of linear time-invariant systems

• the dot product as a convolution
• $k(\cdot, x)$ as the Dirac delta
• $f(\cdot)$ as a system input
• the equation above is the basis of all linear time invariant systems theory

leads to reproducing Kernel Hilbert Spaces
Definition: a Hilbert space is a complete dot-product space (vector space + dot product + limit points of all Cauchy sequences)

Definition: Let $\mathcal{H}$ be a Hilbert space of functions $f: \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is a RKHS endowed with dot-product $\langle .,. \rangle_*$ if

$\exists k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

1. $k$ spans $\mathcal{H}$, i.e., $\exists \{x_i\}, \{\alpha_i\}$, such that

$\mathcal{H} = \text{span}\{k(., x_i)\} = \left\{ f(\cdot) \mid f(\cdot) = \sum \alpha_i k(\cdot, x) \right\}

2. $\langle f(.), k(., x) \rangle_* = f(x), \ \forall f \in \mathcal{H}$,
Mercer kernels

- how different are the spaces $\mathcal{H}_K$ and $\mathcal{H}_M$?

**Theorem:** Let $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a Mercer kernel. Then, there exists an orthonormal set of functions

$$\int \phi_i(x)\phi_j(x)dx = \delta_{ij}$$

and a set of $\lambda_i \geq 0$, such that

1) $\sum_{i=1}^{\infty} \lambda_i^2 = \int \int k^2(x,y)dxdy$

2) $k(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$ (***)
RK vs Mercer maps

Note that for $\mathcal{H}_M$ we are writing

$$\Phi(x) = \sqrt{\lambda_1} \phi_1(x) \tilde{e}_1 + \cdots + \sqrt{\lambda_{d}} \phi_{d}(x) \tilde{e}_{d}$$

but, since the $\phi_i(.)$ are orthonormal, there is a 1-1 map

$$\Gamma : l_2^d \rightarrow \text{span}\{\phi_k(.)\}$$

and we can write

$$\left(\Gamma \circ \Phi\right)(x) = \lambda_1 \phi_1(x) \phi_1(.) + \cdots + \lambda_{d} \phi_{d}(x) \phi_{d}(.)$$

$$= k(., x)$$

from (**)

Hence $k(., x)$ maps $x$ into $\mathcal{M} = \text{span}\{\phi_k(.)\}$
The Mercer picture

we have

\[ \Phi(x_i) \]

\[ T_0 \Phi(x_i) = k(\cdot, x_i) \]

\[ \mathcal{M} \]

\[ \Gamma \]

\[ \phi_1 \]

\[ \phi_2 \]

\[ \phi_3 \]

\[ e_1 \]

\[ e_2 \]

\[ e_3 \]

\[ e_d \]
 RK vs Mercer maps

- define the dot product in $\mathcal{M}$ so that

  $$
  \langle \phi_j, \phi_k \rangle^{**} = \frac{1}{\lambda_j} \int \phi_j(x) \phi_k(x) \, dx = \frac{1}{\lambda_j} \delta_{ij}
  $$

- then $\{\phi_k(.)\}$ is a basis, $\mathcal{M}$ is a vector space, any function in $\mathcal{M}$ can be written as

  $$
  f(x) = \sum_k \alpha_k \phi_k(x)
  $$

  and

  $$
  \langle f(\cdot), k(\cdot, x) \rangle^{**} = \left\langle \sum_i \alpha_i \phi_i(\cdot), \sum_j \lambda_j \phi_j(\cdot) \phi_j(x) \right\rangle^{**}
  $$

  $$
  = \sum_{ij} \alpha_i \lambda_j \phi_j(x) \langle \phi_i, \phi_j \rangle^{**} = \sum_i \alpha_i \phi_i(x) = f(x)
  $$

  i.e., $k$ is a reproducing kernel on $\mathcal{M}$
RK vs Mercer maps

Furthermore, since \( k(.,x) \in \mathcal{M} \), any functions of the form

\[
f(.) = \sum_i \alpha_i k(., x_i) \quad g(.) = \sum_j \beta_j k(., x_j)
\]

