1 Preliminaries

Bayesian inference is so thoroughly based on calculating probabilities, that we might be tempted to say that it is nothing else than probability calculus put in practice. Every bayesian analysis involves definitions and calculations of probabilities or probability densities. They are the essential molding mass of probabilistic modeling. Hence, we could start with a quick review of some common concepts and notations (which are loosely used here). Note: these are mainly simply mathematical tools of probability calculus without (necessarily) any interpretation of probability beyond its mathematical definition! The same material can be found more extensively written in text books on elementary probability theory.

Random event, random variable. By an event A we can denote some actual event that may occur in a series of repeatable experiments, e.g. $A =$"a tail occurs in a coin toss", or $A =$"a measurement X is larger than 46". In the latter case, the event would be a statement regarding a specific (random) variable X - the value of the measurement. The value of X determines whether event A happened (is true) or not (is false). Generally, the 'event' is a logical proposition that is either true or false, and can also describe an unrepeatable state of affairs, e.g. $A =$"Jack is taller than John". (Using random variable notation: if $X =$"height of Jack", $Y =$"height of John", then $A =$"X>Y").

Probability for event A (in bayesian context) denotes the degree of uncertainty we have about the truth value of A. Hence, if you know Jack (short) and John (long) well, you might have $P(X > Y) = 0$, but for an uninformed outsider, it could well be that $P(X > Y) = 0.5$ (before receiving any more information than the names). Mathematically, probability is a *measure* that takes values between zero and one. With two events, A and B, the probability that *both* occur (is true) is written $P(A, B)$, or with specific variables: $P(X = x, Y = y)$, or $P(X \in S_1, Y \in S_2)$, where upper case letter denotes the random variable, and lower case letter denotes a specific value of it.

Probability distribution. For a discrete variable X, taking values in the set $\{x_1, x_2, \ldots\}$, this is the numerative collection of point probabilities $P(X = x_i) = P_i \geq 0$ so that $\sum P_i = 1$. Likewise, for a continuous variable X, taking values x in some set $S \subset \mathbb{R}^n$, this is the *probability density function* $\pi(x) \geq 0$ so that $\int_S \pi(x)\mathbf{d}x = 1$. Note that probability *density* is not the same as probability, because $P(x) = 0$ for all x, but the density is $\pi(x) \geq 0$. If a function does not integrate to one but to some other constant C, $(-\infty < C < \infty)$, it can always be *normalized* to make a **proper** probability distribution. If it integrates to infinity, it is said to be an **improper** probability distribution. Surprisingly, these can sometimes be used too! The *support* of a density is the set of x values for which $\pi(x) > 0$.

Cumulative probability distribution function: $F(x) = P(X \leq x)$, where X can be either discrete or continuous. Note: $F(-\infty) = 0$, and $F(\infty) = 1$. It may sometimes be useful to calculate things like: $P(a < X \leq b) = F(b) - F(a)$, or $P(X > c) = 1 - F(c)$.

Transformation of variable. If $\pi(x)$ is a probability density, and $y = g(x)$ is a continuous smooth function of x, $(x = g^{-1}(y))$, then the probability density of y is $\pi(g^{-1}(y)) \mid \frac{dx}{dy} \mid$. (Note that the support of this new density is usually different from the original).

Conditional probability for events A and B, and conditional probability density for values of $X = x$ and $Y = y$

$$P(A \mid B) = \frac{P(A,B)}{P(B)}, \text{ and } \pi(x \mid y) = \frac{\pi(x,y)}{\pi(y)}$$

Product rule. Due to symmetry of $P(A,B)$ and $\pi(x,y)$ we have: $P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$, and $\pi(x,y) = \pi(x \mid y)\pi(y) = \pi(y \mid x)\pi(x)$. The product rule leads to the most important equation in bayesian modelling: the bayes formula itself.

Sum rule. $P(A \text{ or } B) = P(A) + P(B)$ if A and B are mutually distinct, i.e. $P(A,B) = 0$. Otherwise, more generally, $P(A \text{ or } B) = P(A) + P(B) - P(A,B)$.

