CS:E4830 Kernel Methods in Machine Learning

Lecture 3 : RKHS and Representer Theorem

Rohit Babbar

23rd January, 2019

Couple of Announcements

- Lecture slides of this (3rd) lecture is in Mycourses
- Exercise garage with Eric Bach, tomorrow at 4:15 for help with programming part of assignment 1

Recap of Previous Lecture

What is a Kernel - Definition

Definition

For a non-empty set $\mathcal{X}$, a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is defined to be a kernel if there exists a Hilbert Space $\mathcal{H}$ and a function $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$, $k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j)$$

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}}$$

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i\phi(x_i), a_j\phi(x_j)\rangle_{\mathcal{H}} = || \sum_{i=1}^{n} a_i\phi(x_i)||_{\mathcal{H}}^2 \geq 0$$

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j)\rangle_{\mathcal{H}} = || \sum_{i=1}^{n} a_i \phi(x_i) ||_{\mathcal{H}}^{2} \geq 0$$

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Proof outline (Proof is not required for the exam)

Though, we just need to prove the existence of the feature (Hilbert) space and feature map. However, in this case, it is a proof by construction, meaning

- We construct **the** feature map and **the** feature space

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Proof outline (Proof is not required for the exam)

Though, we just need to prove the existence of the feature (Hilbert) space and feature map. However, in this case, it is a proof by construction, meaning

- We construct **the** feature map and **the** feature space
- **Key aspect** : The feature (Hilbert) space will be a space of functions, i.e., the points in the space are, in fact, **functions**. How does that look like ?

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Proof outline (Proof is not required for the exam)

Though, we just need to prove the existence of the feature (Hilbert) space and feature map. However, in this case, it is a proof by construction, meaning

- We construct **the** feature map and **the** feature space
- **Key aspect** : The feature (Hilbert) space will be a space of functions, i.e., the points in the space are, in fact, **functions**. How does that look like ?
- A possible candidate form of the function space is given by the set

$$\mathcal{H} = \left\{ \sum_{i=1}^{\ell} \alpha_i k(x_i, .) : \ell \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, \ldots, \ell \right\}$$

Function in the candidate space of functions

How do typical functions $f(.)$ and $g(.)$ in $\mathcal{H}$ look like?

Function in the candidate space of functions

How do typical functions $f(.)$ and $g(.)$ in $\mathcal{H}$ look like?

- Consider the following :
  1. $f(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .)$
  2. $g(.) = \sum_{j=1}^{m} \beta_j k(y_j, .)$

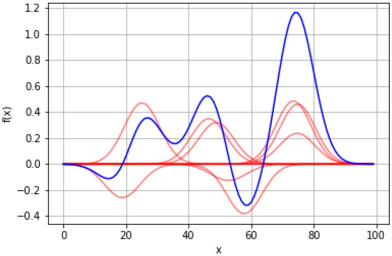


Figure: Pictorial depiction of a function $f(.)$ (in blue) as a linear combination of kernel functions $k(x_i, .)$ for Gaussian kernel evaluated at 9 points (Figure by Eric Bach)

Defining the Feature (function) space

Proof (1/3) - Defining the function Space

- **Key aspect** : The feature (Hilbert) space will be a space of functions, i.e., the points in the space are, in fact, **functions**.
- A candidate for the function space is given by the set of function given by $\mathcal{H}$

$$\mathcal{H} = \left\{ \sum_{i=1}^{\ell} \alpha_i k(x_i, .) : \ell \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, \ldots, \ell \right\}$$

Key points :

- These are non-linear functions in the input space (as shown in picture in last slide)
- Linear functions in the (possibly infinite dimensional) feature space
  - That is, these are of the form $f(x) = f_1\phi_1(x) + f_2\phi_2(x) + \ldots$ (more on this part later)

