Advanced Probability Theory: Question Sheet

Bugs/queries to mosb@robots.ox.ac.uk
For all related material visit
at www.robots.ox.ac.uk/∼mosb/c24

MT 2014
Michael Osborne

1 Let's revisit the engine reliability data from lectures, where counts for the numbers of engines (of type either A or $\neg A$) that were either reliable (R) or not ($\neg R$) that may (N) or may not ($\neg A$) have included a new component are below.

	A			$\neg A$	
	R	$\neg R$		R	$\neg R$
N	180	120	N	20	80
$\neg N$	70	30	$\neg N$	90	210

Rather than treating frequencies as probabilities, as we did for this data in lectures, let's treat this data in a Bayesian fashion. Our task in this question is to select a model to predict reliability given engine type and whether it contains a new component (that is, we can assume that our observations are of reliability for known engine type and known component-containment). Take independent uniform priors over all parameters, as in lectures.

(a) Evaluate the log model evidence of the model that assumes that reliability is the result of a fair coin toss, independent of engine type and whether it has the new component. Draw a Bayes net for this model.

(b) Evaluate the log model evidence of the model that assumes that reliability is independent of engine type and whether it has the new component. Draw a Bayes net for this model.

(c) Evaluate the log model evidence of the model that assumes that reliability is fully dependent upon both engine type and whether it has the new component. Draw a Bayes net for this model.

(d) Explain the advantages and disadvantages of each of the three models. Which should we choose? Why?

2 (a) Show that the probability distribution over $X \in [1, 2, \ldots, N]$ that maximises the entropy has $P(X = i)$ constant for all i. Note that $\sum_i P(X = i) = 1$. Hint: you probably want to use the method of Lagrange multipliers.

(b) Prove that the mutual information is symmetric.

(c) Gibbs' inequality states that, for any two probability distributions $Q(z)$ and $R(z)$,

$$\sum_i R(z_i) \log \frac{R(z_i)}{Q(z_i)} \geq 0.$$

Using Gibbs' inequality, show that the mutual information is non-negative,

$$I[X; Y] \geq 0.$$

See MacKay (2003), Exercise 8.4, p140 for hints.

3 As in lectures, imagine you have 12 numbered balls, which are otherwise identical to the eye. All are the same weight, aside from one that is either heavier or lighter. Given scales that can compare the weight of any set of balls against any other set of balls, we aim to determine the odd ball.

Define X as the index of the odd ball.

(a) Compute the mutual information between X and the outcome of an experiment that weighs six of the balls against the remaining six.

(b) Compute the mutual information between X and the outcome of an experiment that weighs four of the balls against another four.

4 (a) Prove that the sum of two covariance functions is another valid covariance function.

(b) Take a prior for a function that is a weighted sum of evenly-spaced Gaussians, $f(x) = \sum_{i=1}^{N} w_i \mathcal{N}(x; \alpha\, i; \sigma^2)$, $x \in \mathbb{R}$. Assume identical, independent, zero-mean Gaussian priors for each w_i that have (constant) variances τ^2. Note that $f(x)$ is not, in general, a pdf: $f(x)$ may be negative, and its integral is unlikely to be one. Assuming that σ, τ and α are known, show that f is hence modelled with a Gaussian process, and derive its prior mean and covariance functions.

(c) What commonly used covariance function can you unearth as $N \Rightarrow \infty$ while holding αN and $\tau^2 \alpha$ constant? What can you conclude?

5 Consider the following data,

x	$y(x)$
1	1
2	2
3	3
4	3
5	2.

The Gamma distribution is defined as

$$\Gamma(x; k,\, \theta) = \frac{1}{(k-1)!\,\theta^k} x^{k-1} e^{-\frac{x}{\theta}}, \quad \text{for } k \in \mathbb{N}^+, \theta \in \mathbb{R}^+,$$

which has mean equal to $k\theta$ and variance equal to $k\theta^2$. Our goal is to model the function $y(x)$. Define a model $\mathcal{M} = 1$ as

$$p\big(\mathbf{y} \mid \mathbf{x},\, c,\, \mathcal{M} = 1\big) = \prod_{x=1}^{5} \Gamma\big(y(x); 1, e^c\big)$$
$$p(c \mid \mathcal{M} = 1) = \mathcal{N}(c; 0,\, 1).$$

(a) Describe the model by interpreting the two terms $p\big(\mathbf{y} \mid \mathbf{x},\, m,\, c,\, \mathcal{M} = 1\big)$ and $p(c \mid \mathcal{M} = 1)$.

(b) Evaluate the model evidence $p(\mathbf{y} \mid \mathcal{M} = 1, \mathbf{x})$ using a Laplace approximation. You may assume that $0.664 \simeq 11e^{-0.664} - 5$, if useful.

6 Now consider a second model, $\mathcal{M} = 2$, for the data presented in Question 5: a Gaussian process with zero prior mean function and covariance function equal to

$$K(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right).$$

For the following, use a computer to help you with the linear algebra.

(a) Evaluate the model evidence $p(\mathbf{y} \mid \mathcal{M} = 2, \mathbf{x})$, and comment on the relative suitability of models 1 and 2.

(b) Compute the prediction (posterior mean and variance) for $y(6)$.

7 Assume that we aim to draw samples from a distribution $p(\mathbf{x})$ for $x \in \mathbb{R}^N$ which is proportional to

$$f(\mathbf{x}) = \begin{cases} 1 & 0 \leq \rho(\mathbf{x}) \leq r \\ 0 & 0 < \rho(\mathbf{x}) \end{cases}$$

where $\rho(\mathbf{x})$ is the Euclidean norm of $\mathbf{x}$ (the radius). We don't know f (in particular, we don't know r) a priori: all we can do is evaluate it for our choice of $\mathbf{x}$. We will use a Gaussian proposal distribution $g(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma^2 \boldsymbol{I})$ (where $\boldsymbol{I}$ is the N-dimensional identity matrix) for use in either importance or rejection sampling.

(a) The volume of a hypersphere of radius r in N dimensions is

$$V(r, N) = \frac{\pi^{N/2}}{(N/2)!} r^N.$$

Show that the fraction, f, of this volume which lies in a thin shell, thickness, ϵ, at the surface of the sphere is

$$f = 1 - \left(1 - \frac{\epsilon}{r}\right)^N.$$

(b) Evaluate this result for $\epsilon/r = 0.1$ for $N \in [1, 10, 100, 1000]$. What happens as N gets very large?

(c) If ρ is the Euclidean norm of $\mathbf{x}$ drawn from $g(\mathbf{x})$, it can be shown that

$$p(\rho^2) \simeq \mathcal{N}(\rho^2; N\sigma^2, 2N\sigma^4).$$

Define h as the ratio of the standard deviation of this distribution to its mean. Evaluate h for $N \in [1, 10, 100, 1000]$. What happens as N gets very large?

(d) What does this tell you about sampling strategies in high-dimensional spaces? Why might this favour approaches, such as the *Metropolis-Hasting* algorithm, which allow for an *adaptive* proposal distribution?