MATH 462 LECTURE NOTES: ORTHOGONAL FUNCTION REGRESSION

ADAM M. OBERMAN

Reference: Approximation theory Burden Faires Numerical Analysis 9th Burden Faires, Chapter 8. [Bur11].

1. Regression with full density

Previously, we looked at regression with a vector of data. If we knew the full function values, we can do a continuous version.

To better understand the functional point of view, let's do some examples. The easiest functions to work with are polynomials, and trignometric functions, since we can easily compute the intergrals.

$$L(h, \rho) = \mathbb{E}\left[(h(x) - y)^2\right]$$

this helps us to understand generalization. If we think of the inner product as an approximation to

$$(g, h)_\rho = \int_X g(x)h(x)\rho(x)$$

Then we are comparing the empirical inner product with the population inner product.

2. Functional analysis

Here we do a little introduction to thinking about features as functions. Then the normal equation involves the inner products of each feature component.

2.1. **Inner product of functions.** Let $X = [a, b]$ be an interval. Given a non-negative weight function $p(x)$

Define the vector space $V = V(X)$ of functions

$$V = \{f : X \to \mathbb{R}\}$$

with the inner product on V given by

$$(h, g) = (h, g)_\rho = \mathbb{E}[hg] = \int_a^b g(x)h(x)p(x)dx$$

Along with the norm

$$\|h\|_p$$

Given d functions, $f_1, \dots, f_d$, define

$$V_0 = \operatorname{span}\{f_1, \dots f_d\}$$

Using the definition of the inner product,

Date: November 16, 2023.

Definition 2.1. Given the features, f_i and a function $y \in V$, define

$$C_{ij} = (f_i, f_j), \qquad b_i = (f_i, y)$$

Let w be the solution of $Cw = b$, and define $h \in V_0$ by

$$h(x) = w \cdot f(x) = (C^{-1}b)^\top f(x)$$

2.2. **Derivation of the normal equations.** Let

$$h = Proj_{V_0}(y) = \arg\min_{f \in V} \|f - y\|_V^2$$

Define

$$e = y - h$$

to be the error (or residual).

Without going into the details about projections, which can be found in other references, we know the following:

$$(e, f) = 0, \quad \forall f \in V_0$$

the error is orthogonal to the space.

In particular, since h is the projection, for each basis element $f_i \in V$, we have

$$(e, f_i) = 0, \quad \forall i$$

So

$$(h, f_i) = (y, f_i)$$

Write $h = \sum w_j f_j$. Then

$$\left(\sum w_i f_i, f_j\right) = (y, f_j), \quad \forall j$$

or

$$\sum_i w_i (f_i, f_j) = (y, f_j)$$

Using the definition above, $C_{ij} = (f^i, f^j)$ and $b_j = (y, f_j)$.

Thus, using the orthogonality of projection, we recover the normal equation

$$Cw = b$$

2.3. **Orthogonal Functions.** To discuss general function approximation requires the introduction of the notions of weight functions and orthogonality.

Definition 2.2. An integrable function w is called a weight function on the interval I if $w(x) \geq 0$, for all x in I, but $w(x) \not\equiv 0$ on any subinterval of I.

The purpose of a weight function is to assign varying degrees of importance to approximations on certain portions of the interval.

Suppose $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is a set of linearly independent functions on $[a, b]$ and w is a weight function for $[a, b]$. Given $f \in C[a, b]$, we seek a linear combination

$$P(x) = \sum_{k=0}^{n} a_k \phi_k(x)$$

to minimize the error

$$E = E(a_0, \ldots, a_n) = \int_a^b w(x) \left[f(x) - \sum_{k=0}^{n} a_k \phi_k(x)\right]^2 dx.$$

The normal equations associated with this problem are derived from the fact that for each $j = 0, 1, \ldots, n$,

$$0 = \frac{\partial E}{\partial a_j} = 2\int_a^b w(x)\left[f(x) - \sum_{k=0}^{n} a_k\phi_k(x)\right]\phi_j(x)dx.$$

The system of normal equations can be written

$$\int_a^b w(x)f(x)\phi_j(x)dx = \sum_{k=0}^{n} a_k \int_a^b w(x)\phi_k(x)\phi_j(x)dx, \quad \text{for } j = 0, 1, \ldots, n.$$

If the functions $\phi_0, \phi_1, \ldots, \phi_n$ can be chosen so that

$$\int_a^b w(x)\phi_k(x)\phi_j(x)dx = \begin{cases} 0, & \text{when } j \neq k \\ \alpha_j > 0, & \text{when } j = k \end{cases}$$

then the normal equations will reduce to

$$\int_a^b w(x)f(x)\phi_j(x)dx = a_j \int_a^b w(x)\left[\phi_j(x)\right]^2 dx = a_j\alpha_j,$$

for each $j = 0, 1, \ldots, n$. These are easily solved to give

$$a_j = \frac{1}{\alpha_j}\int_a^b w(x)f(x)\phi_j(x)dx.$$

The word orthogonal means right-angled. So in a sense, orthogonal functions are perpendicular to one another.

