REVIEW OF FUNCTIONAL ANALYSIS*

A function space is a space of functions where each function can be thought of as a point in Euclidean space. Functional analysis is loosely speaking a mathematical understanding of function spaces. In the next lecture we will study a very useful function space called a Reproducing kernel Hilbert space (riches) which is used extensively in non-linear regression.

4.1. Hilbert Spaces

Examples. *The following are three examples of function spaces defined on a subset of the real line. In these examples the subset of the real line we consider is* $x \in [a, b]$ *where for example* $a = 0$ *and* $b = 10$.

(1) $C[a, b]$ *is the set of all real-valued continuous functions on* $x \in [a, b]$.
$y = x^3$ *is in* $C[a, b]$ *while* $y = \lceil x \rceil$ *is not in* $C[a, b]$.

(2) $L_2[a, b]$ *is the set of all square integrable functions on* $x \in [a, b]$. *If* $(\int_a^b |f(x)|^2\, dx)^{1/2} < \infty$ *then* $f \in L_2[a, b]$.
$y = x^3$ *is in* $L_2[a, b]$ *and so is* $y = x^3 + \delta(x - c)$ *where* $a < c < b$, *however the second function is not defined at* $x = c$.

(3) $L_1[a, b]$ *is the set of all functions whose absolute value is integrable on* $x \in [a, b]$.
$y = x^3$ *is in* $L_1[a, b]$ *and so is* $y = x^3 + \delta(x - c)$ *where* $a < c < b$, *however the second function is not defined at* $x = c$.

Definition. *A normed vector space is a space* $\mathcal{F}$ *in which a norm is defined. A function* $\|\cdot\|$ *is a norm iff for all* $f, g \in \mathcal{F}$

(1) $\|f\| \geq 0$ *and* $\|f\| = 0$ iff $f = 0$
(2) $\|f + g\| \leq \|f\| + \|g\|$
(3) $\|\alpha f\| = |\alpha|\ \|f\|$.

Note, if all conditions are satisfied except $\|f\| = 0$ iff $f = 0$ *then the space has a seminorm instead of a norm.*

Definition. *An inner product space is a linear vector space* $\mathcal{E}$ *in which an inner product is defined. A real valued function* $\langle \cdot, \cdot \rangle$ *is an inner product iff* $\forall f, g, h \in \mathcal{E}$ *and* $\alpha \in \mathbb{R}$

(1) $\langle f, g \rangle = \langle g, f \rangle$
(2) $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$ *and* $\langle \alpha f, g \rangle = \alpha \langle f, g \rangle$
(3) $\langle f, f \rangle \geq 0$ *and* $\langle f, f \rangle = 0$ *iff* $f = 0$.

Given an inner product space the norm is defined as $\|f\| = \sqrt{\langle f, f \rangle}$ *and an angle between vectors can be defined.*

Definition. *For a normed space* $\mathcal{A}$ *a subspace* $\mathcal{B} \subset \mathcal{A}$ *is dense in* $\mathcal{A}$ *iff* $\mathcal{A} = \bar{\mathcal{B}}$. *Where* $\bar{\mathcal{B}}$ *is the closure of the set* $\mathcal{B}$.

Definition. *A normed space* $\mathcal{F}$ *is separable iff* $\mathcal{F}$ *has a countable dense subset.*

Example. *The set of all rational points is dense in the real line and therefore the real line is separable. Note, the set of rational points is countable.*

Counterexample. *The space of right continuous functions on* $[0, 1]$ *with the sup norm is not separable. For example, the step function*

$$f(x) = U(x - a) \;\; \forall a \in [0, 1]$$

cannot be approximated by a countable family of functions in the sup norm since the jump must occur at a *and the set of all* a *is uncountable.*

Definition. *A sequence* $\{x_n\}$ *in a normed space* $\mathcal{F}$ *is called a Cauchy sequence if* $\lim_{n\to\infty} \sup_{m \geq n} \|x_n - x_m\| = 0$.

Definition. *A normed space* $\mathcal{F}$ *is called complete iff every Cauchy sequence in it converges.*

Definition. *A Hilbert space,* $\mathcal{H}$ *is an inner product space that is complete, separable, and generally infinite dimensional.*
A Hilbert space has a countable basis.

Examples. *The following are examples of Hilbert spaces.*

(1) $\mathbb{R}^n$ *is the textbook example of a Hilbert space. Each point in the space* $x \in \mathbb{R}^n$ *can be represented as a vector* $x = \{x_1, ..., x_n\}$ *and the metric in this space is* $\|x\| = \sqrt{\sum_{i=1}^n |x_i|^2}$. *The space has a very natural basis composed of the* n *basis functions* $e_1 = \{1, 0, ..., 0\}$, $e_2 = \{0, 1, ..., 0\}$,..., $e_n = \{0, 0, ..., 1\}$. *The inner product between a vector* x *and a basis vector* e_i *is simply the projection of* x *onto the* i*th coordinate* $x_i = \langle x, e_i \rangle$.
Note, this is not an infinite dimensional Hilbert space.

