Basis Expansion and RKHS

Yanrong Yang

RSFAS/CBE, Australian National University

9th August 2022

Contents of this week

- Motivation of Nonlinear Modelling
- Approach under study: Basis Expansion
- High-level Extension: Reproducing Kernel Hilbert Space (RKHS)

Why and How to do Nonlinear Modelling?

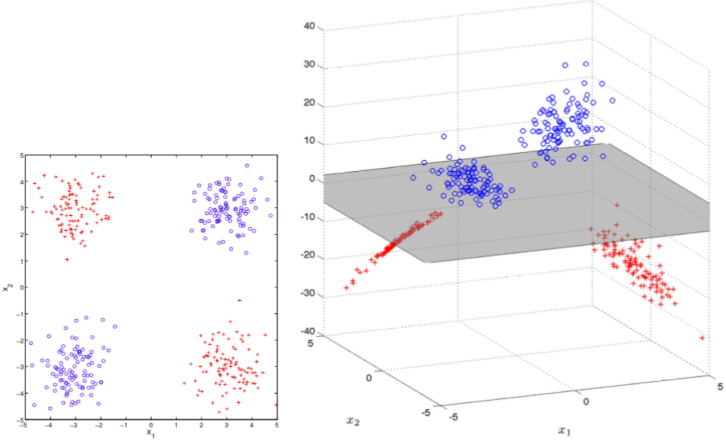
- Nonlinear Relationship is common in supervising learning (e.g. regression, classification) and unsupervised learning (e.g. clustering analysis, dimension reduction).
- Nonlinear Relationship is more popular in Big Data Analysis. As more data is collected, the chance that nonlinear phenomenon appears is larger.
- Various methods for nonlinear regression are available, like the kernel smoothing, local poynomial regression, smoothing splines.

We will discuss a systematic "nonlinear regression" method: basis expansion, which involves smoothing splines as a special case.

Motivation: Example 1 (clustering analysis)

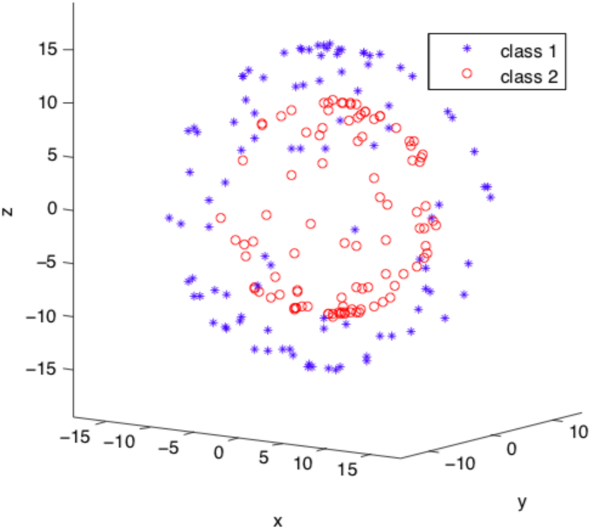
Map a set of 2-dimensional original data (left figure) to 3-dimensional new data (right figure).

$$\phi(x_1, x_2) = (x_1, x_2, x_1 x_2)$$



Motivation: Example 2 (dimension reduction)

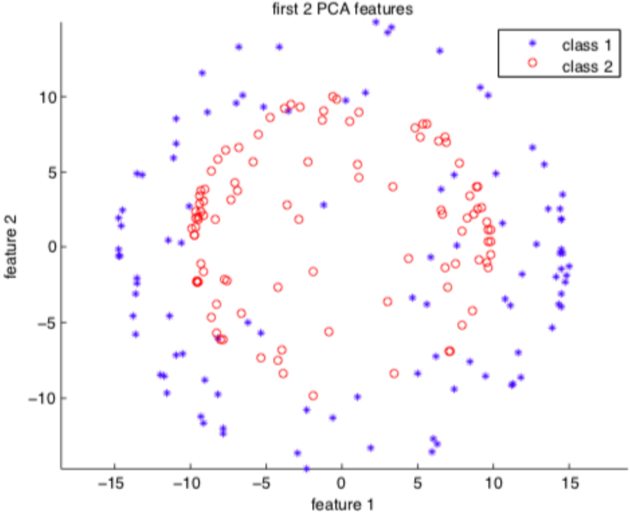
Two sets of spherical data



1

Motivation: Example 2 (dimension reduction)

