Machine Learning Theory

Lecture ?: Sobolev Regression

David A. Hirshberg

May 24, 2024

Emory University

Smoothness constraints

So far, we've talked about two models based on smoothness constraints.

$$\mathcal{M} = \left\{ m : \int_0^1 |m'(x)| dx \le B \right\}$$ The Bounded Variation Model

$$\mathcal{M} = \left\{ m : \max_x |m'(x)| \le B \right\}$$ The Lipschitz Model

If we wanted *stronger* smoothness contraints, e.g. so we don't overfit a small sample, we could use similar bounds on higher order derivatives.

$$\mathcal{M} = \left\{ m : \int_0^1 |m^{(p)}(x)| dx \le B \right\}$$ The Bounded Variation $(p-1)$st Derivative Model

$$\mathcal{M} = \left\{ m : \max_x |m^{(p)}(x)| \le B \right\}$$ The Lipschitz $(p-1)$st Derivative Model

- These use Bounded Variation and Lipschitz constraints on the $(p-1)$st derivative.
- These are fine models, and they all generalize just fine to higher dimensions.
- But we'll focus on one that's similar, but more convenient: the *Sobolev* model.

$$\mathcal{M} = \left\{ m : \int_0^1 m^{(p)}(x)^2 \, dx \le B^2 \right\}.$$

It bounds the mean square of the derivative's absolute value, not the max or mean.

The Sobolev Model

What makes this model convenient

There's an equivalent definition in terms of an *orthogonal basis* for functions on $[0,1]$.

$$\mathcal{M} = \left\{ m : \int_0^1 m^{(p)}(x)^2\, dx \le 1 \right\} = \left\{ \sum_{j=0}^{\infty} b_j \phi_j(x) : \sum_{j=0}^{\infty} \lambda_j b_j^2 \le 1 \right\}$$

$$\text{where} \qquad \int_0^1 \phi_j(x)\phi_k(x)\, dx = 0 \quad \text{for} \quad j \ne k.$$

- We call this a *Fourier series representation*.
- It makes stuff looks a bit like what you'd see in intro classes.
- We can think of the *higher order terms* — ϕ_j where λ_j is large — much like we thought about quadratic terms, interactions, etc., in linear regression.

Advantages

1. It's familiar.
    - It can help us explain things to people with intro-stats level background.
    - And understand their work better.
2. It's easy.
    - We don't need clever model-specific tricks to code up and understand things.
    - We did for using Lipschitz or Bounded Variation or Monotone Regression models.
3. It generalizes very naturally to functions of multi-dimensional covariates.
    - Once we know how to do stuff in 1D, we're good.

What makes this model convenient

There's an equivalent definition in terms of an *orthogonal basis* for functions on $[0, 1]$.

$$\mathcal{M} = \left\{ m : \int_0^1 m^{(p)}(x)^2 \, dx \le 1 \right\} = \left\{ \sum_{j=0}^{\infty} b_j \phi_j(x) : \sum_{j=0}^{\infty} \lambda_j b_j^2 \le 1 \right\}$$

$$\text{where} \qquad \int_0^1 \phi_j(x)\phi_k(x) \, dx = 0 \quad \text{for} \quad j \ne k.$$

- We call this a *Fourier series representation*.
- It makes stuff looks a bit like what you'd see in intro classes.
- We can think of the *higher order terms* — ϕ_j where λ_j is large — much like we thought about quadratic terms, interactions, etc., in linear regression.

Disadvantages

1. It's a bit harder to understand intuitively.
    - I can see from a drawing whether a curve is increasing and whether its derivative is.
    - Or whether it has has small Lipschitz or TV seminorm.
    - With this model, I may have a rough sense, but it's not as easy.
2. Maybe it's not quite what we want.
    - Maybe we know we want a Lipschitz model, e.g. if we're doing RDD.
    - We'd want to ensure it doesn't do anything weird at the data's edge.

A review of orthogonal bases in $\mathbb{R}^n$

- A set of vectors $v_1 \dots v_n$ is a basis if we can write every vector in $\mathbb{R}^n$ as a *unique* weighted average of the vectors in the basis.

$$\text{for all } v \in \mathbb{R}^n \text{, there exists unique } \alpha \in \mathbb{R}^n \text{ such that } v = \sum_{k=1}^{n} \alpha_k v_k.$$

- A basis is *orthogonal* if all pairs of basis vectors have zero inner product.

$$\langle v_j, v_k \rangle = 0 \quad \text{for} \quad j \neq k.$$

- *Eigenvectors* of a symmetric matrix T are an orthogonal for *two inner products*
  1. The usual inner product, the dot product $\langle u, v \rangle_2$.
  2. An inner product involving T, $\langle u, v \rangle_T = \langle Tu, v \rangle_2$.