(2) L_2 *is also a Hilbert space. This Hilbert space is infinite dimensional.*

4.2. Functionals and operators

Definition. *A linear functional on a Hilbert space* $\mathcal{H}$ *is a linear transformation* $T : V \to \mathbb{R}$ *from* $\mathcal{H}$ *into* $\mathbb{R}$.

A linear functional takes an element in a Hilbert space and outputs a real number, integration is an example of a linear functional.

Theorem (Riesz representation theorem)**.** *Let* V *be a finite-dimensional inner product space and let* $T : V \to \mathbb{R}$ *be a linear functional. Then there is a vector* $w \in V$ *such that* $Tv = \langle v, w \rangle$ *for all* $v \in V$.

An integral transformation is one example of an operator (in the rest of the course all examples of operators will be integral transforms). An operator T maps one vector space into another.

Definition. *An integral transform T maps one function into another function as follows*

$$g(u) = (Tf)(u) := \int_{t_1}^{t_2} K(t,u) f(t)\,\mathrm{d}t.$$

LECTURE 5
Reproducing kernel Hilbert spaces

Reproducing Kernel Hilbert Spaces (rkhs) are hypothesis spaces with some very nice properties. The main property of these spaces is the reproducing property which relates norms in the Hilbert space to linear algebra. This class of functions also has a nice interpretation in the context of Gaussian processes. Thus, they are important for computational, statistical, and functional reasons.

5.1. Reproducing Kernel Hilbert Spaces (rkhs)

We will use two formulations to describe rkhs. The first is less general and more constructive. The second is more general and abstract. The key idea in both formulations is that there is a kernel function $K : X \times X \to \mathbb{R}$ and this kernel function has associated to it a Hilbert space $\mathcal{H}_K$ that has wondrous properties for optimization and inference.

The algorithm we will study in detail in the next lecture is the following

$$\widehat{f} := \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K},$$

where $\mathcal{H}_K$ is a rkhs and $\|f\|_{\mathcal{H}_K}$ is specific norm as defined by the reproducing kernel K. The beauty of the rkhs is the optimization problem in the above infinite dimensional function space can be rewritten as a quadratic programming problem which involves only vectors and matrices.

For the remainder of this lecture we constrain the Hilbert spaces to a compact domain X.

5.1.1. Constructive formulation

The development of rkhs in this subsection is seen in most formulations of Support Vector Machines (SVMs) and Kernel Machines. It is less general in that it relies on the reproducing kernel being a Mercer Kernel. It however requires less knowledge of functional analysis and is more intuitive for most people.

We start by defining the kernel or reproducing kernel function.

Definition. *The reproducing kernel (rk), $K(\cdot, \cdot)$ is a symmetric real valued function of two variables $s, t \in X$*

$$K(s, t) : X \times X \to \mathbb{R}.$$

In addition $K(s,t)$ must be positive definite, that is for all real $a_1, ..., a_n$ and $t_1, ..., t_n \in X$

$$\sum_{i,j=1}^{n} a_i a_j K(t_i, t_j) \geq 0.$$

If the above inequality is strict then $K(s,t)$ is strictly positive definite.

In this formulation we consider continuous reproducing kernels $K : X \times X \to \mathbb{R}$. We define an integral operator $L_K : L_2[X] \to C[X]$ by the following integral transform

$$L_K f := \int_X K(s,t) f(t) dt = g(t). \tag{5.1}$$

If K is positive definite then L_K is positive definite (the converse is also true) and therefore the eigenvalues of (5.1) are nonnegative.

We denote the eigenvalues and eigenvectors of (5.1) as $\{\lambda_1, ..., \lambda_k\}$ and $\{\phi_1, ..., \phi_k\}$ respectively, where

$$\int_X K(s,t)\phi_k(t)dt = \lambda_k \phi_k(t) \ \ \forall k.$$

We now state Mercer's theorem.

Theorem. *Given the eigenfunctions and eigenvalues of the integral equation defined by a symmetric positive definite kernel K*

$$\int_X K(s,t)\phi(s)ds = \lambda\, \phi(t).$$

The kernel has the expansion

$$K(s,t) = \sum_j \lambda_j \phi_j(s)\phi_j(t),$$

where convergence is in the $L_2[X]$ norm.