are in \( \mathcal{M} \) and

\[
\langle f, g \rangle^{**} = \left\langle \sum_i \alpha_i k(., x_i), \sum_j \beta_j k(., x_j) \right\rangle^{**}
\]

\[
= \sum_{ij} \alpha_i \beta_j \left\langle \sum_l \lambda_l \phi_l(.) \phi_l(x_i), \sum_m \lambda_m \phi_m(.) \phi_m(x_j) \right\rangle^{**}
\]

\[
= \sum_{ij} \alpha_i \beta_j \sum_{lm} \lambda_l \lambda_m \phi_l(x_i) \phi_m(x_j) \langle \phi_l(.), \phi_m(.) \rangle^{**}
\]

\[
= \sum_{ij} \alpha_i \beta_j \sum_l \lambda_l \phi_l(x_i) \phi_l(x_j) = \sum_{ij} \alpha_i \beta_j k(x_i, x_j)
\]

Note that \( f,g \in \mathcal{H}_K \) and this is the dot product we had in \( \mathcal{H}_K \).
The Mercer picture

- furthermore, note that for \( f \) in \( \mathcal{H}_M \),

\[
f(x) = \sum_k \alpha_k \phi_k(x)
\]

- and since

\[
\langle \phi_j, \phi_k \rangle_{**} = \frac{1}{\lambda_j} \int \phi_j(x) \phi_k(x) \, dx = \frac{1}{\lambda_j} \delta_{ij}
\]

- the dot product on \( \mathcal{H}_M \) is

\[
\langle f, g \rangle_{**} = \sum_{kl} \alpha_k \beta_l \langle \phi_k, \phi_l \rangle_{**}
= \sum_k \frac{\alpha_k \beta_k}{\lambda_k}
\]
The Mercer picture

we have

\[ \langle f, g \rangle_{**} = \sum_{ij} \alpha_i \beta_j k(x_i, x_j) \]

\[ \langle f, g \rangle_{**} = \sum_{k} \frac{\alpha_k \beta_k}{\lambda_k} \]

\[ T_0 \Phi(x_i) = k(., x_i) \]
RK vs Mercer maps

- $\mathcal{H}_K \subset \mathcal{M}$ and $<.,.>*$ in $\mathcal{H}_K$ is the same as $<.,.>**$ in $\mathcal{M}$

Question: is $\mathcal{M} \subset \mathcal{H}_K$?

- need to show that any $f(x) = \sum_k \alpha_k \phi_k(x) \in \mathcal{H}_K$
- from (***)

$$k(., x) = \lambda_1 \phi_1(x) \phi(.) + \cdots + \lambda_d \phi_d(x) \phi_d(.)$$

and, for any sequence $\{x_1, ..., x_d\}$,

$$\begin{bmatrix}
  k(., x_1) \\
  \vdots \\
  k(., x_d)
\end{bmatrix} = \begin{bmatrix}
  \lambda_1 \phi_1(x_1) & \lambda_d \phi_d(x_1) \\
  \vdots & \vdots \\
  \lambda_1 \phi_1(x_d) & \lambda_d \phi_d(x_d)
\end{bmatrix} \begin{bmatrix}
  \phi_1(.) \\
  \vdots \\
  \phi_d(.)
\end{bmatrix}$$