  And they form a basis for $\mathbb{R}^n$.

Proving orthogonality of eigenvectors

Orthogonality in the dot product $\langle \cdot, \cdot \rangle_2$

Orthogonality in the inner product $\langle \cdot, \cdot \rangle_T = \langle T \cdot, \cdot \rangle_2$

Proving orthogonality of eigenvectors

Orthogonality in the dot product $\langle \cdot, \cdot \rangle_2$

Let $v_1 \dots v_n$ be eigenvectors of symmetric T with distinct eigenvalues λ_j: $Tv_k = \lambda_k v_k$.

$$\lambda_j \langle v_j, v_k \rangle_2 = \underset{(Tv_j)^T v_k = v_j^T T^T v_k}{\langle Tv_j, v_k \rangle_2} = \underset{v_j^T (T^T v_k) = v_j^T (Tv_k)}{\langle v_j, Tv_k \rangle_2} = \lambda_k \langle v_j, v_k \rangle$$

Because $\lambda_j \neq \lambda_k$, this is true *only if* v_j, v_k are orthogonal in the dot product $\langle \cdot, \cdot \rangle_2$.

Orthogonality in the inner product $\langle \cdot, \cdot \rangle_T = \langle T \cdot, \cdot \rangle_2$

Orthogonality in the dot product $\langle \cdot, \cdot \rangle_2$

Let $v_1 \ldots v_n$ be eigenvectors of symmetric T with distinct eigenvalues λ_j: $Tv_k = \lambda_k v_k$.

$$\lambda_j \langle v_j, v_k \rangle_2 = \underbrace{\langle Tv_j, v_k \rangle_2}_{(Tv_j)^T v_k = v_j^T T^T v_k} = \underbrace{\langle v_j, Tv_k \rangle_2}_{v_j^T (T^T v_k) = v_j^T (Tv_k)} = \lambda_k \langle v_j, v_k \rangle$$

Because $\lambda_j \neq \lambda_k$, this is true *only if* v_j, v_k are orthogonal in the dot product $\langle \cdot, \cdot \rangle_2$.

Orthogonality in the inner product $\langle \cdot, \cdot \rangle_T = \langle T\cdot, \cdot \rangle_2$

$\langle Tv_j, v_k \rangle = \lambda_j \langle v_j, v_k \rangle_2 = 0$ because we have orthogonality in the dot product.

Orthogonal bases for square-integrable functions on $[0, 1]$

- A set of functions $v_1, v_2, \ldots$ is a basis if we can write every square-integrable function on $[0, 1]$ as a *unique* weighted average of the functions in the basis.

  for all $v : \int_0^1 v(x)^2\, dx < \infty$, there exists unique $\alpha_1, \alpha_2, \ldots$ such that $v = \sum_{k=1}^{\infty} \alpha_k v_k$.

- A basis is *orthogonal* if all pairs of basis functions have zero inner product.

$$\langle v_j, v_k \rangle = 0 \quad \text{for} \quad j \neq k.$$

- *Eigenvectors* of a 'symmetric matrix' T are orthogonal for *two inner products*
  1. The usual inner product, $\langle u, v \rangle_{L_2} = \int_0^1 u(x) v(x)\, dx$.
  2. An inner product involving T, $\langle u, v \rangle_T = \langle Tu, v \rangle_{L_2}$.

  And they form a basis, too. Here T is a symmetric matrix if $\langle Tu, v \rangle_{L_2} = \langle u, Tv \rangle_{L_2}$.

Technical Detail

By *a symmetric matrix*, I mean a compact self-adjoint operator.

Theorem (The Spectral Theorem)
Suppose T is a compact self-adjoint operator on a Hilbert space V. Then there is an orthogonal basis of V consisting of eigenvectors of T. Each eigenvalue is real.

A symmetric matrix of interest

It's convenient to think of our functions as 2-periodic functions of $x \in \mathbb{R}$.

- That is, functions with $u(x + 2k) = u(x)$ for $k \in \mathbb{Z}$.
- Since they're really functions on $[0, 1]$, we just define $u(x)$ this way for $x \notin [0, 1]$.
- And then $\langle u, v\rangle_{L_2} = \frac{1}{2}\int_{-1}^{1} u(x)v(x)\,dx = \frac{1}{2}\int_{-1}^{0} u(x)v(x)\,dx + \frac{1}{2}\int_{0}^{1} u(x)v(x)\,dx$.