We can define the rkhs as the space of functions spanned by the eigenfunctions of the integral operator defined by the kernel

$$\mathcal{H}_K = \{\, f \,|f(s) = \sum_k c_k \phi_k(s) \ \text{ and } \ \|f\|_{\mathcal{H}_K} < \infty\},$$

where the rkhs norm $\|\cdot\|_{\mathcal{H}_K}$ is defined as follows

$$\|f(s)\|^2_{\mathcal{H}_K} = \left\langle \sum_j c_j \phi_j(s), \sum_j c_j \phi_j(s) \right\rangle^2_{\mathcal{H}_K} := \sum_j \frac{c_j^2}{\lambda_j}.$$

Similarly the inner product is defined as follows

$$\langle f, g \rangle = \left\langle \sum_j c_j \phi_j(s), \sum_j d_j \phi_j(s) \right\rangle_{\mathcal{H}_K} := \sum_j \frac{d_j c_j}{\lambda_j}.$$

Part of a homework problem will be to prove the representer property

$$\langle f(\cdot, K(\cdot, x)\rangle_{\mathcal{H}_k} = f(x),$$

using Mercer's theorem and the above definition of the rkhs norm.

5.1.2. Kernels and feature space

The rkhs concept has been utilized in the SVM and kernel machines literature in what is unfortunately called the kernel trick.

Points in the domains $x \in X \subset \mathbb{R}^d$ are mapped into a higher dimensional space by the eigenvalues and eigenfunctions of the reproducing kernel (the space is of the dimensionality of the number of nonzero eigenvalues of the integral operator defined by the kernel)

$$x \to \Phi(x) := \{\sqrt{\lambda_1}\phi_1(x), ..., \sqrt{\lambda_k}\phi_k(x)\}.$$

A standard L_2 inner product of two points mapped into the feature space can be evaluated by a kernel due to Mercer's theorem

$$K(s,t) = \langle \Phi(s), \Phi(t) \rangle_{L_2}.$$

5.1.3. Examples of kernel functions

Any (semi) positive definite function can be used as a kernel function. Examples include

(1) Linear kernel: $k(u,v) = \langle u, v \rangle$
(2) Polynomial kernel: $k(u,v) = (\langle u, v \rangle + b)^p$
(3) Gaussian kernel: $k(u,v) = \exp(-\kappa \|u - v\|^2)$
(4) Double exponential kernel: $k(u,v) = \exp(-\kappa \|u - v\|)$

5.2. Abstract formulation

Proposition. *A linear evaluation function L_t evaluates each function in a Hilbert space $f \in \mathcal{H}$ at a point t. It associates $f \in \mathcal{H}$ to a number $f(t) \in \mathbb{R}$, $L_t[f] = f(t)$.*

(1) $L_t[f + g] = f(t) + g(t)$
(2) $L_t[af] = af(t)$.

Example. *The delta function $\delta(x - t)$ is a linear evaluation function for $C[a, b]$*

$$f(t) = \int_a^b f(x)\delta(x - t)dx.$$

Proposition. *A linear evaluation function is bounded if there exists an M such that for all functions in the Hilbert space $f \in \mathcal{H}$*

$$|L_t[f]| = |f(t)| \leq M\|f\|,$$

where $\|f\|$ is the Hilbert space norm.

Example. *For the Hilbert space $C[a, b]$ with the sup norm there exists a bounded linear evaluation function since $|f(x)| \leq M$ for all functions in $C[a, b]$. This is due to continuity and compactness of the domain. The evaluation function is simply $L_t[f] : t \to f(t)$ and $M = 1$.*

Counterexample. *For the Hibert space $L_2[a, b]$ there exists no bounded linear evaluation function. The following function is in $L_2[a, b]$*

$$y = x^3 + \delta(x - c) \quad \text{where } a < c < b.$$

At the point $x = c$ there is no M such that $|f(c)| \leq M$ since the function is evaluated as "∞". This is an example of a function in the space that is not even defined pointwise.

Definition. *If a Hilbert space has a bounded linear evaluation function, L_t, then it is a Reproducing Kernel Hilbert Space (rkhs), $\mathcal{H}_K$.*

The following property of a rkhs is very important and is a result of the Riesz representation theorem.

Proposition. *If $\mathcal{H}_K$ is a rkhs then there exists an element in the space K_t with the property such that for all $f \in \mathcal{H}_K$*

$$L_t[f] = \langle K_t, f \rangle = f(t).$$

The inner product is in the rkhs norm and the element K_t is called the representer of evaluation of t.

