MLPR Tutorial Sheet 7

1. **Pre-processing for Bayesian linear regression and Gaussian processes:**

   We have a dataset of inputs and outputs $\{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^N$, describing N preparations of cells from some lab experiments. The output of interest, $y^{(n)}$, is the fraction of cells that are alive in preparation n. The first input feature of each preparation indicates whether the cells were created in lab `A`, `B`, or `C`. That is, $x_1^{(n)} \in \{\texttt{A}, \texttt{B}, \texttt{C}\}$. The other features are real numbers describing experimental conditions such as temperature and concentrations of chemicals and nutrients.

   a) Describe how you might represent the first input feature and the output when learning a regression model to predict the fraction of alive cells in future preparations from these labs. Explain your reasoning.

   b) Compare using the lab identity as an input to your regression (as you've discussed above), with two baseline approaches: i) Ignore the lab feature, treat the data from all labs as if they came from one lab; ii) Split the dataset into three parts one for lab `A`, one for `B`, and one for `C`. Then train three separate regression models.

      Discuss both simple linear regression and Gaussian process regression. Is it possible for these models, when given the lab identity as in a), to learn to emulate either or both of the two baselines?

   c) There's a debate in the lab about how to represent the other input features: log-temperature or temperature, and temperature in Fahrenheit, Celsius or Kelvin? Also whether to use log concentration or concentration as inputs to the regression. Discuss ways in which these issues could be resolved.

      Harder: there is a debate between two different representations of the output. Describe how this debate could be resolved.

2. **Gaussian processes with non-zero mean:**

   In the lectures we assumed that the prior over any vector of function values was zero mean: $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K)$. We focussed on the covariance or kernel function $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, which evaluates the K_{ij} elements of the covariance matrix (also called the 'Gram matrix').

   If we know in advance that the distribution of outputs should be centered around some other mean $\mathbf{m}$, we *could* put that into the model. Instead, we usually subtract the known mean $\mathbf{m}$ from the $\mathbf{y}$ data, and just use the zero mean model.

   Sometimes we don't really know the mean $\mathbf{m}$, but look at the data to estimate it. A fully Bayesian treatment puts a prior on $\mathbf{m}$ and, because it's an unknown, considers all possible values when making predictions. A flexible prior on the mean vector could be another Gaussian process(!). Our model for our noisy observations is now:

$$\begin{aligned} \mathbf{m} &\sim \mathcal{N}(\mathbf{0}, K_m), && K_m \text{ from kernel function } k_m, \\ \mathbf{f} &\sim \mathcal{N}(\mathbf{m}, K_f), && K_f \text{ from kernel function } k_f, \\ \mathbf{y} &\sim \mathcal{N}(\mathbf{f}, \sigma_n^2\mathbb{I}), && \text{noisy observations.} \end{aligned}$$

   Show that — despite our efforts — the function values $\mathbf{f}$ still come from a function drawn from a zero-mean Gaussian process (if we marginalize out $\mathbf{m}$). Identify the covariance function of the zero-mean process for f.

   Identify the mean's kernel function k_m for two restricted types of mean: 1) An unknown constant $m_i = b$, with $b \sim \mathcal{N}(0, \sigma_b^2)$. 2) An unknown linear trend: $m_i = m(\mathbf{x}^{(i)}) = \mathbf{w}^\top\mathbf{x}^{(i)} + b$, with Gaussian priors $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_w^2\mathbb{I})$, and $b \sim \mathcal{N}(0, \sigma_b^2)$.

Sketch three typical draws from a GP prior with kernel:

$$k(x^{(i)}, x^{(j)}) = 0.1^2 \exp\big(-(x^{(i)} - x^{(j)})^2/2\big) + 1.$$

Hints in footnote[1].

3. **Laplace approximation:**

   The Laplace approximation fits a Gaussian distribution to a distribution by matching the mode of the log density function and the second derivatives at that mode. See the w8b lecture notes for more pointers.

   Exercise 27.1 in MacKay's textbook (p342) is about inferring the parameter λ of a Poisson distribution based on an observed count r. The likelihood function and prior distribution for the parameter are:

   $$P(r \,|\, \lambda) = \exp(-\lambda)\frac{\lambda^r}{r!}, \qquad p(\lambda) \propto \frac{1}{\lambda}.$$

   Find the Laplace approximation to the posterior over λ given an observed count r.

   Now reparameterize the model in terms of $\ell = \log\lambda$. After performing the change of variables[2], the improper prior on $\log\lambda$ becomes uniform, that is $p(\ell)$ is constant. Find the Laplace approximation to the posterior over $\ell = \log\lambda$.

   Which version of the Laplace approximation is better? It may help to plot the true and approximate posteriors of λ and ℓ for different values of the integer count r.

[1] The covariances of two Gaussians add, so think about the two Gaussian processes that are being added to give this kernel. You can get the answer to this question by making a tiny tweak to the Gaussian process demo code provided with the class notes.

[2] Some review of how probability densities work: Conservation of probability mass means that: $p(\ell)\mathrm{d}\ell = p(\lambda)\mathrm{d}\lambda$ for small corresponding elements $\mathrm{d}\ell$ and $\mathrm{d}\lambda$. Dividing and taking limits: $p(\ell) = p(\lambda)|\mathrm{d}\lambda/\mathrm{d}\ell|$, evaluated at $\lambda = \exp(\ell)$. The size of the derivative $|\mathrm{d}\lambda/\mathrm{d}\ell|$ is referred to as a Jacobian term.