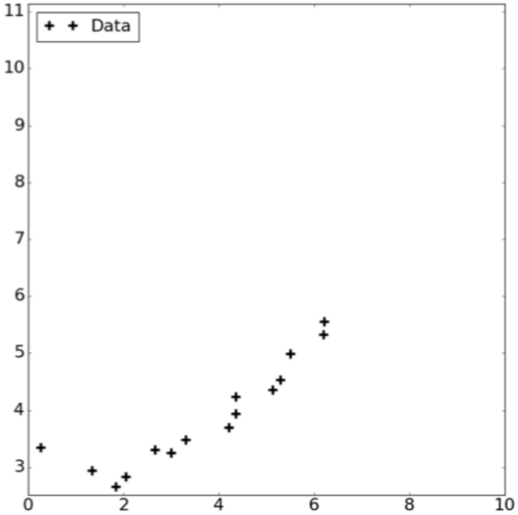
Traditional PCA is applied and new data (after PCA) is below.



2

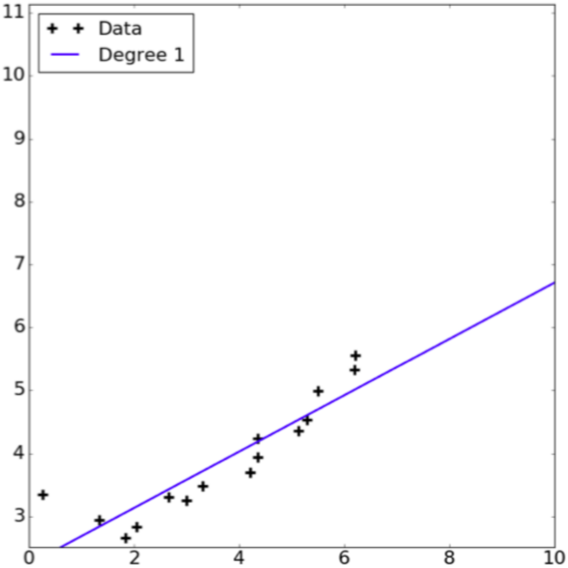
Motivation: Example 3 (regression)

Regression Problem: original data



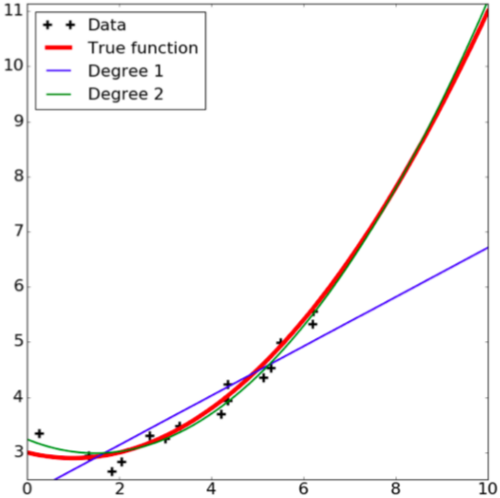
Motivation: Example 3 (regression)

