ANOVA decomposition of a Gaussian Process model

November 18, 2020

1 ANOVA docomposition of a general function

Functional ANOVA decomposition represents a high-dimensional function as a function of the form

$$f(x_1, x_2, \dots, x_D) = f_0 + \sum_{i=1}^{D} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \sum_{i<j<k} f_{ijk}(x_i, x_j, x_k) + \dots + f_{1,\dots,D}(x_1, \dots, x_D).$$

Let's first talk about how to compute the ANOVA components in general and then we focus on how to do it for $f(x_1, \dots, x_D)$ which is evaluated as a mean of a Gaussian process.

The first step is to choose the projection operator P:

$$Pf := \int_{[a,b]} f(x) d\mu(x)$$

We are going to use the projection operator P using Lebegue measure $Pf := \int_{[a,b]} f(x)dx$, so all intergrals should work out as expected.

1.1 Two-dimensional case

Now we use the projection operator to define the constant and the main effects. We assume $D = 2$ for now and then introduce some more notation to generalize:

$$f_0 = \int_{[a_1,b_1]} \int_{[a_2,b_2]} f(x_1, x_2) dx_1 dx_2 \tag{1}$$

$$f_1(x_1) = \int_{[a_2,b_2]} \Big(f(x_1, x_2) - f_0 \Big) dx_2 \tag{2}$$

$$f_2(x_2) = \int_{[a_1,b_1]} \Big(f(x_1, x_2) - f_0 \Big) dx_1 \tag{3}$$

The interaction effect $f_{1,2}(x_1, x_2)$ is defined as the remainder to make the ANOVA decomposition to work out correctly:

$$f_{1,2}(x_1, x_2) = f(x_1, x_2) - f_0 - f_1(x_1) - f_2(x_2).$$

The **total variance** (TV) of the predictor is defined as

$$\sigma^2(f) := \int_{[a_1,b_1]} \int_{[a_2,b_2]} (f(x_1,x_2) - f_0)^2 dx_1 dx_2$$

One can show that TV is decomposible into the sum of variances of main effects and interactions defined above:

$$\sigma_1^2(f_1) := \int_{[a_1,b_1]} (f_1(x_1))^2 dx_1 \tag{4}$$

$$\sigma_2^2(f_2) := \int_{[a_2,b_2]} (f_2(x_2))^2 dx_2 \tag{5}$$

$$\sigma_{1,2}^2(f_{1,2}) := \int_{[a_1,b_1]} \int_{[a_2,b_2]} (f_{1,2}(x_1,x_2) - f_0 - [f_1(x_1) - f_0] - [f_2(x_2) - f_0])^2 \, dx_1 dx_2 \tag{6}$$

$$\sigma^2(f) = \sigma_1^2(f_1) + \sigma_2^2(f_2) + \sigma_{1,2}^2(f_{1,2}). \tag{7}$$

And so, by dividing individual variances by TV we can express these compoents as percentages.

1.2 General case

Using subsets $u \subseteq \{1,\ldots,D\}$, we can establish a shorthand notation for ANOVA components, where f_u and x_u represents a subset of vector x with components $x_i, i \in u$. The we have

$$f(x_1,\ldots,x_D) = \sum_{u \subseteq \{1\ldots,D\}} f_u(x_u), \tag{8}$$

$$f_u(x_u) = \int_{\Omega^{D-|u|}} \Big(f(x) - \sum_{v \subsetneq u} f_v(x_v) \Big) dx_{-u} \tag{9}$$

$$\sigma^2(f_u) = \int_{\Omega^D} \Big(f_u(x_u) \Big)^2 dx \tag{10}$$

$$\sigma^2(f) = \int_{\Omega^D} \Big(f(x) - f_0 \Big)^2 dx = \sum_{u \subseteq \{1,\ldots,D\}, u \neq \emptyset} \sigma^2(f_u). \tag{11}$$

1.3 ANOVA for Gaussian process

For Bayesian optimization we are using the Gaussian proces defined by a parametrized RBF kernel of the form

$$K(x,x'|\theta_0,\vec{\theta}_1) = \theta_0^2 \exp\Big(-(x-x')^t D(\vec{\theta}_1)(x-x')) \Big),$$

where $[D(\vec{\theta}_1)]_{ij} = (\theta_{1,i}^2 + \epsilon)\delta_{ij}$ is a diagonal matrix of scaling coefficients, ϵ beeing a constant and δ_{ij} beeing Dirac-delta.

$$K(x,x'|\theta_0,\vec{\theta}_1) = \theta_0^2 \exp\Big(-\sum_{d=1}^{D} (\theta_{1d}^2 + \epsilon)(x_d - x_d')^2 \Big) = \theta_0^2 \prod_{d=1}^{D} \exp\Big(-(\theta_{1d}^2 + \epsilon)(x_d - x_d')^2 \Big).$$

I drop θ's from the definition of K for shorthand.