Defining Inner product on the space

Proof (2/3) - Verifying elementary properties and defining inner product

- It is a function (vector) space which satisfies the requirements of closure under scalar multiplication and addition
  - For $f \in \mathcal{H}, \gamma \in \mathbb{R} \implies \gamma f \in \mathcal{H}$
  - $f, g \in \mathcal{H} \implies (f + g) \in \mathcal{H}$
- Define Inner product on $\mathcal{H}$ as follows :
  - Let $f, g \in \mathcal{H}$ be given by

$$f(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .) \text{ and } g(.) = \sum_{j=1}^{m} \beta_j k(y_j, .)$$

  - The inner product between f and g is defined by the following

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(y_j) \tag{1}$$

which satisfies the symmetry and linearity properties of the inner product

Reproducing property

Proof (3/3) Using positive definiteness for positivity of IP with itself

- Now, we need $\langle f, f \rangle_{\mathcal{H}} \geq 0, \forall f \in \mathcal{H}$
    - This follows from the positive definiteness of the given function

$$\langle f, f \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

- An important property that follows from equation (1) is obtained by taking $g = k(x, .)$,

$$\langle f, k(x, .) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x)$$

- The above is called the **reproducing property** of the kernel $k(., .)$.
- Also, $f(x)$ has the form $\langle f, k(x, .) \rangle_{\mathcal{H}}$, i.e, the evaluation of f at x is in the form of an inner product (linear function) between f and a feature map $\phi(x) = k(x, .)$, which is called the **canonical feature map** for kernel $k(., .)$

Reproducing kernel Hilbert Space

RKHS - Definition I

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathcal{R}^{\mathcal{X}})$ is defined to be an **Reproducing kernel Hilbert Space (RKHS)** (with $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as the reproducing kernel), if the following conditions are satisfied

- $\forall x \in \mathcal{X}, k(., x) \in \mathcal{H}$ i.e., the space $\mathcal{H}$ contains all functions of the form $k(., x)$ for every element x in the input space $\mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, the following property holds :$f(x) = \langle f, k(., x)\rangle_{\mathcal{H}}$ (it is called the reproducing property of the kernel).

RKHS - Definition I

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathcal{R}^{\mathcal{X}})$ is defined to be an **Reproducing kernel Hilbert Space (RKHS)** (with $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as the reproducing kernel), if the following conditions are satisfied

- $\forall x \in \mathcal{X}, k(., x) \in \mathcal{H}$ i.e., the space $\mathcal{H}$ contains all functions of the form $k(., x)$ for every element x in the input space $\mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, the following property holds :$f(x) = \langle f, k(., x)\rangle_{\mathcal{H}}$ (it is called the reproducing property of the kernel).

The reproducing kernel $k(., .)$ of the Hilbert space $\mathcal{H}$ is a kernel

Proof.

$k(x, x') = \langle k(., x), k(., x')\rangle_{\mathcal{H}}$, from the reproducing property □

RKHS - Definition II

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathcal{R}^{\mathcal{X}})$ is defined to be an **Reproducing kernel Hilbert Space (RKHS)** if and only if, for any element x in the input space $\mathcal{X}$, the following function F, which takes a function f from the Hilbert Space $\mathcal{H}$, and maps it to its value $f(x) \in \mathbb{R}$

$$\begin{aligned} F: \quad \mathcal{H} &\mapsto \quad \mathbb{R} \\ f &\mapsto \quad f(x) \end{aligned}$$

is continuous

Why RKHS are nice function spaces?

Two functions which are identical in the RKHS norm agree at every point :

- This follows from

$$|f(x) - g(x)| \leq \lambda_x ||f - g||_{\mathcal{H}}, \forall f, g \in \mathcal{H}$$

- It turns out that for Reproducing Kernel Hilbert Spaces, λ_x is bounded (not very large), and if $||f - g||_{\mathcal{H}} = 0$, then $f(x) = g(x), \forall x \in \mathcal{X}$
- **Crucially** : This property enables generalization i.e., those functions found from training data are also good on test data
- Furthermore, when we are allow our search (via learning algorithm) to find functions over arbitrary function classes (which may not be RKHS), then closeness in norm **does not** imply identical pointwise evaluations.