Expected value of a random variable $X \in S \subset \mathbb{R}^n$

$$E(X) = \sum_i x_i P(X = x_i) \qquad \text{or} \qquad \int_S x\pi(x)\mathbf{d}x,$$

where P denotes probability and π denotes probability density.

Variance of a random variable:

$$V(X) = E(X - E(X))^2 = \sum_i (x_i - E(X))^2 P(X = x_i) \qquad \text{or} \qquad \int_S (x - E(X))^2 \pi(x)\mathbf{d}x$$

Variance can also be written in this form:

$$V(X) = E(X^2) - (E(X))^2.$$

Independence and conditional independence. Variables X and Y are said to be independent if $P(X,Y) = P(X)P(Y)$. They are said to be conditionally independent, given Z, if $P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$. (Likewise with probability densities $\pi(X,Y \mid Z)$).

If variables X & Y are independent, then the following equations hold:

$$E(XY) = E(X)E(Y) \text{ and } V(X+Y) = V(X) + V(Y).$$

The following equations hold for any random variables, whether they are independent or not:

$$E(X+Y) = E(X) + E(Y) \text{ and } V(cX) = c^2 V(X) \text{ and } E(cX) = cE(X),$$

where c is a constant.

Conditional expected value $E(X \mid Y)$. This is obtained from the previous formulation of $E(X)$ by substituting the distribution of X by the conditional distribution of X. The marginal expected value can be written as $E(X) = E(E(X \mid Y))$, where the outer expected value is taken with respect to Y.

Conditional variance $V(X \mid Y)$. This is similar to conditional expected value. But now we have: $V(X) = E(V(X \mid Y)) + V(E(X \mid Y))$.

Marginal probability:

$$P(X) = \sum_i P(X, y_i) \quad \text{if } X \in \{x_1, \ldots\} \text{ and } Y \in \{y_1, \ldots\} \text{ discrete.}$$

$$\pi(x) = \int_{S_y} \pi(x, y)\mathbf{d}y \quad \text{if } X \text{ and } Y \text{continuous.}$$

Marginal probability is computed similarly from multivariate models; by 'integrating out' the other variables. The marginal probability can also be computed for a k-dimensional vector variable that is part of a n-dimensional larger vector, $(n > k)$, for which the *joint distribution* is $P(X_1, \ldots, X_n)$. Using marginal distributions is an essential practical method for computing and visualizing results from multidimensional joint distributions. This will be used in nearly all practical bayesian applications!

A special random variable: indicator variable

$$I_{\{A(x)\}}(x) = \begin{cases} 1 & \text{if } A(x) \text{ is true} \\ 0 & \text{if } A(x) \text{ is false} \end{cases}$$

For an indicator variable we obtain:

$$E(I_{\{A(x)\}}) = 1P(A(x) \text{ is true}) + 0P(A(x) \text{ is false}) = P(A(x) \text{ is true})$$

The indicator variable is sometimes convenient in mathematical manipulations. Moreover, it will later provide us a simple tool for calculating many probabilities in WinBUGS by taking the average of a suitable indicator variable.

Completion of squares. This is a mathematical routine that is often used in solving posterior densities with Gaussian (normal) models. A square that needs to be completed is typically of the form $(a-b)^2 = a^2 - 2ab + b^2$. An incomplete square is thus completed by adding and subtracting one of the missing terms, e.g.:

$$a^2 - 2ab = a^2 - 2ab + b^2 - b^2 = (a-b)^2 - b^2.$$

In matrix algebra, if a and b are vectors (of size $n \times 1$):

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

and $a^T = (a_1, \ldots, a_n)$ and $b^T = (b_1, \ldots, b_n)$ are transposes of the vectors (of size $1 \times n$), the square (a scalar) is:

$$(a-b)^T(a-b) = a^Ta - a^Tb - b^Ta + b^Tb = a^Ta - 2a^Tb + b^Tb$$

Special functions:

Gamma-function, some useful properties: $\Gamma(N+1) = N!$, and $\Gamma(N+1) = N\Gamma(N)$ for integers N.

Beta-function: $\text{beta}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$