Linear Regression Fitting



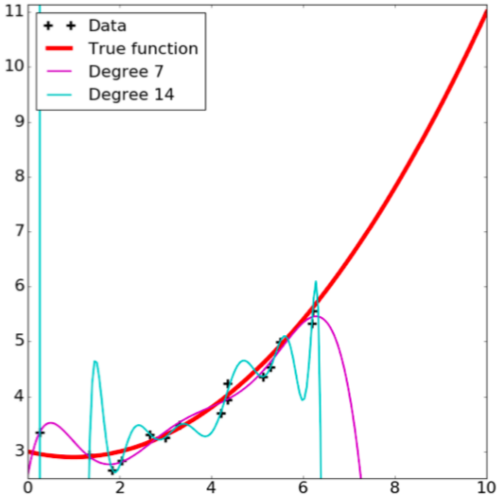
Motivation: Example 3 (regression)

$$\phi(x) = (1, x, x^2)^\top,$$
$$f(x) = \omega_0 + \omega_1 x + \omega_2 x^2 = (\omega_0, \omega_1, \omega_2)^\top \phi(x).$$



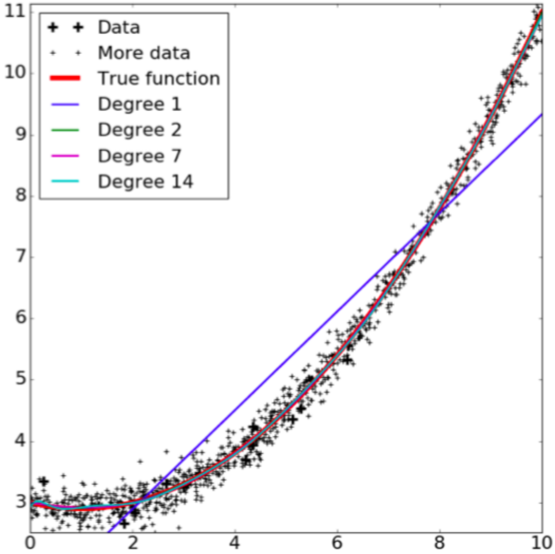
Motivation: Example 3 (regression)

$$\phi(x) = (1, x, \ldots, x^d)^\top,$$
$$f(x) = \omega_0 + \omega_1 x + \ldots + \omega_d x^d = (\omega_0, \omega_1, \ldots, \omega_d)^\top \phi(x).$$



Motivation: Example 3 (regression)

Getting more data is able to avoid overfitting.



Moving beyond Linearity

- Augment the vector of inputs X with additional variables.
- These are transformations of X

$$h_m(X) : \mathbb{R}^p \to \mathbb{R}$$

  with $m = 1, \ldots, M$.
- Then model the relationship between X and Y

$$f(X) = \sum_{m=1}^{M} \beta_m \, h_m(X) = \sum_{m=1}^{M} \beta_m \, Z_m$$

  as a linear basis expansion in X.

Common Basis Functions

- Linear:

$$h_m(X) = X_m, \; m = 1, \ldots, p$$

- Polynomial:

$$h_m(X) = X_j^2, \quad \text{or} \quad h_m(X) = X_j \, X_k$$

- Non-linear transformation of single inputs:

$$h_m(X) = \log(X_j), \sqrt{X_j}, \ldots$$

- Non-linear transformation of multiple input:

$$h_m(X) = \|X\|$$

- Use of Indicator functions:

$$h_m(X) = \mathrm{Ind}(L_m \leq X_k < U_m)$$

Control of Flexibility

- **Restriction Methods**
  Limit the class of functions considered. Use additive models

$$f(X) = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{jm}\, h_{jm}(X_j)$$

- **Selection Methods**
  Scan the set of h_m and only include those that contribute significantly to the fit of the model - Boosting, CART.

- **Regularization Methods**
  Let

$$f(X) = \sum_{j=1}^{M} \beta_j\, h_j(X)$$

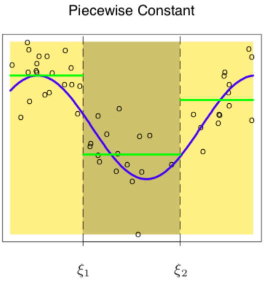
  but when learning the β_j's restrict their values in the manner of *ridge regression* and *lasso*.

