Glossary

Analytical solution An analytical solution to a particular mathematical problem (e.g. optimising a quantity or evaluating an integral) is one in which the solution can be obtained exactly. Many of the problems that we will deal with will not have analytical solutions, necessitating the use of iterative algorithms or sampling techniques.

Biased An estimator (e.g. $\widehat{\sigma^2}$ in Chapter 2) is said to be biased if, on average, it does not equal the true value.

Binomial distribution A popular probability distribution that describes the number of successes in a set of binary trials.

Burn-in When generating samples using MCMC, it is common to throw away the first N as the algorithm may not have converged and hence these are not representative. Determining N is not straightforward.

Conditional independence Two (or more) random variables A and B are said to be conditionally independent if their joint distribution, conditioned on C, can be factorised as $P(A, B \mid C) = P(A \mid C)P(B \mid C)$. Conditional independence does not imply unconditional independence.

Conditional probabilities Conditional probabilities are used to describe the probability of events that depend on the outcome of other events. For example, if the value of the random variable A depends upon the value of the random variable B, we can write the probability of A given the value of C as $P(A \mid C)$.

Conjugate A prior and likelihood are said to be conjugate if they result in a posterior of the same form as the prior.

Continuous random variables Random variables defined on a sample space that cannot be systematically enumerated. For example, random variables defined over all real numbers.

Convergence (sampler) A sampler is said to have converged when the samples it is generating are all coming from the same distribution. Before the sampler has converged, the samples should not be used.

Covariance Covariance is the generalisation of variance to the distributions over several variables. The covariance matrix describes how the different variables co-vary – how they are related.

Cross-validation A technique used for validation and model selection. The data is randomly partitioned into K groups. The model is then trained K times, each time with one of the groups left out.

Decision boundary A line separating two classes in a classification problem.

Deterministic Something that is not random. For example, our model in Chapter 1, $t = \mathbf{w}^\mathsf{T}\mathbf{x}$, is deterministic. The same value of $\mathbf{x}$ will always give the same value of t.

Discrete random variables Random variables defined over a sample space that can be systematically enumerated.

Discriminative classifier A classifier that explicitly defines (and optimises) decision boundarys between the classes.

Expectation For a (discrete) random variable, X, the expected value of some function of X, $f(X)$, is defined as:

$$\mathbf{E}_{p(X)}\left\{f(X)\right\} = \sum_x P(x)f(x).$$

This can be thought of as an average weighted by how likely the different values of X are. For continuous random variables, the summation is exchanged for an integral.

Feature selection In some classification problems it is useful to reduce the number of attributes. This process is known as feature selection. Common techniques for feature selection are scoring functions (pick the attributes/features) with the highest scores, clustering (cluster the attributes and use the cluster means as the new attributes) and projection techniques such as principal components analysis.

Fisher information The Fisher information is a measure of how much information a random variable provides about a particular model parameter.

Function A way of defining a relationship between two or more variables. For example,

$$t = f(x)$$

tells us that t depends on x – if we know x we can compute t.

Generalisation Generalisation is the ability to take something that has been learnt from one set of objects and apply it to previously unseen objects. For example, our Olympic model in Chapter 1 is generalising well if it makes good predictions for future Olympic sprints. In other words, an algorithm that exhibits good generalisation performance is one that is able to make good predictions on previously unseen data.

Generative model A generative model defines a process that could have generated the observed data. Thinking in terms of potential generative processes is often a useful abstraction when building models.

Global optimum For a function that can have many maxima (or minima), the global optimum is described as the highest (or lowest).

Graphical model A graphical representation of a probability distribution in which nodes correspond to random variables and directed edges to dependency relationships.

Hessian matrix The matrix of second derivatives of a function with respect to each pair of variables. Developed and named after Ludwig Otto Hesse, a 19th centutry German mathematician.

Hyper-parameter A parameter controlling the prior over another parameter in an hierarchical Bayesian model.

Information theory The quantitative study of information. In particular, the information content of a random variable is linked to its probability distribution. A distribution that is very uncertain has a high information content.

Joint probability The joint probability of two random variables A and B is the probability that they each take a specific value. For example, the probability that A takes value a *and* B takes value b. This is written as $P(A = a, B = b)$.

Likelihood The value of the density function (or distribution if the data are discrete) of the data, conditioned on any model parameters, evaluated at the data. This is a single numerical value, which is optimised with respect to the model parameters to produce the maximum likelihood solution.

Linear A function $t = f(x)$ is said to be linear if it satisfies the following conditions:

$$f(x_1 + x_2) = f(x_1) + f(x_2)$$
$$f(ax) = af(x)$$

A common example is $f(x) = wx$.

Mahanalobis distance The Mahanalobis distance between two objects $\mathbf{x}_n$ and $\mathbf{x}_m$ is defined as:

$$(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A}(\mathbf{x}_i - \mathbf{x}_j).$$

Substituing $\mathbf{A} = \mathbf{I}$, we recover the standard squared Euclidean distance. The matrix $\mathbf{A}$ creates a warping of the space such that distances are not the same in all directions. The set of points that have a particular squared Euclidean distance away from, say, $\mathbf{x}_n$, form a circle. The set of points a certain Mahanalobis distance away from $\mathbf{x}_n$ form an ellipse, the shape of which is defined by $\mathbf{A}$.

Marginal likelihood The denominator of Bayes' rule. A useful quantity for model comparison and choice.

Marginalisation The act of removing a random variable from a joint distribution by summing (or integrating if it is continuous) the joint distribution over all possible values that the random variable can take. For example:

$$P(A = a) = \sum_b P(A = a, B = b).$$

Maximum likelihood A popular parameter estimation scheme, where parameters are chosen that maximise the likelihood of the observed data.