- if there is an invertible $P$, then $\phi_k(x) = \sum_k \alpha_k k(.,x_k) \in \mathcal{H}_K$ and $\mathcal{M} \subset \mathcal{H}_K$
RK vs Mercer maps

since $\lambda_i > 0$

$$P = \begin{bmatrix}
\phi_1(x_1) & \phi_d(x_1) \\
\vdots & \vdots \\
\phi_1(x_d) & \phi_d(x_d)
\end{bmatrix}
\begin{bmatrix}
\lambda_1 & 0 \\
\vdots & \lambda_i \\
0 & \lambda_d
\end{bmatrix}$$

is invertible when $\Pi$ is. If $\Pi$ is not invertible, then

$$\forall i \exists \alpha \neq 0 \text{ s.t. } \sum_k \alpha_k \phi_k(x_i) = 0$$

if there is no sequence for which $\Pi$ is invertible then

$$\forall x \exists \alpha \neq 0 \text{ s.t. } \sum_k \alpha_k \phi_k(x) = 0$$

the $\phi_k(x)$ cannot be orthonormal. Hence there must be invertible $P$ and $\mathcal{M} \subset \mathcal{H}_K$. ■
The Mercer picture

we have

\[ \langle f, g \rangle_{**} = \sum_{ij} \alpha_i \beta_j k(x_i, x_j) \]

\[ \langle f, g \rangle_{**} = \sum_k \frac{\alpha_k \beta_k}{\lambda_k} \]
In summary

- $\mathcal{H}_K = M$ and $<.,.>^*$ in $\mathcal{H}_K$ is the same as $<.,.>^{**}$ in $M$
- the reproducing kernel map and the Mercer kernel map lead to the same RKHS, Mercer gives us an o.n. basis

- Reproducing kernel map
  $$\mathcal{H}_K = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \right\}$$
  $$\langle f, g \rangle^* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)$$
  $$\Phi_r : x \rightarrow k(., x)$$

- Mercer kernel map
  $$\mathcal{H}_M = l_2^d = \left\{ x \mid \sum_{i=1}^{d} x_i^2 < \infty \right\}$$
  $$\langle f, g \rangle^* = f^T g$$
  $$\Phi_M : x \rightarrow \left( \sqrt{\lambda_1} \phi_1(x), \sqrt{\lambda_2} \phi_2(x), \cdots \right)^T$$

- $\Gamma \circ \Phi_M = \Phi_r$

- this turns out to be true for any RKHS obtained from $k(x,y)$ (1-1 relationship)
Q: RKHS: why do we care?

A: regularization

- we want to do well outside of the training set
- minimizing empirical risk is not enough
- need to penalize complexity
- example: regression
  - give me n points, I will give you a model of zero error (polynomial of order n-1)
  - incredibly "wiggly"
  - poor generalization
Regularization

we need to “regularize” the risk

\[ R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda \Omega[f] \]

- \( f \) is the function we are trying to learn
- \( \lambda > 0 \) is the regularization parameter
- \( \Omega \) is a complexity penalizing functional
  - larger when \( f \) is more “wiggly”

the regularized risk can be justified in various ways

- constrained minimization
- Bayesian inference
- structural risk minimization
- ...

23
Constrained minimization

- regularized risk as the solution to the problem

\[ f^* = \arg\min_f R_{\text{emp}}[f] \quad \text{subject to} \quad \Omega[f] \leq t \]

- we will talk a lot more about this
  - solution is the solution of the Lagrangian problem
    \[ f^* = \arg\min_f \{ R_{\text{emp}}[f] + \lambda \Omega[f] \} \]
  - \( \lambda \) is a Lagrange multiplier
  - changing \( \lambda \) is equivalent to changing \( t \) (the constraint on the complexity of the optimal solution)
Bayesian inference

In Bayesian inference we have

- likelihood function $P_{X|\theta}(x|\theta)$
- prior $P_{\theta}(\theta)$
- and search for the maximum a posteriori (MAP) estimate

$$\theta^* = \arg \max_{\theta} P_{\theta|X}(\theta | x)$$

$$= \arg \max_{\theta} \frac{P_{X|\theta}(x | \theta)P_{\theta}(\theta)}{P_X(x)}$$

$$= \arg \max_{\theta} \left\{ P_{X|\theta}(x | \theta)P_{\theta}(\theta) \right\}$$

$$= \arg \max_{\theta} \left\{ \log[P_{X|\theta}(x | \theta)] + \log P_{\theta}(\theta) \right\}$$
Bayesian inference

if \( P_{X|\theta}(x|f) = e^{-R_{emp}[f]} \) and \( P_{\theta}(f) = e^{-\lambda \Omega[f]} \)

