The Representer Theorem

David S. Rosenberg

New York University

February 13, 2018

Contents

Inner Product Spaces and Projections (Hilbert Spaces)

Inner Product Space (or "Pre-Hilbert" Spaces)

An **inner product space** (over reals) is a vector space $\mathcal{V}$ and an **inner product**, which is a mapping

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbf{R}$$

that has the following properties $\forall x, y, z \in \mathcal{V}$ and $a, b \in \mathbf{R}$:

- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$
- Linearity: $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$
- Positive-definiteness: $\langle x, x \rangle \geqslant 0$ and $\langle x, x \rangle = 0 \iff x = 0$.

Norm from Inner Product

For an inner product space, we define a norm as

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

Example

$\mathbf{R}^d$ with standard Euclidean inner product is an inner product space:

$$\langle x, y\rangle := x^T y \qquad \forall x, y \in \mathbf{R}^d.$$

Norm is

$$\|x\| = \sqrt{x^T x}.$$

What norms can we get from an inner product?

Theorem (Parallelogram Law)

A norm $\|\cdot\|$ can be written in terms of an inner product on $\mathcal{V}$ iff $\forall x, x' \in \mathcal{V}$

$$2\|x\|^2 + 2\|x'\|^2 = \|x+x'\|^2 + \|x-x'\|^2,$$

and if it can, the inner product is given by the ***polarization identity***

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

Example

ℓ_1 norm on $\mathbf{R}^d$ is NOT generated by an inner product. [Exercise]

Is ℓ_2 norm on $\mathbf{R}^d$ generated by an inner product?

Orthogonality (Definitions)

Definition

Two vectors are **orthogonal** if $\langle x, x' \rangle = 0$. We denote this by $x \perp x'$.

Definition

x is orthogonal to a set S, i.e. $x \perp S$, if $x \perp s$ for all $x \in S$.

Pythagorean Theorem

Theorem (Pythagorean Theorem)

If $x \perp x'$, *then* $\|x+x'\|^2 = \|x\|^2 + \|x'\|^2$.

Proof.

We have

$$\begin{aligned} \|x+x'\|^2 &= \langle x+x', x+x' \rangle \\ &= \langle x,x \rangle + \langle x,x' \rangle + \langle x',x \rangle + \langle x',x' \rangle \\ &= \|x\|^2 + \|x'\|^2. \end{aligned}$$

$\square$

Projection onto a Plane (Rough Definition)

- Choose some $x \in \mathcal{V}$.
- Let M be a subspace of inner product space $\mathcal{V}$.
- Then m_0 is the **projection of** x **onto** M**,**
  - if $m_0 \in M$ and is the closest point to x in M.
- In math: For all $m \in M$,

$$\|x - m_0\| \leqslant \|x - m\|.$$

Hilbert Space

- Projections exist for all finite-dimensional inner product spaces.
- We want to allow infinite-dimensional spaces.
- Need an extra condition called **completeness**.
- A space is **complete** if all Cauchy sequences in the space converge.

Definition

A **Hilbert space** is a complete inner product space.

Example

Any finite dimensional inner product space is a Hilbert space.

The Projection Theorem

Theorem (Classical Projection Theorem)

- *$\mathcal{H}$ a Hilbert space*
- *M a closed subspace of $\mathcal{H}$ (picture a hyperplane through the origin)*
- *For any $x \in \mathcal{H}$, there **exists a unique** $m_0 \in M$ for which*

$$\|x - m_0\| \leqslant \|x - m\| \; \forall m \in M.$$

- *This m_0 is called the **[orthogonal] projection of** x **onto** M.*
- *Furthermore, $m_0 \in M$ is the projection of x onto M iff*

$$x - m_0 \perp M.$$

Projection Reduces Norm

Theorem

Let M be a closed subspace of $\mathcal{H}$. For any $x \in \mathcal{H}$, let $m_0 = Proj_M x$ be the projection of x onto M. Then

$$\|m_0\| \leqslant \|x\|,$$

with equality only when $m_0 = x$.

Proof.

$$\begin{aligned} \|x\|^2 &= \|m_0 + (x - m_0)\|^2 \text{ (note: } x - m_0 \perp m_0 \text{ by Projection theorem)} \\ &= \|m_0\|^2 + \|x - m_0\|^2 \text{ by Pythagorean theorem} \\ \|m_0\|^2 &= \|x\|^2 - \|x - m_0\|^2 \end{aligned}$$

Then $\|x - m_0\|^2 \geqslant 0$ implies $\|m_0\|^2 \leqslant \|x\|^2$. If $\|x - m_0\|^2 = 0$, then $x = m_0$, by definition of norm. □

Representer Theorem

Generalize from SVM Objective

- SVM objective:

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[\langle w, x_i\rangle\right]\right).$$

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle\right),$$

where

  - $R : [0, \infty) \to \mathbf{R}$ is nondecreasing (**Regularization term**)
  - and $L : \mathbf{R}^n \to \mathbf{R}$ is arbitrary. (**Loss term**)

General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right),$$

  where
  - $w, x_1, \ldots, x_n \in \mathcal{H}$ for some Hilbert space $\mathcal{H}$. (We typically have $\mathcal{H} = \mathbf{R}^d$.)
  - $\|\cdot\|$ is the norm corresponding to the inner product of $\mathcal{H}$. (i.e. $\|w\| = \sqrt{\langle w, w \rangle}$)
  - $R : [0, \infty) \to \mathbf{R}$ is nondecreasing (**Regularization term**), and
  - $L : \mathbf{R}^n \to \mathbf{R}$ is arbitrary (**Loss term**).