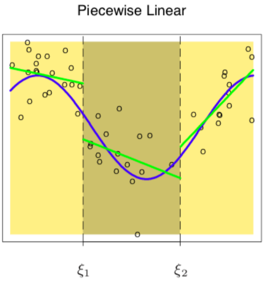
Example of Basis Expansion: Splines

To obtain a piecewise polynomial function $f(X)$

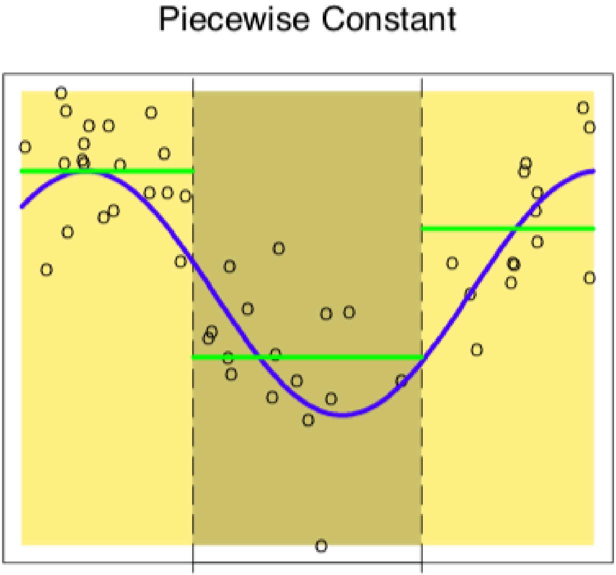
- Divide the domain of X into contiguous intervals.
- Represent f by a separate polynomial in each interval.

Examples





Piecewise Constant



Piecewise Constant

- Divide $[a, b]$, the domain of X, into three regions

  $[a, \xi_1),\ [\xi_1, \xi_2),\ [\xi_2, b]$ with $\xi_1 < \xi_2 < \xi_3$ ξ_i's are referred to as **knots**

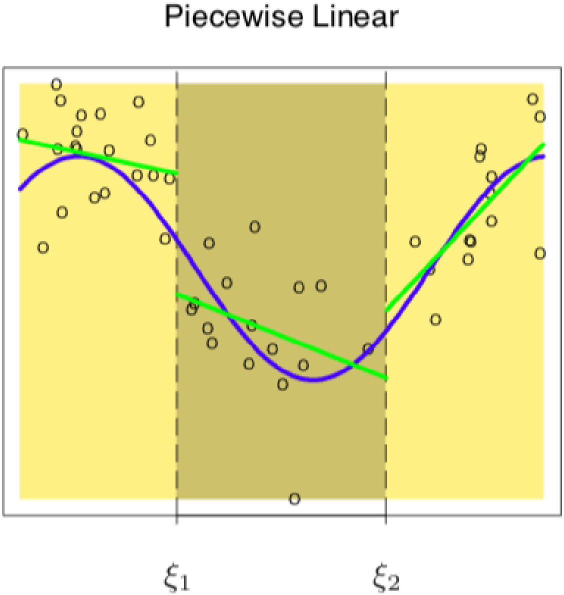
- Define three basis functions

  $h_1(X) = \mathrm{Ind}(X < \xi_1),\ \ h_2(X) = \mathrm{Ind}(\xi_1 \leq X < \xi_2),\ \ h_3(X) = \mathrm{Ind}(\xi_2 \leq X)$

- The model $f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$ is fit using least-squares.

- As basis functions don't overlap $\implies$ $\hat{\beta}_m =$ mean of y_i's in the mth region.