RKHS norm controls smoothness

RKHS norm and smoothness

$$\begin{aligned}|f(x)-f(x')| &= |\langle f, k(x,.)\rangle - \langle f, k(x',.)\rangle| \quad \text{(reproducing property applied to } f\text{)}\\ &= |\langle f, k(x,.) - k(x',.)\rangle| \quad \text{(linearity of dot product)}\\ &\leq ||k(.,x) - k(.,x')||_{\mathcal{H}} ||f||_{\mathcal{H}} \quad \text{(by Cauchy-Schwarz inequality)}\end{aligned}$$

- $||f||_{\mathcal{H}}$ controls how much the values at two points x and x' differ compared to their distance
- Larger value of $||f||_{\mathcal{H}}$ allows higher variations (potentially non-smooth functions)

Smaller RKHS norm $\implies$ Smooth functions

- The same happens in finite diemsions when we add regularization $||w||^2$ for linear regression and SVM

What if we are not in RKHS?

- If we are not in RKHS (meaning we are not using the RKHS norm to measure similarity or dis-similarity), then norm convergence does not imply pointwise convergence.
- Let $\mathcal{F} = L_2([0,1])$, i.e. it represents class of functions for which $\int_0^1 |f(x)^2 dx| < \infty$.
- Suppose, we measure distance between functions in the following way :

$$||f_1 - f_2||_{L_2([0,1])} = \left(\int_0^1 |f_1(x) - f_2(x)|^2 dx \right)^{1/2}$$

- Under this measure of distance between two functions, a function which is zero for all inputs, and one which is non-zero at finitely many points has distance 0.

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- Typically, in a machine learning setup, we do not have access to the true underlying data distribution, and instead we have access to a fixed training set $(x_i, y_i)_{i=1}^n$
- Assume that the data lies in $[0, 1]$, i.e., $x \in \mathcal{X} = [0, 1]$
  - Input x is chosen uniformly at random on $\mathcal{X}$,
  - the label y is chosen in a deterministic way as follows :

$$y = \begin{cases} -1 & \text{if } x < 0.5 \\ +1 & \text{otherwise} \end{cases}$$

- Consider, a potential classifier based on n training samples given as follows :

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1 \ldots n \\ +1 & \text{otherwise} \end{cases}$$

- What is it error on the training set?

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- Typically, in a machine learning setup, we do not have access to the true underlying data distribution, and instead we have access to a fixed training set $(x_i, y_i)_{i=1}^n$
- Assume that the data lies in $[0, 1]$, i.e., $x \in \mathcal{X} = [0, 1]$
  - Input x is chosen uniformly at random on $\mathcal{X}$,
  - the label y is chosen in a deterministic way as follows :

$$y = \begin{cases} -1 & \text{if } x < 0.5 \\ +1 & \text{otherwise} \end{cases}$$

- Consider, a potential classifier based on n training samples given as follows :

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1 \ldots n \\ +1 & \text{otherwise} \end{cases}$$

- What is it error on the training set?
  - training error $= 0$

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- Typically, in a machine learning setup, we do not have access to the true underlying data distribution, and instead we have access to a fixed training set $(x_i, y_i)_{i=1}^n$
- Assume that the data lies in $[0,1]$, i.e., $x \in \mathcal{X} = [0,1]$
    - Input x is chosen uniformly at random on $\mathcal{X}$,
    - the label y is chosen in a deterministic way as follows :

$$y = \begin{cases} -1 & \text{if } x < 0.5 \\ +1 & \text{otherwise} \end{cases}$$

- Consider, a potential classifier based on n training samples given as follows :

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1 \dots n \\ +1 & \text{otherwise} \end{cases}$$

- What is it error on the training set?
    - training error $= 0$
    - Has it learnt anything?

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- What is it error on the training set?
  - training error $= 0$ (minimum possible)
  - Has it learnt anything?
- What is its test error?