Remark. The above property is somewhat amazing in that it says if a Hilbert space has a bounded linear evaluation function then there is an element in this space that evaluates all functions in the space by an inner product.
In the space $L_2[a, b]$ we say that the delta function evaluates all functions in $L_2[a, b]$

$$L_t[f] = \int_a^b f(x)\delta(x - t)dx.$$

However, the delta function is not in $L_2[a, b]$.

There is a deep relation between a rkhs and its reproducing kernel. This is characterized by the following theorem.

Theorem. *For every Reproducing Kernel Hilbert Space (rkhs) there exists a unique reproducing kernel and conversely given a positive definite function K on $X \times X$ we can construct a unique rkhs of real valued functions on X with K as its reproducing kernel (rk).*

Proof.
If $\mathcal{H}_K$ is a rkhs then there exists an element in the rkhs that is the representer evaluation by the Reisz representer theorem. We define the rk

$$K(s, t) := \langle K_s, K_t \rangle$$

where K_s and K_t are the representers of evaluation at s and t. The following hold by the properties of Hilbert spaces and the representer property

$$\begin{aligned}
\left\| \sum_j a_j K_{t_j} \right\|^2 &\geq 0 \\
\left\| \sum_j a_j K_{t_j} \right\|^2 &= \sum_{i,j} a_i a_j \langle K_{t_i}, K_{t_j} \rangle \\
\sum_{i,j} a_i a_j K(t_i, t_j) &= \sum_{i,j} a_i a_j \langle K_{t_i}, K_{t_j} \rangle.
\end{aligned}$$

Therefore $K(s, t)$ is positive definite.

We now prove the converse. Given a rk $K(\cdot, \cdot)$ we construct $\mathcal{H}_K$. For each $t \in X$ we define the real valued function

$$K_t(\cdot) = K(t, \cdot).$$

We can show that the rkhs is simply the completion of the space of functions spanned by the the set K_{t_i}

$$\mathcal{H} = \{f \,|\, f = \sum_i a_i K_{t_i} \quad \text{where} \quad a_i \in \mathbb{R},\ t_i \in X,\ \text{and} \quad i \in \mathbb{N}\}$$

with the following inner product

$$\left\langle \sum_i a_i K_{t_i}, \sum_i a_i K_{t_i} \right\rangle = \sum_{i,j} a_i a_j \langle K_{t_i}, K_{t_j} \rangle = \sum_{i,j} a_i a_j K(t_i, t_j).$$

Since $K(\cdot, \cdot)$ is positive definite the above inner product is well defined. For any $f \in \mathcal{H}_K$ we can check that

$$\langle K_t, f \rangle = f(t)$$

because for any function in the above linear space norm convergence implies pointwise convergence

$$|f_n(t) - f(t)| = |\langle f_n - f, K_t \rangle| =\leq \|f_n - f\| \|K_t\|,$$

the last step is due to Cauchy-Schwartz. Therefore every Cauchy sequence in this space converges and it is complete. □

LECTURE 6
Non-linear regression

The algorithm we will study in detail in the next lecture is the following

$$\widehat{f} := \arg\min_{f\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda\|f\|^2_{\mathcal{H}_K}.$$

We will see the above minimizer as a particular form well suited for optimization due to the representer theorem.

6.1. A result of the representer theorem

The following are the three standard regularization methods:

(1) Tikhonov regularization: indirectly constrain the hypothesis space by adding a penalty term.

$$\min_{f\in\mathcal{H}}\left[n^{-1}\sum_{i=1}^{n}V(f,z_i)+\lambda\Omega(f)\right].$$

(2) Ivanov regularization: directly constrain the hypothesis space

$$\min_{f\in\mathcal{H}}\left[n^{-1}\sum_{i=1}^{n}V(f,z_i)\right] \quad \text{subject to} \quad \Omega(f)\le\tau.$$

(3) Phillips regularization: directly constrain the hypothesis space

$$\min_{f\in\mathcal{H}}\Omega(f) \quad \text{subject to} \quad \left[n^{-1}\sum_{i=1}^{n}V(f,z_i)\right]\le\kappa.$$

Consider the rkhs norm will be as the regularization functional

$$\Omega(f)=\|f\|^2_{\mathcal{H}_K}.$$

This defines the following optimization problems:

$$(P1) \qquad \min_{f\in\mathcal{H}} \left[n^{-1}\sum_{i=1}^{n} V(f,z_i) + \lambda \|f\|^2_{\mathcal{H}_K} \right],$$

$$(P2) \qquad \min_{f\in\mathcal{H}} \left[n^{-1}\sum_{i=1}^{n} V(f,z_i) \right] \quad \text{subject to} \quad \|f\|^2_{\mathcal{H}_K} \leq \tau,$$

$$(P3) \qquad \min_{f\in\mathcal{H}} \|f\|^2_{\mathcal{H}_K} \quad \text{subject to} \quad \left[n^{-1}\sum_{i=1}^{n} V(f,z_i) \right] \leq \kappa.$$

All the above optimization problems above are over spaces of functions that contain an infinite number of functions. Using the formulation in section 5.1.1 we can write any function in the rhks as

$$\mathcal{H}_K = \left\{ f \,|f(x) = \sum_k c_k \phi_k(x) \right\},$$

so the optimization procedure is over the coefficients c_k. The number of nonzero coefficients in the expansion defines the dimensionality of the rkhs and this can be infinite, for example the Gaussian kernel.