Piecewise Linear



Piecewise Linear

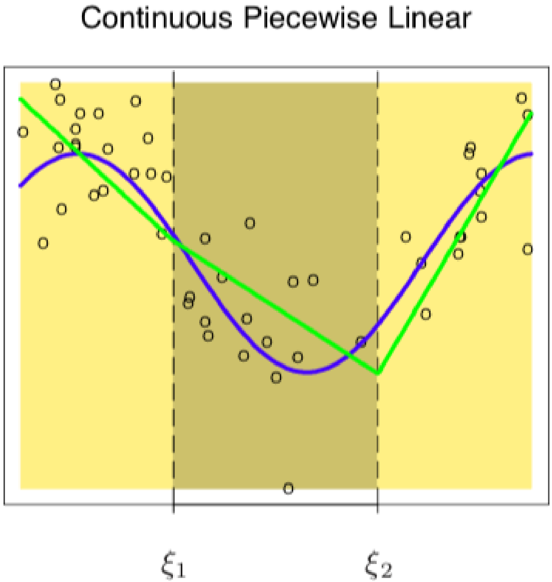
- In this case define 6 basis functions

$$h_1(X) = \mathrm{Ind}(X < \xi_1), \quad h_2(X) = \mathrm{Ind}(\xi_1 \leq X < \xi_2), \quad h_3(X) = \mathrm{Ind}(\xi_2 \leq X)$$

$$h_4(X) = X\, h_1(X), \qquad h_5(X) = X\, h_2(X), \qquad h_6(X) = X\, h_3(X)$$

- The model $f(X) = \sum_{m=1}^{6} \beta_m\, h_m(X)$ is fit using least-squares.
- As basis functions don't overlap $\implies$ fit a separate linear model to the data in each region.

Continuous Piecewise Linear



Continuous Piecewise Linear

- Additionally impose the constraint that $f(X)$ is continuous as ξ_1 and ξ_2.
- This means

$$\beta_1 + \beta_2\xi_1 = \beta_3 + \beta_4\xi_1, \text{ **and**}$$

$$\beta_3 + \beta_4\xi_2 = \beta_5 + \beta_6\xi_2$$

- This reduces the $\#$ of dof of $f(X)$ from 6 to 4.

Smooth Function

Can achieve a smoother $f(X)$ by increasing the order

- of the local polynomials
- of the continuity at the knots

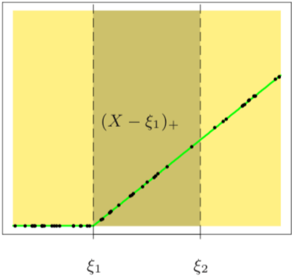
Representation of Basis Functions

- To impose the continuity constraints directly can use this basis instead:

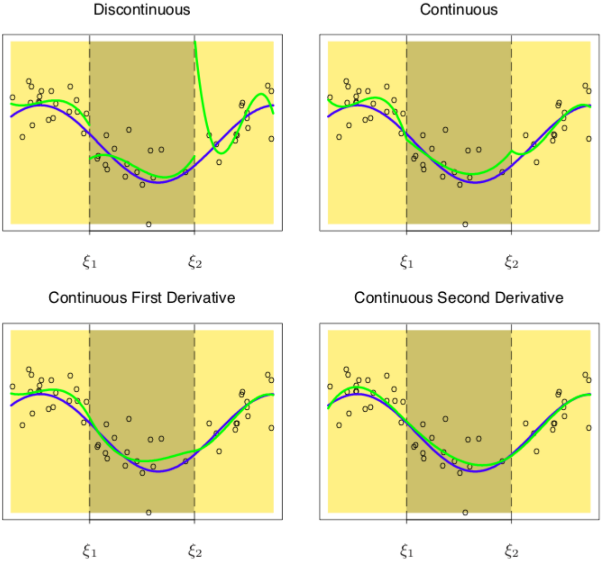
$$h_1(X) = 1 \qquad h_2(X) = X$$

$$h_3(X) = (X - \xi_1)_+ \qquad h_4(X) = (X - \xi_2)_+$$

Piecewise-linear Basis Function



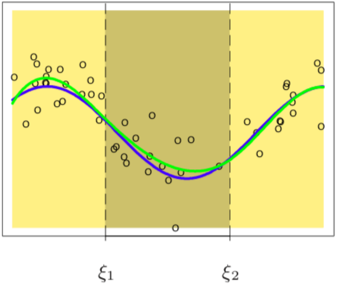
Comparison


Discontinuous
ξ_1
ξ_2
Continuous
ξ_1
ξ_2
Continuous First Derivative
ξ_1
ξ_2
Continuous Second Derivative
ξ_1
ξ_2

