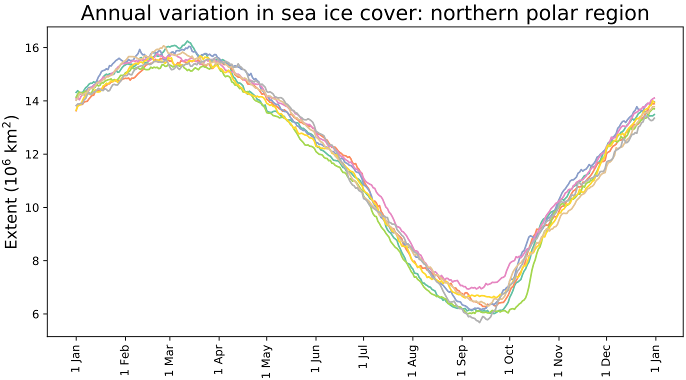
Statistics and Machine Learning 1

Lecture 7F: Gaussian Process Regression

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 7

How could we fit general functions?



The curves[1] above show data from the non-leap years in the 1990’s and prompt a natural question: how could we model them to make predictions and check for changes?

[1] The data came from Kaggle, https://www.kaggle.com, a company that runs machine-learning contests.

Gaussian Processes: functions from MVNs

- The smooth curves plotted in these slides are really just piecewise linear things, threading a long list of pairs

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_N, y_N)$$

- Imagine fixing the x_j and letting the y_j be random variables drawn from a multivariate normal distribution, then we'd have a way to generate *random curves*. This is the idea at the heart of Gaussian Process (GP) regression.
- We'd then need to specify a mean, $\mu \in \mathbb{R}^N$ and an $N \times N$ covariance matrix Σ for the vector

$$(y_1, y_2, \ldots, y_N).$$

This covariance matrix will turn out to provide a way to place constraints on the smoothness of the functions that our GP can produce.

Defining Σ with a kernel

We'll Σ using a *kernel*, a function $K(x_i, x_j)$ that measures how well-correlated we want the values of our functions to be at x_i and x_j. Typically it's a decreasing function of $|x_i - x_j|$, so that y values for nearby points tend to be highly correlated, while those for more distant ones are almost independent of each other.

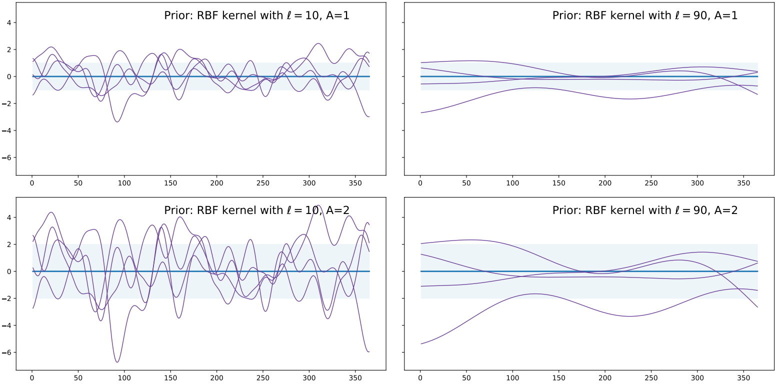
Today I'll introduce just one kind of covariance matrix, based on the RBF[2] or *squared-exponential* kernel. It depends on two parameters, an amplitude A and a length-scale ℓ and its entries take the form

$$\Sigma_{ij} = K_{RBF}(x_i, x_j) = A^2 \exp(-(x_i - x_j)^2/2\ell^2).$$

Once combined with a value for μ—and $\mu = 0$ is a common and reasonable choice—this function defines a multivariate normal that we can think of as a distribution over functions.

[2] This is short for *Radial Basis Function* and is used for historical reasons.

The influence of A and ℓ



Above: four examples of the family of functions generated by the RBF kernel. All panels show: the mean (blue, solid) of the distribution over functions, a region (light blue, shaded) that extends $\pm A$ (which is one standard deviation) above and below the mean, and a few sampled curves (purple).

GP regression is a Bayesian update

We can think of the distribution sketched on the previous slides as a Bayesian prior over functions. If we then get some data, say the value of $y_j^\star$ at some of our points x_j, then we can use them to update our beliefs about the possible values of the function.

- An MVN prior, combined with an MVN likelihood, leads to an MVN posterior, so we'll still be working with a Gaussian Process even after we've incorporated data.
- Once we have some data, we split the y-values into two groups, the ones where we have values—call these $\mathbf{y}_a$—and the ones about which we are still uncertain, $\mathbf{y}_b$. The calculations needed to make predictions amount to computing $f(\mathbf{y}_b \mid \mathbf{y}_a = \mathbf{y}_a^\star)$, which we saw how to do earlier in the lecture.

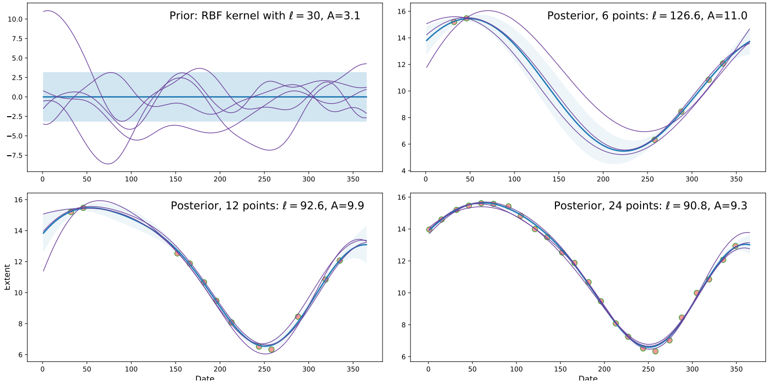
GP regression with noisy data

- If the data we're folding into our regression has some error on it, if $y_j^\star \sim \mathcal{N}(\mu_j^\star, \sigma_j^2)$, then we can simply modify the entries in the covariance matrix to be

$$\Sigma_{ij} = \begin{cases} A^2 \exp(-(x_i - x_j)^2/2\ell^2) & \text{if } i \neq j \\ A^2 + \sigma_j^2 & \text{if } i = j \end{cases}$$

- GP regression as I've described it is not entirely Bayesian. The issue is that need some way to estimate/optimise the parameters A and ℓ that we started with. This is usually done by maximising the posterior likelihood of the data.

How can we fit arbitrary functions?



Above: convergence of a GP regression for the arctic ice data. All panels show: the mean (blue, solid) of the distribution over functions, a region (light blue, shaded) that extends $\pm A$ (which is one standard deviation) above and below the mean, and a few sampled curves (brown).

Further reading

- This section of today's lecture owes a great deal to Chapter 8 of:

  S. Rogers and M. Girolami (2017), *A First Course in Machine Learning*, 2nd edition, Chapman & Hall/CRC. ISBN: 978-1-4987-3848-4.

  Available `online` through the University Library.
- Last year's students found and liked a video about Gaussian Processes at `https://www.youtube.com/watch?v=BS4Wd5rwNwE`
- The standard theoretical reference about Gaussian Processes is

  C. E. Rasmussen and C. K. Williams (2006), *Gaussian Processes for Machine Learning*, MIT Press. ISBN: 9-780262-182539.