Definition 2.3. $\{\phi_0, \phi_1, \ldots, \phi_n\}$ is said to be an orthogonal set of functions for the interval $[a, b]$ with respect to the weight function w if

$$\int_a^b w(x)\phi_k(x)\phi_j(x)dx = \begin{cases} 0, & \text{when } j \neq k, \\ \alpha_j > 0, & \text{when } j = k. \end{cases}$$

If, in addition, $\alpha_j = 1$ for each $j = 0, 1, \ldots, n$, the set is said to be orthonormal.

This definition, together with the remarks preceding it, produces the following theorem.

Theorem 2.4. *If $\{\phi_0, \ldots, \phi_n\}$ is an orthogonal set of functions on an interval $[a, b]$ with respect to the weight function w, then the least squares approximation to f on $[a, b]$ with respect to w is*

$$P(x) = \sum_{j=0}^{n} a_j\phi_j(x)$$

where, for each $j = 0, 1, \ldots, n$,

$$a_j = \frac{\int_a^b w(x)\phi_j(x)f(x)dx}{\int_a^b w(x)\left[\phi_j(x)\right]^2 dx} = \frac{1}{\alpha_j}\int_a^b w(x)\phi_j(x)f(x)dx.$$

There are broad classes of orthogonal functions. We begin with orthogonal sets of polynomials. The next theorem, which is based on the Gram-Schmidt process, describes how to construct orthogonal polynomials on $[a, b]$ with respect to a weight function w.

Theorem 2.5. *The set of polynomial functions $\{\phi_0, \phi_1, \ldots, \phi_n\}$ defined in the following way is orthogonal on $[a, b]$ with respect to the weight function w.*

$$\phi_0(x) \equiv 1, \quad \phi_1(x) = x - B_1, \quad \text{for each } x \text{ in } [a, b],$$

where

$$B_1 = \frac{\int_a^b x w(x) \left[\phi_0(x)\right]^2 dx}{\int_a^b w(x) \left[\phi_0(x)\right]^2 dx},$$

and when $k \geq 2$,

$$\phi_k(x) = (x - B_k)\,\phi_{k-1}(x) - C_k \phi_{k-2}(x), \quad \text{for each } x \text{ in } [a, b],$$

where

$$B_k = \frac{\int_a^b x w(x) \left[\phi_{k-1}(x)\right]^2 dx}{\int_a^b w(x) \left[\phi_{k-1}(x)\right]^2 dx}$$

and

$$C_k = \frac{\int_a^b x w(x) \phi_{k-1}(x) \phi_{k-2}(x) dx}{\int_a^b w(x) \left[\phi_{k-2}(x)\right]^2 dx}$$

Theorem above provides a recursive procedure for constructing a set of orthogonal polynomials. The proof of this theorem follows by applying mathematical induction to the degree of the polynomial $\phi_n(x)$.

2.4. **Classification and orthogonal functions.** Now consider a classification problem. Let X be the data domain. Let $\rho(x)$ be the distribution of data. Let $y_1(x), \ldots, y_k(x)$ be the indicator functions of the class labels and let c_j be the normalized versions, so that

$$c_j(x) = \begin{cases} \frac{1}{\sqrt{K}} & y_j(x) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume that each element $x \in X$ is a member of exactly one class.

Theorem 2.6. *The class label functions, y_j are orthogonal with respect to the weight function $\rho(x)$, as are c_j,*

$$(y_i, y_j)_\rho = 0, \quad i \neq j$$

If the classes are balanced, then the normalized label functions are orthonormal,

$$(y_i, y_i)_\rho = \frac{1}{K}, \qquad \|c_i\|_\rho = 1$$

Proof. No element has two different class labels, so if $i \neq j$, we have $y_j(x) y_i(x) = 0$, this means $(y_j, y_i)_\rho 0$. If there are equally balanced classes then,

$$\int y_i(x) \rho(x) = 1/K, \quad i = 1, \ldots, K$$

this means that $\|c_i\|_\rho^2 = \int c_i^2(x)\rho(x) = \frac{1}{K} \int y_i \rho(x) = 1$ as claimed. ∎

3. Orthogonal Polynomials and Least Squares Approximation

Here we consider polynomial regression, with $d+1$ features, which corresponds to

$$f(x) = (f_1(x) \dots, f_{d+1}(x)) = (1, x, \dots, x^d)$$

along with

$$h(x) = f(x)^\top w = w_0 + w_1 x + \dots w_d x^d$$

3.1. **definitions.**

The previous section considered the problem of least squares approximation to fit a collection of data. The other approximation problem mentioned in the introduction concerns the approximation of functions.