One of the amazing aspects of the all above optimization problems is that a minimizer satisfies the form

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

So the optimization procedure is over n real variables. This is formalized in the following "Representer Theorem."

Theorem. *Given a set of points $\{(x_i,y_i)\}_{i=1}^n$ a function of the form*

$$\widehat{f}(x) = \sum_{i=1}^{n} c_i K(x,x_i),$$

is a minimizer of the following optimization procedure

$$c\left((f(x_1),y_1),...,(f(x_n),y_n)\right) + \lambda g(\|f\|_{\mathcal{H}_K}),$$

where $\|f\|_{\mathcal{H}_K}$ is a rkhs norm, $g(\cdot)$ is monotonically increasing, and c is an arbitrary cost function.

Procedure (P1) is special case of the optimization procedure stated in the above theorem.

Proof. For ease of notation all norms and inner products in the proof are rkhs norms and inner products.

Assume that the function f has the following form

$$f = \sum_{i=1}^{n} b_i \phi_i(x_i) + v,$$

where

$$\langle \phi_i(x_i), v \rangle = 0 \quad \forall i = 1,..,n.$$

The orthogonality condition simple ensures that v is not in the span of $\{\phi_i(x_i)\}_{i=1}^n$.

So for any point x_j $(j = 1, ..., n)$

$$f(x_j) = \left\langle \sum_{i=1}^{n} b_i \phi(x_i) + v, \phi(x_j) \right\rangle = \sum_{i=1}^{n} b_i \langle \phi(x_i), \phi(x_j) \rangle,$$

so v has no effect on the cost function

$$c\left((f(x_1), y_1), ..., (f(x_n), y_n)\right).$$

We now look at the rkhs norm

$$g(\|f\|) = g\left(\left\|\sum_{i=1}^{n} b_i \phi_i(x_i) + v\right\|\right) = g\left(\sqrt{\left\|\sum_{i=1}^{n} b_i \phi_i(x_i)\right\|^2 + \|v\|^2}\right) \geq g\left(\sqrt{\left\|\sum_{i=1}^{n} b_i \phi_i(x_i)\right\|^2}\right).$$

So the extra factor v increases the rkhs norm and has effect on the cost functional and therefore must be zero and the function has the form

$$\widehat{f} = \sum_{i=1}^{n} b_i \phi_i(x_i),$$

and by the reproducing property

$$\widehat{f}(x) = \sum_{i=1}^{n} a_i K(x, x_i). \ \square$$

Homework: proving a representer theorem for the other two regularization formulations.

6.2. Kernel ridge-regression

The Kernel ridge-regression (KRR) algorithm has been invented and reinvented many times and has been called a variety of names such as Regularization networks, Least Square Support Vector Machine (LSSVM), Regularized Least Square Classification (RLSC).

We start with Tikhonov regularization

$$\min_{f \in \mathcal{H}_K} \left[n^{-1} \sum_{i=1}^{n} V(f, z_i) + \lambda \Omega(f) \right]$$

and then set the regularization functional to a RKHS norm

$$\Omega(f) = \|f\|^2_{\mathcal{H}_K}$$

and use the square loss functional

$$n^{-1} \sum_{i=1}^{n} V(f, z_i) = n^{-1} \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$

The resulting optimization problem is

$$\min_{f \in \mathcal{H}_K} \left[n^{-1} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K} \right], \tag{6.1}$$

the minimizer of which we know by the Representer theorem has the form

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

This implies that we only need to solve the optimization problem for the c_i. This turns the problem of optimizing over functions which maybe infinite-dimensional into a problem of optimizing over n real numbers.

Using the representer theorem we derive the optimization problem actually solved for Kernel ridge-regression.