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- What is it error on the training set?
  - training error = 0 (minimum possible)
  - Has it learnt anything?
- What is its test error?
- What is this an example of ?

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- What is it error on the training set?
  - training error $= 0$ (minimum possible)
  - Has it learnt anything?
- What is its test error?
- What is this an example of ?
- Why does overfitting happen?

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- What is it error on the training set?
    - training error $= 0$ (minimum possible)
    - Has it learnt anything?
- What is its test error?
- What is this an example of ?
- Why does overfitting happen?
    - Because we allow any function (could be highly non-smooth) in our function space
- In order to generalize, we need to **restrict our function class**
    - Controlling the RKHS norm of the function $||f||_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ does exactly that
    - Also, called regularization

Bias-Variance Tradeoff

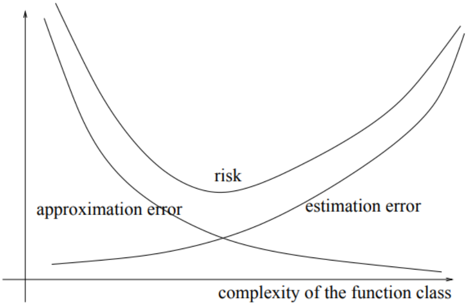


Figure: Bias-Variance Tradeoff (picture from the Learning theory tutorial by Von Luxburg and Schoelkopf, more in next lecture)

RKHS - Equivalence of Two Definitions I

Reproducing property $\implies$ Evaluation functionals are continuous

Proof.

$$\begin{aligned} |f(x)| &= |\langle f, k(.,x)\rangle_{\mathcal{H}}| \quad \text{(reproducing property applied to } f\text{)} \\ &\leq ||k(.,x)||_{\mathcal{H}}||f||_{\mathcal{H}} \quad \text{(by Cauchy-Schwarz inequality)} \\ &= \langle\sqrt{k(.,x)}, \sqrt{k(.,x)}\rangle||f||_{\mathcal{H}} \quad \text{(by definition of norm)} \\ &= \sqrt{k(x,x)} \times ||f||_{\mathcal{H}} \quad \text{(reproducing property applied to } k(.,.)\text{)} \end{aligned} \tag{2}$$

□

Therefore, the mapping $f \in \mathcal{H} \mapsto f(x) \in \mathbb{R}$ δ_x is continuous, i.e. $f \to 0 \implies f(x) \to 0$

RKHS - Equivalence of Two Definitions II

Bounded evaluation functionals $\implies$ Reproducing property

- We will skip this part of the proof, which uses the following theorem from functional analysis

Riesz representation Theorem

In a Hilbert space $\mathcal{H}$, all bounded linear functionals are of the form $\langle ., f\rangle_{\mathcal{H}}$, for $f \in \mathcal{H}$.

Uniqueness of RKHS and reproducing kernel

- Recall the following example from lecture 1 - For the linear kernel, with many possible Hilbert spaces $\mathcal{H}$, and feature maps such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$
  - For the linear kernel $k(x, x') := \langle x, x' \rangle$, two of the many possible choices of feature maps and Hilbert space are :
    - $\phi_1(x) = x$, and $\mathcal{H}_1 = \mathbb{R}$

Uniqueness of RKHS and reproducing kernel

- Recall the following example from lecture 1 - For the linear kernel, with many possible Hilbert spaces $\mathcal{H}$, and feature maps such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$
  - For the linear kernel $k(x, x') := \langle x, x' \rangle$, two of the many possible choices of feature maps and Hilbert space are :
    - $\phi_1(x) = x$, and $\mathcal{H}_1 = \mathbb{R}$
    - $\phi_2(x) = \frac{1}{\sqrt{2}}(x, x)$, and $\mathcal{H}_2 = \mathbb{R}^2$
- However, if $k(.,.)$ is a reproducing kernel for some Hilbert space, then it is **the only kernel** to exhibit the reproducing property.
- Also, $k(.,.)$ can be a reproducing kernel for a **unique Hilbert space**, which is its RKHS. That is, there cannot be multiple Hilbert spaces for which $k(.,.)$ is a reproducing kernel.