Suppose $y = \in C[a,b]$ (meaning it is a continuous function defined on the interval $[a,b]$, and that a polynomial $h(x) = h_n(x)$ of degree at most n is required that will minimize the error

$$\int_a^b [y(x) - h(x)]^2 \, dx$$

To determine a least squares approximating polynomial; that is, a polynomial to minimize this expression, let

$$\begin{aligned} h(x) &= w \cdot f(x) && \text{ML Notation} \\ &= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = \sum_{k=0}^n a_k x^k && \text{math notation} \end{aligned}$$

Remark 3.1. We are using two notations here: (i) ML notation to make it consistent with the rest of the ML presentation, (ii) math notation to make it consistent with the reference [Bur11].

Define

$$\begin{aligned} L(w) &= \int_a^b (y(x) - w \cdot f(x))^2 \, dx && \text{ML Notation} \\ E(a) &= \int_a^b \left(y(x) - \sum_{k=0}^n a_k x^k \right)^2 dx && \text{math notation} \end{aligned}$$

3.2. **Calculus derivation of the minimizer.**

The problem is to find real coefficients $a_0, a_1, \dots, a_n$ that will minimize E. A necessary condition for the numbers $a_0, a_1, \dots, a_n$ to minimize E is that

$$\frac{\partial E}{\partial a_j} = 0, \quad \text{for each } j = 0, 1, \dots, n$$

Since

$$E = \int_a^b [f(x)]^2 dx - 2 \sum_{k=0}^n a_k \int_a^b x^k f(x) dx + \int_a^b \left(\sum_{k=0}^n a_k x^k \right)^2 dx,$$

we have

$$\frac{\partial E}{\partial a_j} = -2 \int_a^b x^j f(x) dx + 2 \sum_{k=0}^n a_k \int_a^b x^{j+k} dx.$$

Hence, to find $P_n(x)$, the $(n+1)$ linear normal equations

$$\sum_{k=0}^n a_k \int_a^b x^{j+k} dx = \int_a^b x^j f(x) dx, \quad \text{for each } j = 0, 1, \dots, n,$$

must be solved for the $(n+1)$ unknowns a_j. The normal equations always have a unique solution provided that $f \in C[a,b]$. (See Exercise)

Example 3.2. Find the least squares approximating polynomial of degree 2 for the function $f(x) = \sin \pi x$ on the interval $[0,1]$. Solution The normal equations for $P_2(x) = a_2x^2 + a_1x + a_0$ are

$$a_0 \int_0^1 1dx + a_1 \int_0^1 xdx + a_2 \int_0^1 x^2dx = \int_0^1 \sin \pi x dx,$$
$$a_0 \int_0^1 xdx + a_1 \int_0^1 x^2dx + a_2 \int_0^1 x^3dx = \int_0^1 x\sin \pi x dx,$$
$$a_0 \int_0^1 x^2dx + a_1 \int_0^1 x^3dx + a_2 \int_0^1 x^4dx = \int_0^1 x^2\sin \pi x dx.$$

Performing the integration yields

$$a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 = \frac{2}{\pi}, \quad \frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 = \frac{1}{\pi}, \quad \frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 = \frac{\pi^2-4}{\pi^3}.$$

These three equations in three unknowns can be solved to obtain

$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx -0.050465 \text{ and } a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx 4.12251.$$

Consequently, the least squares polynomial approximation of degree 2 for $f(x) = \sin \pi x$ on $[0,1]$ is $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$.

The example illustrates a difficulty in obtaining a least squares polynomial approximation. An $(n+1)\times(n+1)$ linear system for the unknowns $a_0, \dots, a_n$ must be solved, and the coefficients in the linear system are of the form

$$(f_j(x), f_k(x))_\rho = (x^j, x^k)_\rho = \int_a^b x^{j+k}dx = \frac{b^{j+k+1} - a^{j+k+1}}{j+k+1},$$

This leads to a linear system. The matrix in the linear system is known as a Hilbert matrix.

3.3. **Legendre Polynomials.** The set of Legendre polynomials, $\{P_n(x)\}$, is orthogonal on $[-1,1]$ with respect to the weight function $w(x) \equiv 1$. The classical definition of the Legendre polynomials requires that $P_n(1) = 1$ for each n, and a recursive relation is used to generate the polynomials when $n \geq 2$. This normalization will not be needed in our discussion, and the least squares approximating polynomials generated in either case are essentially the same.