- the MAP estimate is the minimum of the regularized risk

\[
    f^* = \arg\min_{f} \{ R_{emp}[f] + \lambda \Omega[f] \}
\]

- the prior \( P_{\theta}(f) \) assigns low probability to function \( f \) with large \( \Omega[f] \)
- this reflects our prior belief that it is “unlikely” that the solution will be very complex

example:

- if \( \Omega[f] = || f-f_0 ||^2 \), the prior is a Gaussian centered at \( f_0 \) with variance \( 1/\lambda^{1/2} \)
- we believe that the most likely solution is \( f = f_0 \)
- the larger the \( \lambda \), the more we penalize solutions which are different from this
Structural risk minimization

• start from a nested collection of families of functions

\[ S_1 \subset \cdots \subset S_k \]

where \( S_i = \{ h_i(x, \alpha), \text{ for all } \alpha \} \)

• for each \( S_i \), find the function (set of parameters) that minimizes the empirical risk

\[ R_{emp}^i = \min_{\alpha} \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}[y_k, h_i(x_k, \alpha)] \]

• select the function class such that

\[ R^* = \min_{i} \left\{ R_{emp}^i + \Phi(h_i) \right\} \]

• \( \Phi(h) \) is a function of VC dimension (complexity) of the family \( S_i \)

• VC have shown that this is equivalent to minimizing a bound on the risk, and provides generalization guarantees

• regularization with the “right” regularizer!
The regularizer

what is a good regularizer?

intuition: “wigglier” functions have a larger norm than smoother functions

for example, in $\mathcal{H}_K$ we have

$$f(x) = \sum_i \alpha_i k(., x_i)$$

$$= \sum_i \alpha_i \sum_l \lambda_l \phi_l(x) \phi_l(x_i) =$$

$$= \sum_l \left[ \lambda_l \sum_i \alpha_i \phi_i(x_i) \right] \phi_l(x)$$

$$= \sum_l c_l \phi_l(x)$$
The regularizer

and

$$\| f(x) \|^2 = \sum_{lk} c_l c_k \langle \phi_l(x), \phi_k(x) \rangle_{**} = \sum_{lk} c_l c_k \frac{\delta_{lk}}{\lambda_l} = \sum_l \frac{c_l^2}{\lambda_l}$$

with $c_l = \lambda_l \sum_i \alpha_i \phi_l(x_i)$

hence,

- $\| f \|^2$ grows with the number of $c_l$ different than zero
- this is the case in which $f$ is more complex, since it becomes a sum of more basis function $\phi_l(x)$
- identical to what happens in Fourier type decompositions
- more coefficients means more “high-frequencies” or “less smoothness”
Regularization

- OK, regularization is a good idea from multiple points of view
- problem: minimizing the regularized risk

\[ R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda \Omega[f] \]

over the set of all functions seems like a nightmare

- it turns out that it is not, under some reasonable conditions on the regularizer \( \Omega \)
- the key is the “representer theorem”
Representer theorem

Theorem: Let

- $\Omega: [0, \infty) \rightarrow \mathcal{H}$ be a strictly monotonically increasing function,
- $\mathcal{H}$ the RKHS associated with a kernel $k(x,y)$
- $L[y,f(x)]$ a loss function

then, if

$$f^* = \arg \min_f \left[ \sum_{i=1}^n L[y_i, f(x_i)] + \lambda \Omega(\|f\|^2) \right]$$

$f^*$ admits a representation of the form

$$f^* = \sum_{i=1}^n \alpha_i k(., x_i) \quad (i.e. \ f^* \in \mathcal{H})$$
Proof