General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right),$$

- What's "linear"?
- The prediction/score function $x \mapsto \langle w, x_i \rangle$ is linear – in what?
  - in parameter vector w, and
  - in the feature vector x_i.
- Why? [Real-valued] inner products are linear in each argument.
- **The important part is the linearity in the parameter** w.

General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right),$$

- Ridge regression and SVM are of this form.
- What if we penalize with $\lambda \|w\|_2$ instead of $\lambda \|w\|_2^2$? Yes!.
- What if we use lasso regression? No! ℓ_1 norm does not correspond to an inner product.

The Representer Theorem

Theorem (Representer Theorem)

Let

$$J(w) = R\left(\|w\|\right) + L\left(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle\right),$$

where

- $w, x_1, \ldots, x_n \in \mathcal{H}$ *for some Hilbert space* $\mathcal{H}$. *(We typically have* $\mathcal{H} = \mathbf{R}^d$.*)*
- $\|\cdot\|$ *is the norm corresponding to the inner product of* $\mathcal{H}$. *(i.e.* $\|w\| = \sqrt{\langle w, w\rangle}$*)*
- $R : [0, \infty) \to \mathbf{R}$ *is nondecreasing (**Regularization term**), and*
- $L : \mathbf{R}^n \to \mathbf{R}$ *is arbitrary (**Loss term**).*

Then

- *If* $M = span(x_1, \ldots, x_n)$, *then* $J(Proj_M w) \leqslant J(w)$ *for any* $w \in \mathcal{H}$.
- *If* $J(w)$ *has a minimizer, then it **has a minimizer of the form*** $w^* = \sum_{i=1}^n \alpha_i x_i$.
- *If R is strictly increasing, then all minimizers have this form. (Proof in homework.)*

The Representer Theorem (Proof)

1. Fix any $w \in \mathcal{H}$.
2. Let $w_M = \text{Proj}_M w$.
3. Then $w_M^\perp := w - w_M$ is orthogonal to M.
4. So $\langle w, x_i \rangle = \left\langle w_M + w_M^\perp, x_i \right\rangle = \langle w_M, x_i \rangle \; \forall i$, and
5. $L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right) = L\left(\langle w_M, x_1 \rangle, \ldots, \langle w_M, x_n \rangle\right)$.
6. Projections decrease norms: $\|w_M\| \leqslant \|w\|$.
7. Since R is nondecreasing, $R(\|w_M\|) \leqslant R(\|w\|)$.
8. $J(w_M) \leqslant J(w)$. [Proves first result.]
9. If w^* minimizes $J(w)$, then $w_M^* = \text{Proj}_M w^*$ is also a minimizer, since $J(w_M^*) \leqslant J(w^*)$.
10. So $\exists \alpha$ s.t. $w_M^* = \sum_{i=1}^n \alpha_i x_i$ is a minimizer of $J(w)$.

Q.E.D.

Sufficient Condition for Existence of a Minimizer

Theorem

[a] *Let*

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

and let $M = \text{span}(x_1, \ldots, x_n)$. Then under the same conditions given in the Representer theorem, if w_M^* minimizes $J(w)$ ***over the set*** *M, then w_M^* minimizes $J(w)$ over all $\mathcal{H}$.*

[a] Thanks to Mingsi Long for suggesting this nice theorem and proof.

- One consequence of the Representer theorem only applies if $J(w)$ has a minimizer over $\mathcal{H}$. This theorem tells us that it's sufficient to check for a constrained minimizer of $J(w)$ over M. If one exists, then it's also an unconstrained minimizer of $J(w)$ over $\mathcal{H}$. If there is no constrained minimizer over M, then $J(w)$ has no minimizer over $\mathcal{H}$ (by the Representer theorem).
- Bottom Line: We can jump straight to minimizing over M, the "span of the data".

Sufficient Condition for Existence of a Minimizer (Proof)

1. Let $w_M^* \in \arg\min_{w \in M} J(w)$. [the constrained minimizer]
2. Consider any $w \in \mathcal{H}$.
3. Let $w_M = \text{Proj}_M w$.
4. By the Representer theorem, $J(w_M) \leqslant J(w)$.
5. $J(w_M^*) \leqslant J(w_M)$ by definition of w_M^*.
6. Thus for any $w \in \mathcal{H}$, $J(w_M^*) \leqslant J(w)$.
7. Therefore w_M^* minimizes $J(w)$ over $\mathcal{H}$

QED