Using the Gram-Schmidt process with $P_0(x) \equiv 1$ gives

$$B_1 = \frac{\int_{-1}^1 xdx}{\int_{-1}^1 dx} = 0 \quad \text{and} \quad P_1(x) = (x - B_1)\, P_0(x) = x$$

Also,

$$B_2 = \frac{\int_{-1}^1 x^3dx}{\int_{-1}^1 x^2dx} = 0 \quad \text{and} \quad C_2 = \frac{\int_{-1}^1 x^2dx}{\int_{-1}^1 1dx} = \frac{1}{3},$$

so

$$P_2(x) = (x - B_2)\, P_1(x) - C_2P_0(x) = (x-0)x - \frac{1}{3}\cdot 1 = x^2 - \frac{1}{3}.$$

The higher-degree Legendre polynomials shown in Figure 8.9 are derived in the same manner. Although the integration can be tedious, it is not difficult with a Computer Algebra System.

4. Trigonometric polynomial approximation

The first observation in the development of Fourier series is that, for each positive integer n, the set of functions $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$, where

$$\phi_0(x) = \frac{1}{2},$$
$$\phi_k(x) = \cos kx, \quad \text{for each } k = 1, 2, \ldots, n,$$

and

$$\phi_{n+k}(x) = \sin kx, \quad \text{for each } k = 1, 2, \ldots, n-1,$$

is an orthogonal set on $[-\pi, \pi]$ with respect to $w(x) \equiv 1$. This orthogonality follows from the fact that for every integer j, the integrals of $\sin jx$ and $\cos jx$ over $[-\pi, \pi]$ are 0 , and we can rewrite products of sine and cosine functions as sums by using the three trigonometric identities

$$\sin t_1 \sin t_2 = \frac{1}{2}\left[\cos(t_1 - t_2) - \cos(t_1 + t_2)\right],$$
$$\cos t_1 \cos t_2 = \frac{1}{2}\left[\cos(t_1 - t_2) + \cos(t_1 + t_2)\right],$$
$$\sin t_1 \cos t_2 = \frac{1}{2}\left[\sin(t_1 - t_2) + \sin(t_1 + t_2)\right].$$

4.1. **Orthogonal Trigonometric Polynomials.**

Let $\mathcal{T}_n$ denote the set of all linear combinations of the functions $\phi_0, \phi_1, \ldots, \phi_{2n-1}$. This set is called the set of trigonometric polynomials of degree less than or equal to n.

For a function $f \in C[-\pi, \pi]$, we want to find the continuous least squares approximation by functions in $\mathcal{T}_n$ in the form

$$S_n(x) = \frac{a_0}{2} + a_n \cos nx + \sum_{k=1}^{n-1} \left(a_k \cos kx + b_k \sin kx\right).$$

Since the set of functions $\{\phi_0, \phi_1, \ldots, \phi_{2n-1}\}$ is orthogonal on $[-\pi, \pi]$ with respect to $w(x) \equiv 1$, it follows from Theorem that the appropriate selection of coefficients is

$$a_k = \frac{\int_{-\pi}^{\pi} f(x) \cos kx dx}{\int_{-\pi}^{\pi} (\cos kx)^2 dx} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, \quad \text{for each } k = 0, 1, 2, \ldots, n,$$

and

$$b_k = \frac{\int_{-\pi}^{\pi} f(x) \sin kx dx}{\int_{-\pi}^{\pi} (\sin kx)^2 dx} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx, \quad \text{for each } k = 1, 2, \ldots, n-1.$$

The limit of $S_n(x)$ when $n \to \infty$ is called the Fourier series of f. Fourier series are used to describe the solution of various ordinary and partial-differential equations that occur in physical situations.

Example 4.1. Determine the trigonometric polynomial from $\mathcal{T}_n$ that approximates

$$f(x) = |x|, \quad \text{for } -\pi < x < \pi.$$

Solution We first need to find the coefficients

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| dx = -\frac{1}{\pi}\int_{-\pi}^{0} x dx + \frac{1}{\pi}\int_{0}^{\pi} x dx = \frac{2}{\pi}\int_{0}^{\pi} x dx = \pi,$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \cos kx dx = \frac{2}{\pi}\int_{0}^{\pi} x \cos kx dx = \frac{2}{\pi k^2}\left[(-1)^k - 1\right],$$

for each $k = 1, 2, \ldots, n$, and

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} |x| \sin kx dx = 0, \qquad \text{for each } k = 1, 2, \ldots, n-1.$$

That the b_k 's are all 0 follows from the fact that $g(x) = |x| \sin kx$ is an odd function for each k, and the integral of a continuous odd function over an interval of the form $[-a, a]$ is 0 . (See Exercises 13 and 14.) The trigonometric polynomial from $\mathcal{T}_n$ approximating f is therefore,

$$S_n(x) = \frac{\pi}{2} + \frac{2}{\pi}\sum_{k=1}^{n} \frac{(-1)^k - 1}{k^2} \cos kx.$$

5. TODO: Discrete (data) least squares trig approximation

This section is good because it has discrete data.

References

[Bur11] Richard L Burden. *Numerical analysis*. Brooks/Cole Cengage Learning, 2011.