This isn't anything meaningful—it's all just a trick to simplify notation.

For periodic functions, we can express a first-order Sobolev derivative constraint in terms of the second derivative. We use integration by parts.

A symmetric matrix of interest

It's convenient to think of our functions as 2-periodic functions of $x \in \mathbb{R}$.

- That is, functions with $u(x+2k) = u(x)$ for $k \in \mathbb{Z}$.
- Since they're really functions on $[0,1]$, we just define $u(x)$ this way for $x \notin [0,1]$.
- And then $\langle u, v \rangle_{L_2} = \frac{1}{2}\int_{-1}^{1} u(x)v(x)\,dx = \frac{1}{2}\int_{-1}^{0} u(x)v(x)\,dx + \frac{1}{2}\int_{0}^{1} u(x)v(x)\,dx.$

This isn't anything meaningful—it's all just a trick to simplify notation.

For periodic functions, we can express a first-order Sobolev derivative constraint in terms of the second derivative. We use integration by parts.

$$
\begin{aligned}
\int_{-1}^{1} m'(x)^2\,dx &= \int_{-1}^{1} u(x)v'(x) && \text{for } u = m',\ v = m \\
&= u(x)v(x) \mid_{-1}^{1} - \int_{-1}^{1} u'(x)v(x) && \text{integrating by parts} \\
&= 0 - \int_{-1}^{1} m''(x)m(x) && \text{substituting and using periodicity} \\
&= 2\langle \text{-}\Delta\, m, m \rangle_{L_2} && \text{where } \text{-}\Delta\, u = -u''
\end{aligned}
$$

The negated second derivative operator

We can show the second derivative operator -$\Delta\, u = -u''$ is a self-adjoint operator.

The negated second derivative operator

We can show the second derivative operator -$\Delta\, u = -u''$ is a self-adjoint operator.

$$
\begin{aligned}
-2\langle u, \text{-}\Delta\, v\rangle_{L_2} &= \int_{-1}^{1} u(x)v''(x)\,dx \\
&= u(x)v'(x)\mid_{-1}^{1} - \left(u'(x)v'(x)\mid_{-1}^{1} - \int_{-1}^{1} u''(x)v(x)\,dx \right) \\
&= \int_{-1}^{1} u''(x)v(x)\,dx = -2\langle \text{-}\Delta\, u, v\rangle_{L_2}.
\end{aligned}
$$

The negated second derivative operator

We can show the second derivative operator $\text{-}\Delta\, u = -u''$ is a self-adjoint operator.

$$
\begin{aligned}
-2\langle u, \text{-}\Delta\, v\rangle_{L_2} &= \int_{-1}^{1} u(x)v''(x)\,dx \\
&= u(x)v'(x)\,|_{-1}^{1} - \left(u'(x)v'(x)\,|_{-1}^{1} - \int_{-1}^{1} u''(x)v(x)\,dx \right) \\
&= \int_{-1}^{1} u''(x)v(x)\,dx = -2\langle \text{-}\Delta\, u, v\rangle_{L_2}.
\end{aligned}
$$

Implications

- This means the *eigenvectors* of $\text{-}\Delta$ are an orthogonal basis for our space of periodic functions.
- And they're orthogonal in the sense of the usual inner product and the inner product of derivatives.

$$\langle \text{-}\Delta\, u, v\rangle_{L_2} = \langle u', v'\rangle_{L_2}.$$

The Sobolev model and the Negated Second Derivative Operator

We can characterize our model very simply in terms of these eigenvectors.

$$\mathcal{M} = \{m : \rho_{\text{-}\Delta}(m) \leq B\} \quad \text{for} \quad \rho(m) = \int_0^1 m'(x) = \dots$$

The Sobolev model and the Negated Second Derivative Operator

We can characterize our model very simply in terms of these eigenvectors.

$$\mathcal{M} = \{m : \rho_{\text{-}\Delta}(m) \leq B\} \quad \text{for} \quad \rho(m) = \int_0^1 m'(x) = \dots$$

$$\begin{aligned}
\int_0^1 m'(x)^2\,dx &= \langle m', m' \rangle_{L_2} = \langle \text{-}\Delta\, m, m \rangle \\
&= \langle \text{-}\Delta \sum_j m_j \phi_j, \sum_j m_j \phi_j \rangle \qquad \text{for the series expansion } m(x) = \sum_j m_j \phi_j(x) \\
&= \langle \sum_j m_j \lambda_j \phi_j, \sum_j m_j \phi_j \rangle \qquad \text{because } \phi_j \text{ is an eigenvector} \\
&= \sum_j m_j^2 \lambda_j
\end{aligned}$$

The sobolev model is the set of linear combinations of eigenvectors with coefficients in an infinite-dimensional *ellipse*. That's almost something we can implement.