Therefore, for kernels, we can talk about **the** Reproducing Kernel Hilbert Space (like the RKHS of Gaussian kernel, we study next)

Reproducing Kernel and its functions

Properties of functions in the RKHS induced by the kernel depend solely on the kernel. For instance,

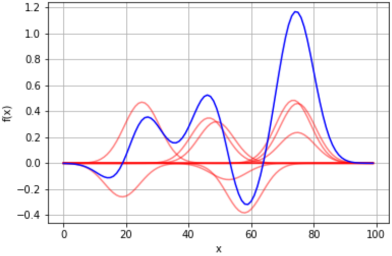
- Every $f \in \mathcal{H}$ is continuous if and only if $k(., x)$ is continous for all $x \in \mathcal{X}$
- Every $f \in \mathcal{H}$ is m-times continously differentiable if $k(., x)$ is m-times continously differentiable

Examples of RKHS

Functions in the RKHS induced by Gaussian kernel

- We have seen that functions in the RKHS can be expressed as linear combinations of kernel evaluations at points x_i

$$
\begin{aligned}
f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) &= \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \right\rangle_{\mathcal{H}}
\end{aligned}
$$



Figure

Mercer Theorem and Gaussian Kernel

- Mercer's theorem [1] gives the following expansion of kernel function:

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x')$$

[1] lets not worry about the exact statement of the theorem for now and the exam

Mercer Theorem and Gaussian Kernel

- Mercer's theorem [1] gives the following expansion of kernel function:

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x') = \langle \phi(x), \phi(x') \rangle$$

  where $\phi(x) = [\sqrt{\lambda_1} e_1(x), \ldots, \sqrt{\lambda_\ell} e_\ell(x), \ldots]^T$

- Furthermore, the eigen functions $e_\ell(x)$ need to be orthogonal in the following way :

$$\int_{\mathcal{X}} e_\ell(x) e_m(x) d\mu(x) = \begin{cases} 1 & \text{if } \ell = m \\ 0 & \text{otherwise} \end{cases}$$

- Since we are dealing with positive definite functions (instead of positive definite matrices), we have eigen functions (instead of eigen vectors)

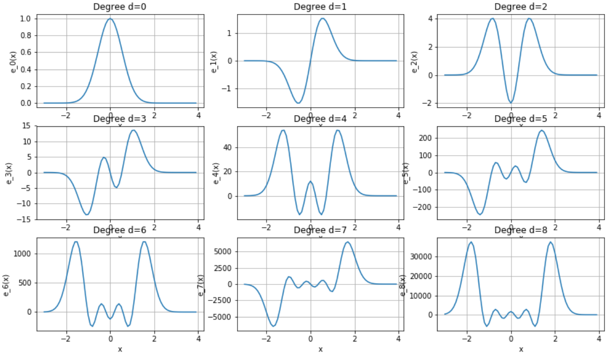
[1] lets not worry about the exact statement of the theorem for now and the exam

Mercer Theorem and Gaussian Kernel

For the Gaussian kernel $k(x, x') = \exp\left(-\frac{||x-x'||^2}{\sigma^2}\right)$

-
$$\lambda_\ell \propto b^\ell, 0 < b < 1$$

- The eigenfunctions are functions of Hermite polynomials, the first 9 of them are plotted below



Explicit Coefficients of the infinite dimensional Feature Vector

- Lets recall $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$
- From Mercer Theorem $\phi(x) = [\sqrt{\lambda_1} e_1(x), \ldots, \sqrt{\lambda_\ell} e_\ell(x), \ldots]^T$
- Further, using reproducing property $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} f_\ell \sqrt{\lambda_\ell} e_\ell(x)$

Explicit Coefficients of the infinite dimensional Feature Vector

- Lets recall $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$
- From Mercer Theorem $\phi(x) = [\sqrt{\lambda_1} e_1(x), \ldots, \sqrt{\lambda_\ell} e_\ell(x), \ldots]^T$
- Further, using reproducing property $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} f_\ell \sqrt{\lambda_\ell} e_\ell(x)$

What we now want is the explicit coefficients f_ℓ which are the components of the infinite linear function in the feature space $\mathcal{H}$.