- decompose any $f$ into the part contained in the span of the $k(.,x_i)$ and the part in the orthogonal complement

$$f(x) = f_0(x) + f_\perp(x) = \sum_{i=1}^{n} \alpha_i k(., x_i) + f_\perp(x)$$

where $f \in w_0$ and $f \in w_\perp$ with

$$w_0 = \{ f \mid f \in \text{span}\{k(., x_i)\}\}$$

$$w_\perp = \{ g \mid \langle g, f \rangle = 0, \forall f \in w_0 \}$$

- then

$$\Omega\left[\|f\|^2\right] = \Omega\left(\left\| \sum_{i=1}^{n} \alpha_i k(., x_i) + f_\perp(x) \right\|^2\right)$$

$$(\Omega \text{ mon. increasing})$$

$$= \Omega\left(\left\| \sum_{i=1}^{n} \alpha_i k(., x_i) \right\|^2 + \|f_\perp\|^2\right) \geq \Omega\left(\left\| \sum_{i=1}^{n} \alpha_i k(., x_i) \right\|^2\right)$$
Proof

• this shows that the second term in
  \[ f^* = \arg \min_f \left[ \sum_{i=1}^{n} L[y_i, f(x_i)] + \lambda \Omega(\|f\|^2) \right] \]
  is minimized by a function of the stated form.

• the first, using the reproducing property

  \[ f(x_i) = \langle f(\cdot), k(\cdot, x_i) \rangle = \langle f_0(\cdot), k(\cdot, x_i) \rangle + \langle f_\perp(\cdot), k(\cdot, x_i) \rangle \]
  \[ = \sum_{j=1}^{n} \alpha_j k(x_i, x_j) \]
  is always a function of the stated form.

• hence, the minimum must be of this form as well.
The picture

- due to reproducing property:
  - $f(x_i) \in \mathcal{H}$

- if $f^*$ is the solution in $\mathcal{H}$
  - $\|f\| > \|f^*\|$
  - is a more complex function
  - which does not reduce the loss

- hence, $f^*$ is the optimal solution
Relevance

the remarkable consequence of the theorem is that:

• we can reduce the minimization over the (infinite dimensional) space of functions

• to a minimization on a finite dimensional space!

to see this note that, because \( f^* = \sum_{i=1}^{n} \alpha_i k(., x_i) \), we have

\[
\left\| f^* \right\|^2 = \left\langle f^*, f^* \right\rangle = \sum_{ij} \alpha_i \alpha_j \left\langle k(., x_i), k(., x_j) \right\rangle
\]

\[
= \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha
\]

\[
f^*(x_i) = \sum_{j} \alpha_j k(x_i, x_j) = K \alpha
\]
Regularization

this proves the following theorem

Theorem:

• if \( \Omega: [0, \infty) \rightarrow \mathcal{R} \) is a strictly monotonically increasing function,
• then for any dot-product kernel \( k(x,y) \) and loss function \( L[y,f(x)] \)

the solution of

\[
\mathbf{f}^* = \arg\min_{\mathbf{f}} \left[ \sum_{i=1}^{n} L[y_i, f(x_i)] + \lambda \Omega(\|\mathbf{f}\|) \right]
\]

is

\[
\mathbf{f}^* = \sum_{i=1}^{n} \alpha_i^* k(\cdot, x_i)
\]

with

\[
\alpha^* = \arg\min_{\alpha} \left[ \sum_{i=1}^{n} L[Y, K\alpha] + \lambda \Omega(\alpha^T K\alpha) \right]
\]

K the Gram matrix \([k(x_i, x_j)]\) and \( Y = (...) , y_i, (...) \)
Note

- the result that $f^* \in \mathcal{H}$ holds for any norm, not just $L_2$

- however, the exact form of the problem will no longer be

\[
\alpha^* = \arg \min_{\alpha} \left[ \sum_{i=1}^{n} L[Y_i, K\alpha] + \lambda \Omega(\alpha^T K\alpha) \right]
\]

- the argument of the second term will have a different form
Any Questions?