Cubic Spline

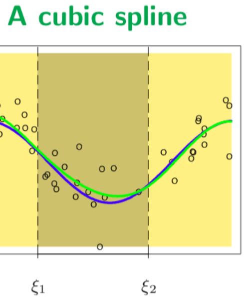
$f(X)$ is a **cubic spline** if

- it is a piecewise cubic polynomial **and**
- has 1st and 2nd continuity at the knots



A cubic spline

Representation of Cubic Spline



The following basis represents a cubic spline with knots at ξ_1 and ξ_2:

$$h_1(X) = 1, \qquad h_3(X) = X^2, \qquad h_5(X) = (X - \xi_1)_+^3$$

$$h_2(X) = X, \qquad h_4(X) = X^3, \qquad h_6(X) = (X - \xi_2)_+^3$$

Regression Spline

- An order M spline with knots $\xi_1, \ldots, \xi_K$ is
  - a piecewise-polynomial of order M **and**
  - has continuous derivatives up to order $M - 2$
- The general form for the truncated-power basis set is

$$h_j(X) = X^{j-1} \quad j = 1, \ldots, M$$
$$h_{M+l}(X) = (X - \xi_l)_+^{M-1}, \quad l = 1, \ldots, K$$

- In practice the most widely used orders are $M = 1, 2, 4$.

Regression Spline

- Fixed-knot splines are known as regression splines.
- For a regression spline one needs to select
  - the order of the spline,
  - the number of knots and
  - the placement of the knots.
- One common approach is to set a knot at each observation x_i.
- There are many equivalent bases for representing splines and the **truncated power basis** is intuitively attractive but not computationally attractive.
- A better basis set for implementation is the B-spline basis set.

Natural Cubic Spline

Problem

The polynomials fit beyond the boundary knots behave wildly.

Solution: Natural Cubic Splines

- Have the additional constraints that the function is linear beyond the boundary knots.
- This frees up 4 dof which can be used by having more knots in the interior region.
- Near the boundaries one has reduced the variance of the fit but increased its bias!

Drawbacks of Regression Splines

1. Regression splines have advantage over polynomial regression due to more flexibility, but they do have one shortcoming: the placement of knots.
2. Choices regarding the number of knots and where they are located are not particularly easy to make in a systematic and data-driven manner.
3. assuming that you place knots at quantiles or equally spaced intervals, models will not be nested inside each other, which complicates hypothesis testing.

Remedy Idea: Roughness Penalty Approach

Recall the aims of regression splines which

1. good measure of fit as most as possible;
2. smoothing curve.

Two criteria can reflect these two aims, respectively

1. $$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{1}$$

2. $$\min_f \int \left[f''(x)\right]^2 dx \tag{2}$$

Alternative Method: Smoothing Splines

Smoothing Spline is the solution to the optimization problem below

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \left[f''(x)\right]^2 dx, \tag{3}$$

where $\lambda > 0$ is a penalized parameter or smoothing parameter.
Note:

1. smoothing splines have a penalized version of the least squares objective function;
2. the first term captures the fit to the data while the second penalizes smoothness of the curve.

Connection with Regression Splines

The smoothing parameter λ controls the trade-off between the two aspects

1. $\lambda = 0$ imposes no restrictions and f will therefore interpolate the data.
2. $\lambda = \infty$ renders curvature impossible, thereby returning us to ordinary linear regression.

It may sound impossible to solve for such an f over all possible functions, but the solution turns out to be surprisingly simple: smoothing spline f is a natural cubic spline.

Interpretation of This Connection (1)

One Theorem for natural cubic splines:
Out of all twice-differentiable functions passing through the points (x_i, y_i), $i = 1, 2, \ldots, n$, the one that minimizes

$$\lambda \int \left[f''(x) \right]^2 dx \tag{4}$$

is a natural cubic spline with knots at every unique value of x_i, $i = 1, 2, \ldots, n$.