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x_i) e_\ell(x) = \sum_{\ell=1}^{\infty} \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_\ell} e_\ell(x_i) \sqrt{\lambda_\ell} e_\ell(x)$$

So, the coefficient $f_\ell = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_\ell} e_\ell(x_i)$

Explicit Coefficients of the infinite dimensional Feature Vector

- Lets recall $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$
- From Mercer Theorem $\phi(x) = [\sqrt{\lambda_1} e_1(x), \ldots, \sqrt{\lambda_\ell} e_\ell(x), \ldots]^T$
- Further, using reproducing property $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} f_\ell \sqrt{\lambda_\ell} e_\ell(x)$

What we now want is the explicit coefficients f_ℓ which are the components of the infinite linear function in the feature space $\mathcal{H}$.

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x_i) e_\ell(x) = \sum_{\ell=1}^{\infty} \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_\ell} e_\ell(x_i) \sqrt{\lambda_\ell} e_\ell(x)$$

So, the coefficient $f_\ell = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_\ell} e_\ell(x_i)$

- For the norm to be well-defined, we need finiteness of $\langle f, f \rangle = \sum_{\ell=1}^{\infty} f_\ell^2 < \infty$,
- What does that require?
- Recall from ℓ_2-regularization for (finite dimensional case) in SVMs given by $||\mathbf{w}||_2^2 = \sum_{d=1}^{D} \mathbf{w}_d^2$

Representer Theorem

Empirical and Expected Loss

- For a loss function L, and dataset D of size n, Empirical loss $\mathcal{R}_{L,D}$ of a classifier f is defined as

$$\mathcal{R}_{L,D} := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

- The classifier obtained by minimizing the empirical loss is given by

$$f_D := \arg\min_{f:\mathcal{X}\mapsto\mathbb{R}} \mathcal{R}_{L,D}$$

- Minimizing Empirical risk over an arbitrary function class can lead to overfitting (as seen earlier in the lecture)
- Empirical risk minimization (ERM) is, therefore, performed over a smaller function class of function, which are typically smooth functions. This is given by

$$f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{R}_{L,D}$$

- Pick to be a function class with bounded RKHS norm i.e. $\{f : ||f||_{\mathcal{H}} \leq \lambda\}$

Constrained versus Penalized Problem

- Contrained formulation (from previous slide)

$$f_{\mathcal{H}} := \arg \min_{\{f:||f||_{\mathcal{H}} \leq \lambda\}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

- Lagrangian formulation

$$f_{\mathcal{H}} := \arg \min_{f} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda ||f||_{\mathcal{H}}^2$$

  for $\lambda > 0$

- The above is optimization problem over potentially an infinite dimensional space, since $\mathcal{H}$ can be an infinite dimensional as in the case of Gaussian kernel.

Representer Theorem

- For the following optimization

$$f_{\mathcal{H}} := \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\theta(||f||_{\mathcal{H}})$$

where $\theta : [0, \infty) \mapsto \mathbb{R}$ is non-decreasing function

- Even though the above problem is potentially an infinite dimensional optimization problem, **Representer Theorem** states its solution can be expressed in the following form

$$f_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i k(., x_i)$$

where $\alpha_i \in \mathbb{R}$

- Infinite to finite dimensional problem

Representer Theorem - Proof

- Decompose the RKHS $\mathcal{H}$ into the following :