We first define some notation. We will use the symbol K to refer to either the kernel function K or the $n \times n$ matrix K where

$$K_{ij} \equiv K(x_i, x_j).$$

Using this definition the function $f(x)$ evaluated at a training point x_j can be written in matrix notation as

$$\begin{aligned} f(x_j) &= \sum_{i=1}^{n} K(x_i, x_j) c_i \\ &= [Kc]_j, \end{aligned}$$

where $[Kc]_j$, is the jth element of the vector obtained in multiplying the kernel matrix K with the vector c. In this notation we can rewrite equation (6.1) as

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} (Kc - y)^2 + \lambda \|f\|_K^2,$$

where y is the vector of y values. Also by the representer theorem the RKHS norm can be evaluated using linear algebra

$$||f||_K^2 = c^T K c,$$

where c^T is the transpose of the vector c. Substituting the above norm into equation (6.1) results in an optimization problem on the vector c

$$\arg \min_{c \in \mathbb{R}^n} \left[g(c) := \frac{1}{\ell} (Kc - y)^2 + \lambda c^T K c. \right]$$

This is a convex, differentiable function of c, so we can minimize it simply by taking the derivative with respect to c, then setting this derivative to 0.

$$\frac{\partial g(c)}{\partial c} = \frac{2}{\ell} K(Kc - y) + 2\lambda K c = 0.$$

We show that the solution of the above equation is the following linear system

$$c = (K + \lambda \ell I)^{-1} y,$$

where I is the identity matrix:

$$\begin{array}{cc} \text{differentiation} & 0 = \frac{2}{\ell} K(Kc - y) + 2\lambda K c \\ \text{multiplication} & K(Kc) + \lambda \ell K c = Ky \\ \text{“left multiplication by } K^{-1}\text{”} & (K + \lambda \ell I) c = y \\ \text{inversion} & c = (K + \lambda \ell I)^{-1} y. \end{array}$$

The matrix $K + \lambda \ell I$ is positive definite and will be well-conditioned if λ is not too small.

A few properties of the linear system are:

(1) The matrix $(K + \lambda \ell I)$ is guaranteed to be invertible if $\lambda > 0$. As $\lambda \to 0$, the regularized least-squares solution goes to the standard Gaussian least-squares solution which minimizes the empirical loss. As $\lambda \to \infty$, the solution goes to $f(\mathbf{x}) = 0$.
(2) In practice, we don't actually invert $(K + \lambda \ell I)$, but instead use an algorithm for solving linear systems.
(3) In order to use this approach, we need to compute and store the entire kernel matrix K. This makes it impractical for use with very large training sets.

Lastly, there is nothing to stop us for using the above algorithm for classification. By doing so, we are essentially treating our classification problem as a regression problem with y values of 1 or -1.

6.2.1. Solving for c

The conjugate gradient (CG) algorithm is a popular algorithm for solving positive definite linear systems. For the purposes of this class, we need to know that CG is an iterative algorithm. The major operation in CG is multiplying a vector v by the matrix A. Note that matrix A need not always be supplied explicitly, we just need some way to form a product Av.

For ordinary positive semidefinite systems, CG will be competitive with direct methods. CG can be much faster if there is a way to multiply by A quickly.

Example. *Suppose our kernel K is linear:*

$$K(x, y) = \langle x, y \rangle.$$

Then our solution x can be written as

$$\begin{aligned} f(x) &= \sum c_i \langle x_i, x \rangle \\ &= \left\langle \left(\sum c_i x_i \right), x \right\rangle \\ &:= \langle w, x \rangle, \end{aligned}$$

and we can apply our function to new examples in time d rather than time nd.

This is a general property of Tikhonov regularization with a linear kernel, not related to the use of the square loss.

We can use the CG algorithm to get a huge savings for solving regularized least-squares regression with a linear kernel ($K(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{x_1} \cdot \mathbf{x_2}$). With an arbitrary kernel, we must form a product Kv explicitly — we multiply a vector by K. With the linear kernel, we note that $K = AA^T$, where A is a matrix with the data points as row vectors. Using this:

$$\begin{aligned} (K + \lambda n I)v &= (AA^T + \lambda n I)v \\ &= A(A^T v) + \lambda n I v. \end{aligned}$$

Suppose we have n points in d dimensions. Forming the kernel matrix K explicitly takes $n^2 d$ time, and multiplying a vector by K takes n^2 time.

If we use the linear representation, we pay nothing to form the kernel matrix, and multiplying a vector by K takes $2dn$ time.

If $d \ll n$, we save approximately a factor of $\frac{n}{2d}$ per iteration. The memory savings are even more important, as we cannot store the kernel matrix at all for

large training sets, and if were to recompute the entries of the kernel matrix as needed, each iteration would cost $n^2 d$ time.

Also note that if the training data are sparse (they consist of a large number of dimensions, but the majority of dimensions for each point are zero), the cost of multiplying a vector by K can be written as $2\bar{d}n$, where $\bar{d}$ is the average number of nonzero entries per data point.