Interpretation of This Connection (2)

One Theorem that connects natural cubic splines with smoothing splines:

Out of all twice-differentiable functions, the one that minimizes

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \left[f^{''}(x) \right]^2 dx, \tag{5}$$

is a natural cubic spline with knots at every unique value of x_i, $i = 1, 2, \ldots, n$.

Solution to Smoothing Splines (1)

Let N_j, $j = 1, 2, \ldots, n$ denote the collection of natural cubic spline basis functions and $\mathbf{N}$ denote the $n \times n$ design matrix consisting of the basis functions evaluated at the observed values:

1. $N_{ij} = N_j(x_i)$.
2. $f(x) = \sum_{j=1}^{n} N_j(x)\beta_j$
3. $f(\mathbf{x}) = \mathbf{N}\boldsymbol{\beta}$.

Solution to Smoothing Splines (2)

The penalized objective function is therefore

$$(\mathbf{y} - \mathbf{N}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{N}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta}, \tag{6}$$

where $\boldsymbol{\Omega}_{jk} = \int N_j''(x) N_k''(x) dx$.
The solution is therefore

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{N}^T \mathbf{N} + \lambda \boldsymbol{\Omega}\right)^{-1} \mathbf{N}^T \mathbf{y} \tag{7}$$

Hilbert Space

A Hilbert Space is a complete inner product space. We will see that a reproducing kernel Hilbert space (RKHS) is a Hilbert space with extra structure that makes it useful for statistics and machine learning.

- Let $\mathcal{K}$ be a set of functions taking values on $\mathbb{R}^1$. A two-variable function $\langle \cdot, \cdot \rangle : \mathcal{K} \times \mathcal{K} \to \mathbb{R}^1$ is said to be an inner product on $\mathcal{K}$ if

$$\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle. \quad (1)$$

$$\langle f, g \rangle = \langle g, f \rangle. \quad (2)$$

$$\langle f, f \rangle \geq 0, \; \textit{and} \; \langle f, f \rangle = 0 \; \textit{if and only if} \; f = 0. \quad (3)$$

- A norm on $\mathcal{K}$ is defined as

$$||f||_{\mathcal{K}} = \sqrt{\langle f, f \rangle}. \quad (4)$$

High-level Extension of Basis Expansion

- Essential idea of basis expansion is to represent a function $f(x)$ by basis functions $\boldsymbol{\phi}(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_d(x))$, i.e.

$$f(x) = \omega_0 + \omega_1\phi_1(x) + \ldots + \omega_d\phi_d(x). \tag{5}$$

  In regression problem, nonparametric estimation for $f(x)$ is equivalent to parametric estimation for parameters $\omega_j, j = 0, 1, \ldots, d$.
- Finding the basis functions $\phi_1(x), \phi_2(x), \ldots, \phi_d(x)$ is important.
- We will utilize kernel function to define feature maps (or basis functions).

Kernel Function

A RKHS is defined by a **Mercer kernel**. A Mercer kernel $K(x, y)$ is a function of two variables that is symmetric and positive definite. This means that, for any function f,

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

(This is like the definition of a positive definite matrix: $x^T Ax \geq 0$ for each x.)

Our main example is the Gaussian kernel

$$K(x,y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

Common Kernel Functions

- Polynomial Kernels

$$K(x, y) = (\langle x, y \rangle + c)^m \,. \tag{6}$$

- Exponential Kernels

$$K(x, y) = \exp\left(\langle x, y \rangle\right) . \tag{7}$$

- Taylor Series Kernels

$$K(x, y) = \sum_{n=0}^{\infty} a_n \langle x, y \rangle^n . \tag{8}$$

Feature Maps

- Given a Kernel Function $K(x, y)$, we define the feature map as

$$\phi(x) = K_y(x), \tag{9}$$

where $K_y(x)$ is the function $K(x, y)$ obtained by fixing y.

