Prerequisites for Foundations of Machine Learning and Data Science

Luke Dickens

September 24, 2018

1 A note on these prerequisites

This module is designed to give you the mathematical intuition behind a number of core/common machine learning algorithms. There is a significant practical aspect too, in which you will learn to use the methods and evaluate their performance on real world data. However, the module is not purely about learning to use machine learning libraries. It is about learning how and why they work too. Incoming students are expected to have **some background in mathematics** (see the next section) and to have **the rudiments of programming in Python**. If you have never programmed in Python before, then you should have all you need by simply working your way through the following online tutorial: DataCamp – Intro to Python for Data Science [1] (it should take you about 4 hours).

2 Mathematical Prerequisites

Below is a brief list of the mathematical notation, results and techniques that you would be expected to be familiar with before starting the module. Much of this material is taught in a UK mathematics A-level (school leaving certificate). However, some of the notation may be unfamiliar, and some ideas might go a little beyond high school mathematics. These represent many of the concepts that we will be building on for the content of the module. You should therefore expect to encounter them early in the module, and for the module material to extend and generalise from these concepts. There are also some questions within the table. These represent ideas that are less essential you know in advance, but still useful.

If there are a small number of items that are hazy or unfamiliar, then there is no reason you cannot familiarise/refamiliarise yourself with these in advance to taking the module. However, if a significant proportion is new to you, then you may find the mathematics in the module too challenging. Please contact me (`l.dickens@ucl.ac.uk`) if you would like some advice/links to supporting material.

Prerequisite	Symbols	Examples & Notes
Sets/Discrete		
Set membership	$\in, \notin$	$x \in X$ (x is member of X)
		$x \notin X$ (x is not a member of X)
Set operations	$\cup$ (union)	If $x \in A \cap B$ then $x \in A$ and $x \in B$

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

[1] `https://www.datacamp.com/courses/intro-to-python-for-data-science`

Prerequisite	Symbols	Examples & Notes
	$\cap$ (intersection)	If $x \in A \cup B$ then $x \in A$ or $x \in B$
	$\setminus$ (set-minus)	If $x \in A \setminus B$ then $x \in A$ and $x \notin B$
Set relations	$\subseteq$ (subset of) $\subset$ (strict subset of)	How would you formalise these concepts using the above notation?
Types of numbers	$\mathbb{N}$ (natural numbers), $\mathbb{Z}$ (integers) $\mathbb{R}$ (real numbers)	The integers are part of the reals, $\mathbb{Z} \subset \mathbb{R}$. Often we denote a range of real numbers as $[a, b] \subset \mathbb{R}$ meaning all numbers between a and b (inclusive).
Partition (of a set)		A set's partition is a grouping of that set's elements into non-empty subsets, such that every element is included in one and only one of the subsets. Example: if $\{A_1, A_2, \ldots, A_n\}$ is a partition of set A, then $A_i \cap A_j = \emptyset$ (the emptyset) for any $i \neq j$ and $A_1 \cup A_2 \cup \ldots \cup A_n = A$.
Summation	$\sum$	$\sum_{i=1}^{n} a_i = a_1 + \ldots + a_n$
Product	$\prod$	$\prod_{i=1}^{n} a_i = a_1 \times \ldots \times a_n$
Functions		
Integer Powers	a^n ($a \in \mathbb{R}$, $n \in \mathbb{Z}$)	$a^n = a \cdot a \cdot \ldots \cdot a$ for $n > 0$
		$b^{-1} = \frac{1}{b}$ for $b \neq 0$ (what if $b = 0$?)
		we normally say $a^0 = 1$ (why?)
General Powers	a^b ($a, b \in \mathbb{R}$)	$b^{m+n} = b^m \cdot b^n$
		$(b^m)^n = b^{m \cdot n}$
		$(b \cdot c)^n = b^n \cdot c^n$
		How do we define a^b for $a, b \in \mathbb{R}$?
Logarithms	ln (natural log)	$\ln a = x$ means $a = e^x$,
	$\log_b$ (log base b)	$\ln_b a = c$ means $a = b^c$,
		$\log_{10} 100 = 2$,
		$\log_2 64 = 8$,
		$\ln xy = \ln x + \ln y$,
		$\ln \frac{x}{y} = \ln x - \ln y$,
		$\ln x^n = n \ln x$
Factorials	!	$n! = n \cdot (n-1) \cdot \ldots \cdot 1$ for $n \in \mathbb{N}$
		We normally say $0! = 1$ (why?)

