Hashim

hafizhashimimtiaz786

January 2020

1 Determination of Covariance Functions

Covariance is a measure of relationship between random variables. It measures that how much two variables change together.
The covariance function $k(x, x')$ models the dependence between any function $f(x)$ values at different input points x and x':

$$k(x, x') = E((f(x) - m(x) - f(x') - m(x'))$$

A covariance function on set S is a function $k : S \times S \rightarrow R$ such that $\forall\ n \in N, \forall x_1, ..., x_n \in S$, the matrix $C = \begin{pmatrix} C_1 1 & C_1 2 & \cdots & C_1 n) \\ C_2 1 & C_2 2 & \cdots & C_2 n \\ \vdots & \vdots & \ddots & \vdots \\ C_n 1 & C_n 2 & \cdots & C_n n \end{pmatrix}, (C_{ij} = k(x_i, x_j))$
should be symmetric and positive semi definite, where $k(x_i, x_j)$ is the covariance function which will evaluate x_i and x_j.
So the covariance function $k(x_i, x_j))$ is valid if the resultant $n \times n$ matrix is positive definite for any n and any $x_1, ..., x_n$.
Conditions for the matrix to be positive semi definite are:

$$u^T C u > 0 \quad \forall\ u \in R^n$$

$$\forall_i\ \lambda_i > 0 \quad \rightarrow \textit{all eigen values are positive}$$

$$|C| > 0$$

Covariance functions should also be positive semi definite and fulfill some conditions to produce valid covariance matrix:

$$(k(x_i, x_j) = k(x_j, x_i)$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0 \qquad a_i a_j \in R, n \in N$$

It is often intractable to show from the definition that covariance function is

positive semi definite. So a common method is to use some already defined and proved kernels and make new one from them. Some kernels are shown below.

$$k(x_i, x_j) = \sigma_f^2 \exp -(\frac{||(x_i - x_j)||^2}{(2\lambda^2)}) \quad Squared\ Exponential\ Kernel$$

$$k(x_i, x_j) = \sigma_f^2 \exp -(\frac{r}{(\lambda)})\ where\ r = \sqrt{(x_i - x_j)^T(x_i - x_j)} \quad Exponential\ Kernel$$

$$k(x_i, x_j) = \sigma_f^2 \exp(1+\frac{\sqrt{3}r}{(\lambda)}) \exp -(\frac{\sqrt{3}r}{(\lambda)})\ where\ r = \sqrt{(x_i - x_j)^T(x_i - x_j)} \quad Matern3/2$$

The kernels that are a function of $(|x_i x_j|)$ are known as stationary kernels. And the kernels that are a function of distance $(||x_i x_j||)$ are known as isotropic kernels. There are many other defined kernels like the shown above which we can use and we can also make new one by using these already proved kernels. There are many operations we can apply to these covariance functions such that after modification positive semi definite property will retain. The common operations are:

- Multiplication by a scalar
- Sum of kernels
- Product of kernels
- Multiplication by a function
- Composition with a function

2 Optimizing hyper-parameters

In Gaussian process regression we need to optimize the hyper-parameters of chosen kernel function. We need to find the best values of hyper-parameters to maximize the marginal likelihood function. The covariance function usually contains hyper-parameters such as the length-scale (λ),signal variance σ_f, and noise variance σ_n, which are unknown and need to be predicted from the observed data. The common method to predict or compute the best hyper-parameters is by maximising the marginal likelihood or minimizing the negative of marginal likelihood. For convenience, we use θ to denote the vector of all hyper-parameters of the model. Let given the data $D = (X; y)$ and hyper-parameters θ such that $\theta \sim (\lambda, \sigma_f, \sigma_n)$.We will chose the parametric prior.

$$p(f|\theta)) = \mathcal{GP}(f; \mu(x; \theta), K(x, x'; \theta))$$

We will measure the quality of the fit to our training data $D = (X; y)$ with the

marginal likelihood. The equation for log marginal likelihood is shown below.

$$\log p(y|X) = -\frac{y^T(K(X,X)+\sigma_n^2 I)^{-1}y}{2} - \frac{\log|K(X,X)+\sigma_n^2 I|}{2} - \frac{n\log 2\pi}{2}$$

The marginal log likelihood can be seen as a penalized fit measure, where the first term $-\frac{y^T(K(X,X)+\sigma_n^2 I)^{-1}y}{2}$ measures the data fit. The second term $= -\frac{\log|K(X,X)+\sigma_n^2 I|}{2}$ is a complexity penalization term and the last one $-\frac{n\log 2\pi}{2}$ is a normalization constant. To find the appropriate parameters, we can maximize our log marginal likelihood function using gradient ascent. We can also minimize the negative of function using gradient descent. For maximizing the function, we take partial derivative of function with respect to each hyper-parameter, multiply it with a learning rate and add it to the previous value of parameter. For minimizing, we take the negative of function and take partial derivative of function with respect to each hyper-parameter, multiply it with a learning rate and subtract it from the previous value of parameter. We Keep doing this until we reach to the maximum for original marginal likelihood function or minimum for negative of function, means where derivative is zero and there is no much difference between previous and present value of the function at the specific value of parameter. We want to find this specific value of all the hypermeters. The notations for gradient acsent and gradient descent are shown below:

$$\theta_j \leftarrow \theta_j + \alpha\frac{\partial}{\partial\theta_j}J(\theta) \sim \textit{until converges}$$

$$\theta_j \leftarrow \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta) \sim \textit{until converges}$$