Maximum a posteriori A popular way of choosing point estimates for parameter values that extends maximum likelihood by adding a regularising prior term.

Metropolis–Hastings A popular algorithm for generating samples from a density that does not require evaluation of the normalising constant.

Model complexity A term used to describe how complex a model is. For example, $t = w_0 + w_1 x$ is less complex than $t = w_0 + w_1 x + w_2 x^2$ and as such, is not able to find as complex patterns in data.

Model selection Model selection is the task of selecting which model to use for a particular task. The model choices could all come from the same family, although they don't have to. For example, if we wish to use a polynomial function $t = \sum_{k=0}^{K} w_k x^k$, choosing a suitable value for K is a model selection problem.

Model A mathematical description of a process. For example, in Chapter 1 we proposed the model $t = w_0 + w_1 x$ to represent the winning time in a 100 m sprint in Olympic year x.

Mode The mode of a distribution over some random variable is the most likely value.

Monotonic function A monotonic function is one that increases or decreases indefinitely. A common example is $\log(x)$ that always increases as x increases. This has the useful property that the value of x that minimises $f(x)$ will also minimise $\log(f(x))$.

Monte Carlo approximation An approximation to an expectation performed by drawing samples from the appropriate distribution. An expectation of the form:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\ dx$$

is approximated by:

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} \approx \frac{1}{S}\sum_{s=1}^{S} f(x^s),$$

where $x^1, \ldots, x^S$ are S samples from $p(x)$.

Multinomial distribution A popular distribution over vectors of integers. For example, if I role a die N times and record the number of times I obtain each face value in a six-dimensional vector, this vector could be modelled as a random variable with a multinomial distribution.

Natural logarithm The logarithm to the base e, referred to here as log but often referred to as ln.

Noise Variability in data that is assumed to be not of interest for the problem at hand. For example, random fluctuations brought about by measurement error.

Over-fitting A model is said to be over-fitting if it is too complex and is using its surplus complexity to fit to noise. Over-fitted models usually generalise badly.

Parameters Variables used to define a model. For example, the model

$$t = w_0 + w_1 x$$

has two parameters – w_0 and w_1.

Partial derivatives Taking partial derivatives of a function of several variables involves differentiation with respect to each variable whilst treating other variables as constant. For example, if the function $t = f(x, y)$ is defined as

$$t = 2x^2 + 3y^3 + xy,$$

the partial derivatives with respect to x and y are:

$$\frac{\partial f(x,y)}{\partial x} = 4x + y$$
$$\frac{\partial f(x,y)}{\partial y} = 9y^2 + x$$

Plate In graphical models, a shorthand used to show that there are several instances of a particular type of random variables.

Polynomial A polynomial function $t = f(x)$ has the form $t = \sum_{k=0}^{K} w_k x^k$. Common examples are the first order (or linear) polynomial $t = w_0 + w_1 x = \sum_{k=0}^{1} w_k x^k$ (called first order because the highest power to which x is raised is 1) and quadratic (second order) polynomial $t = w_0 + w_1 x + w_2 x^2 = \sum_{k=0}^{2} w_k x^k$. Note that $x^0 = 1$.

Posterior distribution The posterior distribution is the distribution over our parameter values after we have observed some data.

Precision In hierarchical Bayesian models it is often convenient to work with the precision rather than the variance. The precision is defined as:

$$\tau = \frac{1}{\sigma^2}.$$

Hence a Gaussian with mean μ and variance σ^2 can also be represented using precision τ as:

$$\mathcal{N}(\mu, \tau^{-1}).$$

Prior distributions Distributions describing our knowledge parameter values before any data has been observed.

Probability density function A probability density function (pdf) describes how the probability mass of a continuous random variable is distributed across its sample space. Probability density functions must always be positive and the integral of the probability density function over the sample space must be equal to 1.

Probability distribution A function or set of values that describes the characteristics of a random variable.

Probability The probability of an event taking place is a number between 0 and 1 that describes how likely the event is to take place.

Projection algorithms A family of Machine Learning algorithms that project data from M dimensions into D dimensions ($D \ll M$). Projection techniques are useful for visualisation (with $D = 2$) and can also be used for data pre-procesing for, for example, classification.

Quadratic A quadratic function $t = f(x)$ is a polynomial function where the highest power to which x is raised is 2. For example, $t = x^2$ and $t = w_0 + w_1x + w_2x^2$ are both quadratic functions.

Random events Events for which we cannot (or do not want or need to) define a deterministic model. For example, rolling a die or tossing a coin. Although we do not know the outcome of such events, it is likely that we will know the relative likelihoods of different outcomes.

Random variable A variable that stores the result of a random event. For example, if we toss a coin and assign the variable X the value 1 if the coin lands with the heads face up and 0 if it lands with the tails face up is a random variable.

Random walk A sequence of samples where each depends on the previous one.

Regularisation Regularisation is the act of placing restrictions on parameter values to limit the maximum complexity of a model.

Sample space The space of possible values that can be taken by a random variable. In other words, the set of the possible outcomes of a particular random event.

Statistics Statistics describes the collection of techniques and principles concerning the collection and interpretation of data.

Supervised learning Machine learning tasks where one is provided with a set of data objects and some associated labels.

Symmetric matrix A square matrix $\mathbf{X}$ is symmetric if $x_{ij} = x_{ji}$ for all j, i. If this is the case, then it follows that $\mathbf{X}^\mathsf{T} = \mathbf{X}$.

Unbiased An estimator (for example, $\widehat{\mathbf{w}}$) is said to be unbiased if, on average, its value is equal to the true value.

Unsupervised learning Learning algorithms that do not require targets or labels. Examples include clustering and projection.

Validation data Data that is used to help choose model type and parameters that is not directly used to train the model.

Variance Variance is the expected squared difference between the random variable and its mean.