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}^{\perp}$$

  - where $\mathcal{H}_0 = \{f \in \mathcal{H} : f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), (\alpha_i)_{i=1}^{n}\}$. It can also been seen as a space spanned by the kernel evaluations at the points x_i, i.e. $\mathcal{H}_0 = \text{span}\{k(., x_1), k(., x_2), \ldots, k(., x_n)\}$
  - $\mathcal{H}^{\perp}$ is the component of $\mathcal{H}$, which is orthogonal to $\mathcal{H}_0$.
- Therefore, the function $f \in \mathcal{H}$ is decomposed as

$$f = f_0 + f^{\perp}$$

-

$$f(x_i) = \langle f, k(., x_i) \rangle_{\mathcal{H}} = \langle f_0, k(., x_i) \rangle_{\mathcal{H}_0} + \langle f^{\perp}, k(., x_i) \rangle_{\mathcal{H}^{\perp}}$$

Representer Theorem - Proof

- Decompose the RKHS $\mathcal{H}$ into the following :

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}^{\perp}$$

  - where $\mathcal{H}_0 = \{f \in \mathcal{H} : f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), (\alpha_i)_{i=1}^{n}\}$. It can also been seen as a space spanned by the kernel evaluations at the points x_i, i.e. $\mathcal{H}_0 = \text{span}\{k(., x_1), k(., x_2), \ldots, k(., x_n)\}$
  - $\mathcal{H}^{\perp}$ is the component of $\mathcal{H}$, which is orthogonal to $\mathcal{H}_0$.
- Therefore, the function $f \in \mathcal{H}$ is decomposed as

$$f = f_0 + f^{\perp}$$

- $$f(x_i) = \langle f, k(., x_i)\rangle_{\mathcal{H}} = \langle f_0, k(., x_i)\rangle_{\mathcal{H}_0} + \langle f^{\perp}, k(., x_i)\rangle_{\mathcal{H}^{\perp}}$$

  The second term in the above loss term expansion is 0.

Representer Theorem - Proof

- For the regularization term use the Pythagoras theorem, i.e.

$$||f||^2_{\mathcal{H}} = ||f^{\perp}||^2_{\mathcal{H}^{\perp}} + ||f_0||^2_{\mathcal{H}_0}$$

  This implies that $||f_0||_{\mathcal{H}} \leq ||f||_{\mathcal{H}}$. Since $\theta(.)$ is a non-decreasing function $\theta(||f_0||_{\mathcal{H}}) \leq \theta(||f||_{\mathcal{H}})$

- Therefore, the optimal solution has no component in $\mathcal{H}^{\perp}$, and has the form

$$f_{\mathcal{H}}(.) = \sum_{i=1}^{n} \alpha_i k(., x_i)$$

Practical Implications of Representer Theorem

- It allows us to look for the solutions of the following form :

$$f_{\mathcal{H}}(.) = \sum_{i=1}^{n} \alpha_i k(., x_i)$$

  which are easier finite dimensional optimization problems as against the equivalent infinite dimensional original problems.

- For $j = 1 \dots n$

$$f_{\mathcal{H}}(x_j) = \sum_{i=1}^{n} \alpha_i k(x_i, x_j) = [K\boldsymbol{\alpha}]_j$$

  which is the j-th element of the matrix-vector product $K\boldsymbol{\alpha}$

- Also,

$$||f||_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$$

Least Square Regression

- Let f be the prediction function, then squared error is given by

$$\ell(f(x), y) = (y - f(x))^2$$

- Let $\mathcal{H}$ be a function class (not necessarily an RKHS) from which we are choosing our function
- Least Square regresion find a function with smallest sqaured error

$$\hat{f} \in \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Possible problems :
  - Can be unstable in high dimensions
  - Overfit if the function space $\mathcal{H}$ is too large