This is often the case for applications relating to text, where the dimensions will correspond to the words in a "dictionary". There may be tens of thousands of words, but only a few hundred will appear in any given document.

6.3. Equivalence of the three forms

The three forms of regularization have a certain equivalence. The equivalence is that given a set of points $\{(x_i, y_i)\}_{i=1}^n$ the parameters $\lambda, \tau,$ and κ can be set such that the same function $f(x)$ minimizes (P1), (P2), and (P3). Given this equivalence and and the representer theorem for (P1) it is clear that a representer theorem holds for (P2) and (P3).

Proposition. *Given a convex loss function function the following optimization procedures are equivalent*

$$(P1) \qquad \min_{f \in \mathcal{H}_K} \left[n^{-1} \sum_{i=1}^n V(f, z_i) + \lambda \|f\|_{\mathcal{H}_K}^2 \right],$$

$$(P2) \qquad \min_{f \in \mathcal{H}_K} \left[n^{-1} \sum_{i=1}^n V(f, z_i) \right] \quad \textit{subject to} \quad \|f\|_{\mathcal{H}_K}^2 \leq \tau,$$

$$(P3) \qquad \min_{f \in \mathcal{H}_K} \|f\|_{\mathcal{H}_K}^2 \quad \textit{subject to} \quad \left[n^{-1} \sum_{i=1}^n V(f, z_i) \right] \leq \kappa.$$

By equivalent we mean that if $f_0(x)$ is a solution of one of the problems then there exist parameters τ, κ, λ for which $f_0(x)$ is a of the others.

Proof.

Let f_0 be the solution of (P2) for a fixed τ and assume that the constraint under the optimization is tight ($\|f_0\|_{\mathcal{H}_K}^2 = \tau$). Let $\left[n^{-1} \sum_{i=1}^n V(f_0, z_i) \right] = b$. By inspection the solution of (P3) with $\kappa = b$ will be f_0.

Let f_0 be the solution of (P3) for a fixed κ and assume that the constraint under the optimization is tight ($[n^{-1} \sum_{i=1}^n V(f_0, z_i)] = \kappa$). Let $\|f_0\|_{\mathcal{H}_K}^2 = b$. By inspection the solution of (P2) with $\tau = b$ will be f_0.

For both (P2) and (P3) the above argument can be adjusted for the case where the constraints are not tight but the solution f_0 is not necessarily unique.

Let f_0 be the solution of (P2) for a fixed τ. Using Lagrange multipliers we can rewrite (P2) as

$$\min_{f \in \mathcal{H}_K, \alpha} \left[n^{-1} \sum_{i=1}^n V(f, z_i) \right] + \alpha \left(\|f\|_{\mathcal{H}_K}^2 - \tau \right), \tag{6.2}$$

where $\alpha \geq 0$ the optimal $\alpha = \alpha_0$. By the Karush-Kuhn-Tucker (KKT) conditions (complimentary slackness) at optimality

$$\alpha_0 \left(\|f_0\|_{\mathcal{H}_K}^2 - \tau \right) = 0.$$

LECTURE 9
Gaussian process regression

The idea behind a Gaussian process regression is to place a distribution over a space of functions say $\mathcal{H}$. Consider for example an rkhs $\mathcal{H}_K$ over which we want to do Bayesian inference. Assume a regression model with the standard noise assumption

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad f \in \mathcal{H}_K.$$

If we knew how to place a prior over the function space we in theory could do Bayesian inference.

9.1. Gaussian process

A Gaussian process is a specification of probability distributions over functions $f(x)$, $f \in \mathcal{H}$ and $x \in \mathcal{X}$ parameterized ny a mean function μ and a covariance function $K(\cdot, \cdot)$. The idea can be informally stated as

$$p(f) \propto \exp\Big(-\frac{1}{2}\|f\|^2_{\mathcal{H}_K}\Big), \quad p(f) \geq 0 \,\forall\, f \in \mathcal{H}, \int_{f\in\mathcal{H}} p(f)\,\mathrm{d}f = 1,$$

where we use the term informal because $\mathrm{d}f$ is not well defined, it is not clear what the normalization constant is for $p(f)$ and what the space of functions $\mathcal{H}$ is not clear not is the relation of $\mathcal{H}$ to $\mathcal{H}_K$ stated clearly. Instead of making all the points clear we will develop Gaussian processes from an alternative perspective. There are many ways to define and think about a Gaussian process. A standard formulation is that a Gaussian process is an infinite version of a multivariate Gaussian distribution and has two parameters: a mean function μ corresponding to the mean vector and a positive definite covariance or kernel function K corresponding to a positive definite covariance matrix.

