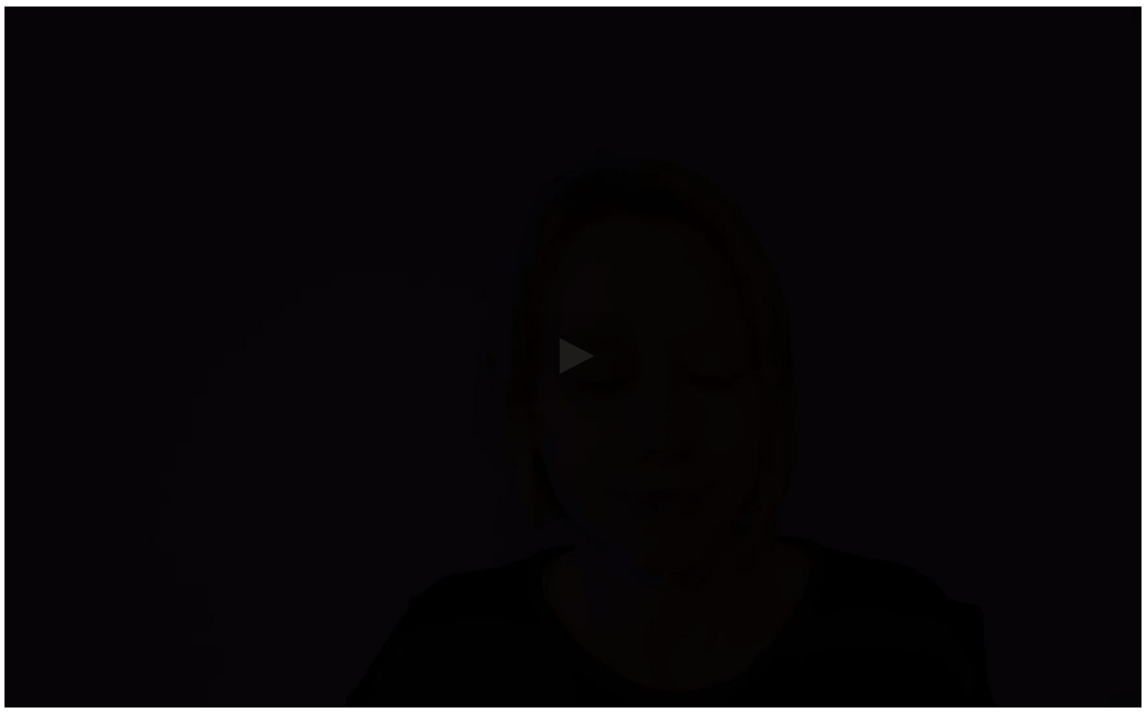
2. Model Selection

Bookmark this page

Model Selection



20:43 / 20:43 1.50x

to fit Gaussian processes on a variety of data

without even that much prior knowledge.

It's still good to know what these different kernels do,

so that you can already come up with a good set of candidate

kernels.

But other than that, you can actually

fit the rest to the data at hand that you have.

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Initially, let's recall our setup where we have a pair of multivariate Gaussian random variables $\mathbf{X}_1 \in \mathbb{R}^d$, and $\mathbf{X}_2 \in \mathbb{R}^{N-d}$. These two random variables are used to represent the temperature at two sets of cities: $\mathbf{X}_1$ are the cities for which we do no have temperature measurements, and $\mathbf{X}_2$ are the cities for which we do have temperate measurements. In addition, we also have access to the means of both of these random variables, which are denoted by μ_1 and μ_2 respectively — these will be the mean temperature at each of the cities.

The random variables are associated with physical locations represented by the variables $\mathbf{Z}_1 \in \mathbb{R}^{M \times d}$ and $\mathbf{Z}_2 \in \mathbb{R}^{M \times (N-d)}$ where we have assumed that we are working on an M-dimensional space; typically, $M = 2$ for spatial data. Further, we have selected a covariance function $k(z_i, z_j)$ that serves as proxy for the relation between two random variables as a function of their spatial locations. We use this kernel function to construct a covariance matrix so that $\Sigma_{ij} = cov(X_i, X_j) = k(z_i, z_j)$. Thus, we build the matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} \in \mathbb{R}^{d \times d} & \mathbf{\Sigma}_{12} \in \mathbb{R}^{d \times (N-d)} \\ \mathbf{\Sigma}_{21} \in \mathbb{R}^{(N-d) \times d} & \mathbf{\Sigma}_{22} \in \mathbb{R}^{(N-d) \times (N-d)} \end{bmatrix}$$

In the previous sections we have shown that the distribution of the random variable $\mathbf{X}_1$ conditioned on $\mathbf{X}_2 = \mathbf{x}_2$ is a Normal distribution with

$$\begin{aligned} \mu_{\mathbf{X}_1|\mathbf{X}_2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2) \\ \Sigma_{\mathbf{X}_1|\mathbf{X}_2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

The main running assumption in this process is to model the variables to be measured – like temperature – as a jointly Normally distributed random variable with correlations determined as a function of location through the kernel function $k(z_i, z_j)$. Once the means have been specified, we may predict the unobserved random variables by computing the marginal distributions conditioned on the observed variables.

Here, we study the fundamental question of how to select this kernel function. One could create a countable number of models, such as Gaussian processes with different kernels, or use the same kernel with a set of different parameters. However, from these sets of kernels, how do we specify and select which model for the kernel is best?

We present two possible approaches for such a problem of model selection. These are not an exhaustive exposition of the possible approaches but are useful for the problems at hand. The interested reader may wish to consult the literature of Gaussian processes and model selection for further approaches.

We will proceed by first constructing an additional abstraction: we will consider the parameters of the kernel function to be some generic value θ. That is, for example in the case of the kernel function

$$k(y_i, y_j) = \exp\left(-\frac{\|y_i - y_j\|^2}{2\ell^2}\right).$$

We can say that $\theta = \{\ell\}$, and our objective is to find the "best" θ in some particular sense that will be defined later.

The two approaches we will explore are:

- *Estimate Generalization error: cross-validation, leave-one-out, or k-fold.* This defines a "good model"' as one that predicts best data that we have not seen before, i.e., generalization. This approach corresponds to the classical tension between having a model that fits the data well, and at the same time, generalizes to unobserved data.
- *Maximize the log marginal likelihood of the data, $p(y|X, \theta)$ to θ.* Here we assume we have a probabilistic model, where we compute how likely the data is that we have seen, under the chosen model. Alternatively, in short, how well the model fits the data as measured by a normalized probability. This approach balances fitting power and the simplicity of the model.

Discussion

Hide Discussion

Topic: Module 5: Environmental Data and Gaussian Processes:Sensing and Analyzing global patterns of dependence / 2. Model Selection

Add a Post

Show all posts ⌄ by recent activity ⌄

There are no posts in this topic yet.

< Previous Next >

edX

About
Affiliates
edX for Business
Open edX
Careers
News

Legal

Terms of Service & Honor Code
Privacy Policy
Accessibility Policy
Trademark Policy
Sitemap

Connect

Blog
Contact Us
Help Center
Media Kit
Donate

GET IT ON Google play

Download on the App Store