Kernel Ridge Regression

- Finding f in an RKHS with a kernel $k(x, x')$
- $$\hat{f} \in \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda ||f||_{\mathcal{H}}$$
- The above formulation has two advantages :
  - Prevents overfitting
  - Representer theorem enables an efficient solution of the form

$$\hat{f}(.) = \sum_{i=1}^{n} \alpha_i k(., x_i)$$

Solving Kernel Ridge Regression

- Lets denote by
  - The label vector $\mathbf{y} \in \mathbb{R}^n$ denoting the true values for the inputs
  - The kernel matrix K, where $K_{i,j} = K(x_i, x_j)$
  - $\boldsymbol{\alpha} \in \mathbb{R}^n$, the co-efficients we want to find
- For the input instance, the prediction by the desired function can be written as follows :

$$(\hat{f}(x_1), \ldots, \hat{f}(x_n))^T = K\boldsymbol{\alpha}$$

- We also know that

$$||f||^2_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$$

- Solving Kernel Ridge Regression involves solving

$$\arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{n}(K\boldsymbol{\alpha} - \mathbf{y})^T (K\boldsymbol{\alpha} - \mathbf{y}) + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$$

Kernel Ridge Regression - Solution

- Desired optimization problem

$$\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{n}(K\boldsymbol{\alpha}-\mathbf{y})^T(K\boldsymbol{\alpha}-\mathbf{y})+\lambda\boldsymbol{\alpha}^T K\boldsymbol{\alpha}$$

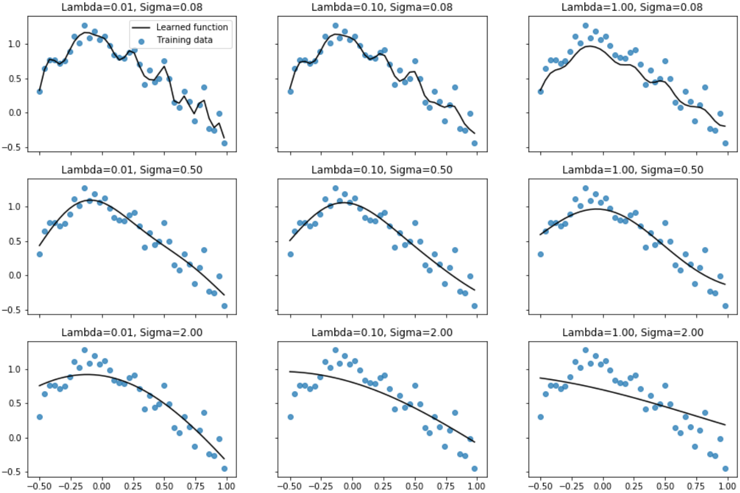
- The above is convex and differntiable w.r.t to $\boldsymbol{\alpha}$, and can be analytically found by setting the gradient

$$\frac{2}{n}K(K\alpha-\mathbf{y})+2\lambda K\alpha=\mathbf{0}$$

- Since K is positive definite (from previous lecture), we can invert $K+\lambda nI$, and hence the solution is given by

$$\boldsymbol{\alpha}=(K+\lambda nI)^{-1}\mathbf{y}$$

Kernel Ridge Regression with Gaussian Kernel

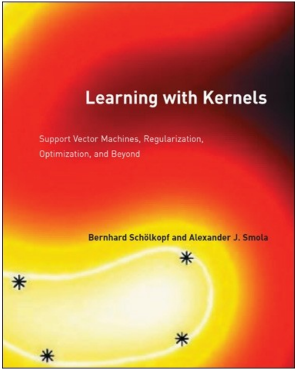


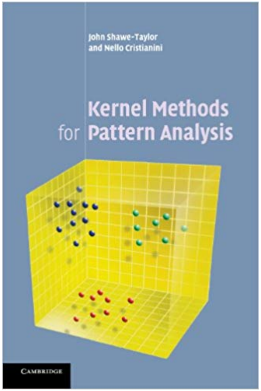
References

- For more detailed material on Gaussian RKHS, proof of RKHS equivalence
  - Lecture notes by Arthur Gretton - `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf`

Books for further study

- Learning with kernels - Schoelkopf and Smola
- Kernel Methods for Pattern Analysis - Shawe-Taylor and Christianini


Learning with Kernels
Support Vector Machines, Regularization, Optimization, and Beyond
Bernhard Schölkopf and Alexander J. Smola


John Shawe-Taylor and Nello Cristianini
Kernel Methods for Pattern Analysis
CAMBRIDGE