Eigenvector and Eigenvalues

- We've expressed this model in terms of the *eigenvectors* $\phi_1, \phi_2, \ldots$ and eigenvalues $\lambda_1, \lambda_2, \ldots$ of the second derivative operator -Δ.

  i.e. the set of solutions (λ_k, ϕ_k) to $-\phi''(x) = \lambda\phi(x)$ for 2-periodic ϕ.

  What are they?

Eigenvector and Eigenvalues

- We've expressed this model in terms of the *eigenvectors* $\phi_1, \phi_2, \ldots$ and eigenvalues $\lambda_1, \lambda_2, \ldots$ of the second derivative operator -Δ.

i.e. the set of solutions (λ_k, ϕ_k) to $-\phi''(x) = \lambda\phi(x)$ for 2-periodic ϕ.

What are they?

They're sines and cosines

$$\phi_{2k}(x) = \sqrt{2}\sin(\pi k x) \quad \text{and} \quad \phi_{2k+1}(x) = \sqrt{2}\cos(\pi k x)$$

with eigenvalues $\lambda_{2k} = \lambda_{2k+1} = \pi^2 k^2$ for $k = 0, 1, 2, \ldots$

$$-\Delta\,\phi_{2k+1}(x) = -\{\sqrt{2}\cos(\pi k x)\}'' = \pi k \cdot \sqrt{2}\sin(\pi k x)' = \pi^2 k^2 \sqrt{2}\cos(\pi k x) = \pi^2 k^2 \phi_{2k+1}(x).$$

- We get them for integers k and not all k because the others aren't 2-periodic.
- We scale by a factor of $\sqrt{2}$ so they have length one, i.e., so $\langle \phi_k, \phi_k \rangle_{L_2} = 1$.

Sobolev Models and Fourier Series

Given a function on $[0,1]$, we can express it as a fourier series.

$$m(x) = \sum_{j=0}^{\infty} m_j \phi_j(x)$$

And we can express the relevant seminorm in terms of the coefficients of that series.

$$\rho_{\text{-}\Delta}(m) = \langle \text{-}\Delta\, m, m \rangle_{L_2} = \left\langle \sum_j \lambda_j m_j \phi_j, \sum_k m_k \phi_k \right\rangle_{L_2} = \sum_j \lambda_j m_j^2 \langle \phi_j, \phi_j \rangle_{L_2}.$$

This is simple because are basis functions ϕ_j are *orthogonal with length one*.

- The cross terms contribute nothing: $\langle \phi_j, \phi_k \rangle_{L_2} = 0$ for $j \neq k$.
- The factor $\langle \phi_j, \phi_j \rangle_{L_2}$ in the diagonal terms is just 1.

Summary: we can write our model in terms of fourier coefficients.

$$\mathcal{M} = \{m : \rho_{\text{-}\Delta}(m) \leq 1\} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) : \sum_{j=0}^{\infty} \lambda_j m_j^2 \leq 1 \right\}.$$

Smoother Sobolev Models and Fourier Series

We can do the same for Sobolev models defined in terms of higher order derivatives.

$$\mathcal{M}^p = \left\{ m : \int_0^1 m^{(p)}(x)^2 \, dx \le 1 \right\} = \left\{ \sum_{j=0}^{\infty} m_j \phi_j(x) : \sum_{j=0}^{\infty} \lambda_j^p m_j^2 \le 1 \right\}.$$

All that changes is the *power* of the eigenvalues λ_j.

Why?

The relevant seminorm involves the pth power of the second derivative operator.

$$\int_0^1 m^{(p)}(x)^2 \, dx = \langle \text{-}\Delta \underbrace{\cdots}_{s \text{ times}} \text{-}\Delta \, m, m \rangle \quad \text{via integration by parts}$$

And the pth power of any matrix T has

- The same eigenvectors ϕ_j as the matrix T
- Powered-up versions λ_j^p of the eigenvalues of T.