Bayesian Machine Learning - Lecture 3

Guido Sanguinetti

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
gsanguin@inf.ed.ac.uk

February 25, 2015

Today's lecture

1. Random functions and Bayesian regression
2. Gaussian Processes
3. Bayesian prediction with GPs

Bayesian regression revisited

- We saw yesterday that Bayesian linear regression places a (Gaussian) prior over the weights vector
- Hence, the regression line is a *random line*: the output prediction at any point is a Gaussian random variable
- This concept can be generalised: taking linear combinations of basis functions with (Gaussian) random coefficients leads to a (Gaussian) random function
- Since Bayesian estimation effectively regularises, we can avoid the pitfalls of overfitting in this way.

Random functions terminology

- A random function is an infinite collection of random variables indexed by the argument of the function
- A popular alternative name is a *stochastic process*
- When considering the random function evaluated at a (finite) set of points, we get a random vector
- The distribution of this random vector is called *finite dimensional marginal*

Important exercise

Let $\phi_1(x), \ldots, \phi_N(x)$ be a fixed set of functions, and let $f(x) = \sum w_i \phi_i(x)$. If $\mathbf{w} \sim \mathcal{N}(0, I)$, compute:

1. The single-point marginal distribution of $f(x)$
2. The two-point marginal distribution of $f(x_1), f(x_2)$

Solution (sketched, take notes)

- Obviously the distributions are Gaussians
- Obviously both distributions have mean zero
- To compute the (co)variance, take products and expectations and remember that $\langle w_i w_j \rangle = \delta_{ij}$
- Defining $\phi(x) = (\phi_1(x), \ldots, \phi_N(x))$, we get that

$$\langle f(x_i) f(x_j) \rangle = \phi(x_i)^T \phi(x_j)$$

The Gram matrix

- Generalising the exercise to more than two points, we get that *any* finite dimensional marginal of this process is multivariate Gaussian
- The covariance matrix of this function is given by evaluating a function of two variables at all possible pairs
- The function is defined by the set of basis functions

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

- The covariance matrix is often called *Gram matrix* and is (necessarily) symmetric and positive definite
- Bayesian prediction in regression then is essentially the same as computing conditionals for Gaussians (more later)

Main limitation of Bayesian regression

- Choice of basis functions inevitably impacts what can be predicted
- Suppose one wishes the basis functions to tend to zero as $x \to \infty$
- Then, necessarily, very large input values will have predicted outputs near zero with high confidence!
- Ideally, one would want a prior over functions which would have the same uncertainty everywhere

Stationary variance

- We have seen that the variance of a random combination of functions depends on space as $\sum \phi_i^2(x)$
- Given any compact set, (e.g. hypercube with centre in the origin), we can find a finite set of basis functions s.t. $\sum \phi_i^2(x) = const$ (partition of unity, e.g. triangulations or smoother alternatives)
- We can construct a sequence of such sets which covers the whole of $\mathbb{R}^D$ in the limit
- Therefore, we can construct a sequence of priors which all have constant prior variance across all space
- Covariances would still be computed by evaluating a Gram matrix (and need not be constant)

Function space view

- The argument before shows that we can put a prior over infinite-dimensional spaces of functions s.t. all finite dimensional marginals are multivariate Gaussian
- The constructive argument, often referred to as *weights space view*, is useful for intuition but impractical
- It does demonstrate the existence of truly infinite dimensional Gaussian processes
- Once we accept that Gaussian processes exist, we are better off proceeding along a more abstract line

GP definition

- A Gaussian Process (GP) is a stochastic process indexed by a continuous variable x s.t. all finite dimensional marginals are multivariate Gaussian
- A GP is uniquely defined by its *mean* and *covariance* functions, denoted by $\mu(x)$ and $k(x, x')$:

$$f \sim \mathcal{GP}(\mu, k) \leftrightarrow \mathbf{f} = (f(x_1), \ldots, f(x_N)) \sim \mathcal{N}(\boldsymbol{\mu}, K),$$

$$\boldsymbol{\mu} = (\mu(x_1), \ldots . \mu(x_N)), \quad K = (k(x_i, x_j))_{i,j}$$

- The covariance function must satisfy some conditions (Mercer's theorem), essentially it needs to evaluate to a symmetric positive definite function for all sets of input points

Covariance functions

- The covariance function encapsulates the basis functions used $\longrightarrow$ it determines the type of functions which can be sampled
- The radial basis functions (RBF or squared exponential) covariance function

$$k(x_i, x_j) = \alpha^2 \exp\left[-\frac{(x_i - x_j)^2}{\lambda^2}\right]$$

corresponds to Gaussian bumps basis functions and yields smooth bumpy samples

- The Ornstein-Uhlenbeck (OU) covariance

$$k(x_i, x_j) = \alpha^2 \exp\left[-\frac{|x_i - x_j|}{\lambda^2}\right]$$

yields rough paths which are nowhere differentiable

- Both RBF and OU are stationary and encode exponentially decaying correlations

Observing GPs

- In a regression case, we assume to have observed the function values at some input values with i.i.d. Gaussian noise with variance σ^2
- What is the effect of observation noise?
- Suppose we have a Gaussian vector $\mathbf{f} \sim \mathcal{N}(\mu, \Sigma)$, and observations $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$
- Exercise: compute the marginal distribution of $\mathbf{y}$

Predicting with GPs

- Suppose we have noisy observations $\mathbf{y}$ of a function value at inputs $\mathbf{x}$, and want to predict the value at a new input x_{new}
- The joint prior probability of function values at the observed and new input points is multivariate Gaussian
- By Bayes' theorem, we have

$$p(f_{new}|\mathbf{y}) \propto \int df(\mathbf{x}) p(f_{new}, f(\mathbf{x})) p(\mathbf{y}|f(\mathbf{x})) \qquad (1)$$

where $f(\mathbf{x})$ is the vector of *true* function values at the input points

- Exercise: compute the distribution $p(f(x_{new})|\mathbf{y})$
- You will need the partitioned inverse formula (see `http://en.wikipedia.org/wiki/Block_matrix_pseudoinverse`

Covariance parameters

- Covariance functions often depend on hyperparameters (e.g. the amplitude and lengthscale of the RBF covariance)
- These can be tuned by optimising the marginal likelihood (called type II maximum likelihood)

$$\mathcal{L} = \log \int df(\mathbf{x}) p(f(\mathbf{x})) p(\mathbf{y}|f(\mathbf{x})) = \log |(K+\sigma^2 I)| - \mathbf{y}^T (K+\sigma^2 I)^{-1} \mathbf{y}$$

- Usually gradient methods are used; Bayesian methods are complicated by the in general complex functional form (no conjugate prior)

Pitfalls of GP prediction

- Addition of a new observation *always* reduces uncertainty at all points $\longrightarrow$ vulnerable to outliers
- Optimisation of hyperparameters often tricky: works well if σ^2 is known, otherwise it can be seriously multimodal
- **MAIN PROBLEM: GP prediction relies on a matrix inversion which scales cubically with the number of points!**
- Sparsification methods have been proposed but in high dimension GP regression is likely to be tricky nevertheless