Representer Theorem

We want to identify approaches like SVM and Ridge Regression that admit Kernels. In particular we want to determine objective functions that allow for Kernels. We will try to spot these objective functions by looking for weights that should lie in the span of the data.

Hilbert Spaces

Recall that a Hilbert space is a vector space with an inner product. What is the connection between inner products and norms?

Inner products are similarity scores between vectors. Norms are measures of length of vectors.

If we have an inner product $\langle \cdot, \cdot \rangle$ then we can form the associated norm

$$\|v\| := (\langle v,v \rangle)^{1/2}$$

What if we start from a norm?

Parallelogram Law

We can set

$$\langle x,y \rangle = \frac{\|x+y\|^2 - \|x\|^2 - \|y\|^2}{2}$$

However do the properties of a norm

- $\|v\| \geq 0$ with equality if and only if $v=0$
- $\|cv\| = |c| \cdot \|v\|$ for all numbers c
- $\|v+w\| \leq \|v\| + \|w\|$

give us the properties of an inner product

- $\langle v,w \rangle = \langle w,v \rangle$
- $\langle v+cw, z \rangle = \langle v,z \rangle + c\langle w,z \rangle$
- $\langle v,v \rangle \geq 0$ with equality if and only if $v=0$

If we can check an identity called the parallelogram rule

$$\|v+w\|^2 + \|v-w\|^2 = 2\|v\|^2 + 2\|w\|^2$$

then we have an inner product. Note that some norms like l_1 do not satisfy the identity.

Orthogonal Vectors

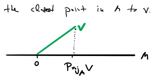
Two vectors v and w are orthogonal when $\langle v,w \rangle = 0$. We can think of orthogonal vectors as being at right angles. For orthogonal vectors we can simplify the parallelogram law

$$\begin{aligned} \|v+w\|^2 &= \langle v,v \rangle + \langle w,w \rangle + 2\langle v,w \rangle \\ &= \|v\|^2 + \|w\|^2 \end{aligned}$$

Orthogonal vectors arise in projections onto subspaces.

Projections

Suppose we have a vector v and a subspace M. The projection $\mathrm{Proj}_M v$ of v onto M is the closest point in M to v.



Note that we have a decomposition

$$v = (v - \mathrm{Proj}_M v) + \mathrm{Proj}_M v$$

into orthogonal vectors. Therefore we have

$$\|\mathrm{Proj}_M v\| \leq \|v\|$$

We can interpret this inequality as saying projections decrease length.

Span of the Data

In the representer theorem we have an objective function of the form

$$L(\langle w,x_1 \rangle, \ldots, \langle w,x_n \rangle) + R(\|w\|)$$

Here L is an arbitrary function of inner products with $\{x_1, \ldots, x_n\}$ training data. Here R is a nondecreasing function.

Take M to be the subspace of feature space generated by $\{x_1, \ldots, x_n\}$. Note

$$\begin{aligned} \langle w, x_i \rangle &= \langle (w - \mathrm{Proj}_M w) + \mathrm{Proj}_M w, x_i \rangle \\ &= \langle \mathrm{Proj}_M w, x_i \rangle \end{aligned}$$

We know that

$$\|w\| \geq \|\mathrm{Proj}_M w\|$$

Therefore

$$R(\|w\|) \geq R(\|\mathrm{Proj}_M w\|)$$

So the optimal solution w^* should be in M. We can restrict the search to vectors of the form $w = \sum \alpha_i x_i$