- For the Gaussian kernel, $K_y(x)$ is a normal density, centered at the point y.
- In summary, given one value y, we have a feature map $\phi(x)$.

Basis Expansion

- We create functions by taking linear combinations of the feature maps:

$$f(x) = \sum_{j=1}^{K} \omega_j \phi_j(x) = \sum_{j=1}^{K} \omega_j K_{y_j}(x). \tag{10}$$

- Let $\mathcal{H}$ denote all such functions

$$\mathcal{H} := \left\{ f : \sum_{j=1}^{K} \omega_j K_{y_j}(x) \right\} \tag{11}$$

Reproducing Kernel Hilbert Space (RKHS)

To make the set $\mathcal{H}$ into a space, we define an inner product and norm:

- For two functions $f(x) = \sum_{j=1}^{K} \alpha_j K_{x_j}(x)$ and $g(x) = \sum_{j=1}^{K} \beta_j K_{y_j}(x)$,

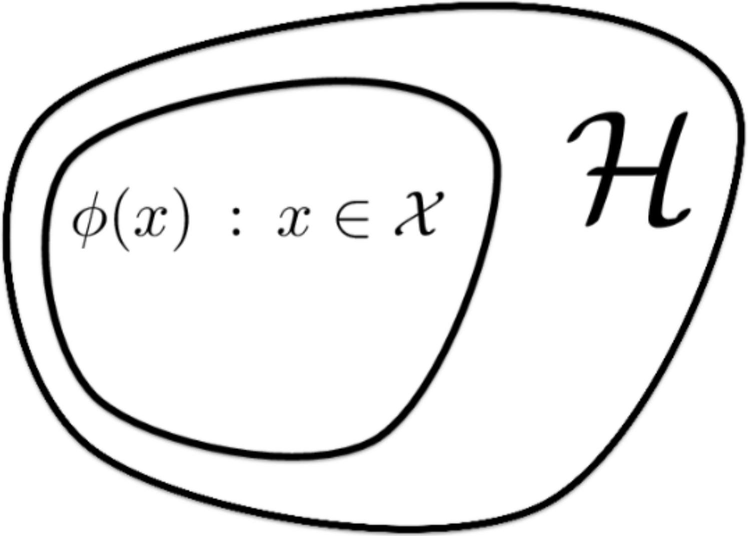
$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j). \tag{12}$$

- This inner product defines a norm

$$||f|| = \sqrt{\langle f, f \rangle} = \sqrt{\sum_{j=1}^{K} \sum_{k=1}^{K} \alpha_j \alpha_k K(x_j, x_k)} = \sqrt{\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}}, \tag{13}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ and $\mathbf{K} = (K(x_j, x_k))_{K \times K}$.

Illustration of Feature Maps and RKHS



Why RKHS? Reproducing Property

- For any function $f(x) = \sum_{j=1}^{K} K_{y_j}(x) \in \mathcal{H}$, we calculate the inner product between $f(\cdot)$ and the feature map $K_x(\cdot)$:

$$\langle f, K_x \rangle = \sum_{j=1}^{K} \alpha_j K(y_j, x) = f(x). \tag{14}$$

  This "evolutional property" is called reproducing property.
- Let $f(x) = K_y(x)$ in (14). Then

$$\langle K_y, K_x \rangle = K(x, y). \tag{15}$$

Application: Regularized Nonparametric Regression

Consider a penalized nonparametric regression problem

$$\widehat{f} = \arg\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + J(||f||^2), \tag{16}$$

where $J(\cdot)$ is any monotone increasing function. Then $\widehat{f}$ has the form

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x). \tag{17}$$

Note that the kernel function $K(x, y)$ here corresponds to the inner product $||\cdot||$ in (16).

Conclusion

Understand

- Basis Expansion methods: regression splines, smoothing splines.
- Reproducing Kernel Hilbert Space (RKHS): basic framework, common kernels, relation with regularization.