A common approach in defining an infinite dimensional object is by defining it's finite dimensional projections. This is the approach we will take with a Gaussian process. Consider $x_1, ..., x_n$ as a finite collection of points in $\mathcal{X}$. For a Gaussian process over functions $f \in \mathcal{H}$ the probability density of $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$ is a multivariate normal with $\boldsymbol{\mu} = \{\mu(x_1), ..., \mu(x_n)\}$ and covariance $\boldsymbol{\Sigma}_{ij} = K(x_i, x_j)$

$$\mathrm{f} \sim \mathrm{N}\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\Big),$$

where $\mu(x) = \mathbb{E}f(x)$ and $K(x_i, x_j) = \mathbb{E}\big[(f(x_i) - \mu(x_i))(f(x_j) - \mu(x_j))\big]$ and

$$f \sim \mathcal{G}P(\mu(\cdot), K(\cdot, \cdot)).$$

Definition. *A stochastic process over domain $\mathcal{X}$ with mean function μ and covariance kernel K is a Gaussian process if and only if for any $\{x_1, ..., x_n\} \in \mathcal{X}$ and $n \in \mathbb{N}$ the distribution of $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$ is*

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim N\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_1, x_n) & \cdots & K(x_n, x_n) \end{bmatrix} \right).$$

9.2. Gaussian process regression

Consider data $D = \{(x_i, y_i)\}_{i=1}^n$ drawn from the model

$$Y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2),$$

we will place a prior on the space of functions using a Gaussian process

$$f \sim \mathcal{G}P(\mu(\cdot), K(\cdot, \cdot)).$$

We are also given some new variables or test data $T = \{x_i^*\}_{i=1}^m$ each of which would have a corresponding y_i^*.

We now provide some notation

$$\mathbf{X} = \begin{bmatrix} - x_1 - \\ \vdots \\ - x_n - \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} - x_1^* - \\ \vdots \\ - x_m^* - \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{Y}^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_m^* \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\varepsilon}^* = \begin{bmatrix} \varepsilon_1^* \\ \vdots \\ \varepsilon_m^* \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad \mathbf{f}^* = \begin{bmatrix} f(x_1^*) \\ \vdots \\ f(x_m^*) \end{bmatrix}.$$

Our ultimate objective will be to specify the predictive distribution on $\mathbf{Y}^*$ which we know will be multivariate normal

$$\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X} \sim \mathrm{N}(\mu^*, \Sigma^*).$$

Now first observe

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}^* \end{bmatrix} \Big| \, \mathbf{X}^*, \mathbf{X} = \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}^* \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) + \sigma^2 \mathbf{I} \end{bmatrix} \right),$$

where $K(\mathbf{X}, \mathbf{X})$ is the $n \times n$ matrix with $\mathbf{K}_{ij} = K(x_i, x_j)$ and $K(\mathbf{X}^*, \mathbf{X}^*)$ is the $m \times m$ matrix with $\mathbf{K}_{ij}^* = K(x_i^*, x_j^*)$.

To get to the predictive distribution on $\mathbf{Y}^*$ we write the conditional $\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X}$. Given the above multivariate normal distribution we simply condition on all the other variables to get the mean and covariance for the normal distribution for the posterior predictive density:

$$\begin{aligned} \mu^* &= K(\mathbf{X}^*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \mathbf{Y} \\ \Sigma^* &= K(\mathbf{X}^*, \mathbf{X}^*) + \sigma^2 \mathbf{I} - K(\mathbf{X}^*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} K(\mathbf{X}, \mathbf{X}^*). \end{aligned}$$

The beauty of Gaussian process regression is that we can place priors over functions using a kernel and evaluating the variance of the function values at a finite number of points, all just based on properties of the multivariate normal

distribution. This is a very powerful non-linear prediction tool. There is a strong relation between the kernels, rkhs and Gaussian processes. There are also some subtle differences. The main difference comes from what is called the Kalianpur $0-1$ law

Theorem (Kallianpur 1970)**.** *If $Z \sim \mathcal{G}P(\mu, K)$ is a Gaussian process with covariance kernel K and mean $\mu \in \mathcal{H}_K$ and $\mathcal{H}_K$ is infinite dimensional then*

$$\mathbf{P}(Z \in \mathcal{H}_K) = 0.$$

The point of the above theorem is that if we specify a kernel K and ensure the mean of the Gaussian process is in the rkhs $\mathcal{H}_K$ corresponding to the kernel K, draws from this Gaussian process will not be in the rkhs. What one can formally show is that if one takes any of the random functions, call them g then the following is true for all g

$$\int_{\mathcal{X}} g(u)K(x,u)\,\mathrm{d}u \in \mathcal{H}_K.$$