Let Σ be the correlation matrix of the training set $\{(x_i, y_i)\}_{i=1}^n$ and $k(x)$ a vector with correlations to the new point x, to evaluate the functional given by the GP and get the mean at the points x we do:

$$\Sigma := \{K(x_i, x_j)\}_{ij} \tag{12}$$

$$k(x) := \{K(x, x_i)\}_{i=1}^n \tag{13}$$

$$f(x) := k(x)^t \Sigma^{-1} y \tag{14}$$

$$= \sum_{i=1}^n K(x, x_i) \Sigma_i^{-1} y, \tag{15}$$

where Σ_i^{-1} is the i'th row of the inverse correlation matrix (of course, we are just solving $\Sigma z = y$ and take the i'th component of z, so I will just use z_i as a shorthand:

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

I'll use wolframalpha to derive integrands. We also define a shorthand

$$c_d := \sqrt{\theta_{1d}^2 + \epsilon}$$

For the constant we have:

$$f_0 = \int_{\Omega_D} f(x) dx = \int_{\Omega_D} \sum_{i=1}^n z_i K(x, x_i) dx \tag{16}$$

$$= \sum_{i=1}^n z_i \theta_0^2 \prod_{d=1}^D \int_{a_d}^{b_d} \exp\Big(-c_d^2 (x_d - x_{id})^2\Big) dx_d \tag{17}$$

$$= \theta_0^2 \sum_{i=1}^n z_i \prod_{d=1}^D \frac{\sqrt{\pi}\Big[\operatorname{erf}\Big(c_d(b_d - x_{id})\Big) - \operatorname{erf}\Big(c_d(a_d - x_{id})\Big)\Big]}{2c_d} \tag{18}$$

For the total variance we have:

$$\sigma^2(f) = \int_{\Omega_d} \Big(f(x) - f_0\Big)^2 dx = \int_{\Omega_d} \Big(\sum_{i=1}^{n} z_i K(x, x_i) - f_0\Big)^2 dx \tag{19}$$

$$= \int_{\Omega_d} \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j K(x, x_i) K(x, x_j) dx - 2 f_0 \int_{\Omega_d} \sum_{i=1}^{n} z_i K(x, x_i) dx + f_0^2 \int_{\Omega_d} dx \tag{20}$$

$$= f_0^2 \Big(\prod_{d=1}^{D} (b_d - a_d) - 2\Big) + \theta_0^4 \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \prod_{d=1}^{D} \int_{a_d}^{b_d} \exp\Big(-c_d^2 \Big[(x_d - x_{id})^2 + (x_d - x_{jd})^2\Big]\Big) dx_d \tag{21}$$

$$= f_0^2(\ldots) + \theta_0^4 \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \prod_{d=1}^{D} \frac{\sqrt{\frac{\pi}{2}} \exp\Big(-1/2 c_d^2 (x_{id} - x_{jd})^2\Big)\Big(\mathrm{erf}\Big[\frac{c_d(-2a_d + x_{id} + x_{jd})}{\sqrt{2}}\Big] - \mathrm{erf}\Big[\frac{c_d(-2b_d + x_{id} + x_{jd})}{\sqrt{2}}\Big]}{2c_d} \tag{22}$$

For the main effects we have:

$$f_t(x_t) = \theta_0^2 \sum_{i=1}^{n} z_i \exp\Big(-c_t^2 (x_t - x_{it})^2\Big) \prod_{d \neq t}^{D} \frac{\sqrt{\pi}\Big[\mathrm{erf}\Big(c_d(b_d - x_{id})\Big) - \mathrm{erf}\Big(c_d(a_d - x_{id})\Big)\Big]}{2c_d} - f_0 \prod_{d=1}^{D} (b_d - a_d) \tag{23}$$

$$=: \theta_0^2 \sum_{i=1}^{n} z_i \exp\Big(-c_t^2 (x_t - x_{it})^2\Big) P_i - F_0 \tag{24}$$

$$\sigma^2(f_t(x_t)) = \int_{\Omega^D} \Big(f_d(x_d)\Big)^2 dx = \int_{\Omega^D} \Big(\theta_0^2 \sum_{i=1}^{n} z_i \exp\Big(-c_t^2 (x_t - x_{it})^2\Big) P_i - F_0\Big)^2 dx \tag{25}$$

$$= \int_{\Omega^D} \theta_0^4 \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j P_i P_j \exp\Big(-c_t \Big[(x_t - x_{it})^2 + (x_t - x_{jt})^2\Big]\Big) \tag{26}$$

$$- 2\theta_0^2 F_0 \sum_{i=1}^{n} z_i P_i \int_{\Omega^D} \exp\Big(-c_t^2 (x_t - x_{it})^2\Big) dx + F_0^2 \int_{\Omega^D} dx \tag{27}$$

$$= \theta_0^4 \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j P_i P_j \frac{\sqrt{\frac{\pi}{2}} e^{-1/2 c_t^2 (x_{it} - x_{jt})^2} \Big[\mathrm{erf}\Big(\frac{c_t(-2a_t + x_{it} + x_{jt})}{\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{c_t(-2b_t + x_{it} + x_{jt})}{\sqrt{2}}\Big)\Big]}{2c_t} \prod_{d \neq t}^{D} (b_d - a_d) \tag{28}$$

$$- 2\theta_0^2 F_0 \sum_{i=1}^{n} z_i P_i \frac{\sqrt{\pi}\Big[\mathrm{erf}\Big(c_t(b_t - x_{it})\Big) - \mathrm{erf}\Big(c_t(a_t - x_{it})\Big)\Big]}{2c_t} \prod_{d \neq t}^{D} (b_d - a_d) + F_0^2 \prod_{d=1}^{D} (b_d - a_d) \tag{29}$$

Note:we can further simplify the computations if we normalize the domain Ω^D to a hypercube $[0,1]^D$.