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

Prerequisite	Symbols	Examples & Notes
Functions	$f : \mathbb{R} \to \mathbb{R}$	A function f which takes a single real number input and outputs a single real number, e.g. $f(x) = y$ where $x, y \in \mathbb{R}$, can be written $f : \mathbb{R} \to \mathbb{R}$.
	$f : \mathbb{R}^2 \to \mathbb{R}$	A function can take 2 inputs, and output a scalar. Here, f takes a two real number inputs and outputs a single real number, e.g. $f(x, y) = z$ where $x, y, z \in \mathbb{R}$.
	$f : \mathbb{R}^m \to \mathbb{R}^n$	More generally, a function can take m inputs (or an m-vector of inputs) and output an n-vector.
Vector/Matrix		
Vectors	$\boldsymbol{v} \in \mathbb{R}^n$	$\boldsymbol{v} \in \mathbb{R}^n$ means $\boldsymbol{v}$ is a *stack* of n real numbers. v_i used to refer to the ith element of $\boldsymbol{v}$
Matrices	$A \in \mathbb{R}^{m \times n}$	$A \in \mathbb{R}^{m \times n}$ is a matrix with m rows and n columns we can visualise it as: $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$ We can refer to the (i, j)th element as $[A]_{ij}$ A^T is the transpose of A (switching rows and columns) so $[A^T]_{ij} = [A]_{ji}$
The dot product	$\boldsymbol{a} \cdot \boldsymbol{b}$ or $\boldsymbol{a}^T\boldsymbol{b}$	$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i$ The dot product is sometimes written $\boldsymbol{a}^T\boldsymbol{b}$ assuming column vectors (why?) The (Euclidean) length of a vector is: $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}} = \sqrt{\boldsymbol{v}^2}$ The (Euclidean) distance between two vectors is: $\|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^T(\boldsymbol{a} - \boldsymbol{b})} = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^2}$
Cosine Rule		If θ is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$: $\boldsymbol{a}^T\boldsymbol{b} = \|\boldsymbol{a}\|\|\boldsymbol{b}\| \cos\theta$

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

Prerequisite	Symbols	Examples & Notes
Matrix multiplication		AB for $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ $$[AB]_{ij} = \sum_k a_{ik} b_{kj}$$
Matrix Determinant	$\det(A)$ or $\|A\|$	The determinant of a square matrix A is a positive number that can be computed from A elements. For $A \in \mathbb{R}^{2 \times 2}$, $$\|A\| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$ Geometrically, $\|A\|$ is the scaling factor of the linear transformation described by matrix A.
Matrix Rank	$\text{rank}(A)$	The rank of a matrix A is the number of linearly independent columns (or rows) of A. Equally, if we treat every column of A as a vector, then $\text{rank}(A)$ is the dimension of the space spanned by these vectors.
Identity Matrix	I, I_n	The square matrix $I \in \mathbb{R}^{n \times n}$ has 1s on the diagonal and zeros elsewhere, so $[I]_{ii} = 1$ and $[I]_{ij} = 0$ when $i \neq j$. I is the matrix analog of unity (1). If the dimension of I is unclear, we can denote it with a subscript, e.g. $I_m \in \mathbb{R}^{m \times m}$. Multiplying by I leaves your matrix (or vector) unchanged, e.g. for $A \in \mathbb{R}^{m \times n}$, $I_m A = A = A I_n$
Matrix Inversion	A^{-1}	If a square matrix $A \in \mathbb{R}^{n \times n}$ is invertible, then its inverse A^{-1} multiplied (left or right) by A gives the identity matrix, i.e. $$A^{-1}A = I = AA^{-1}$$ Not all (square) matrices are invertible. How would you check if a matrix is invertible? The inverse is the matrix analog of the reciprocal $r^{-1} = \frac{1}{r}$ of a scalar $r \in \mathbb{R}$. Do you know how to find the inverse of a 2×2 matrix?

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

Prerequisite	Symbols	Examples & Notes
Eigenvalues & Eigenvectors		An eigenvector $\mathbf{x} \in \mathbb{R}^n$ of a square-matrix $A \in \mathbb{R}^{n \times n}$ satisfies the following equation: $$A\mathbf{x} = \lambda\mathbf{x}$$ for some corresponding (scalar) eigenvalue $\lambda \in \mathbb{R}$. Geometrically an eigenvector, corresponding to a real nonzero eigenvalue, points in a direction that is stretched by the transformation and the eigenvalue is the factor by which it is stretched.
Calculus		
Differentiation	$f'(x)$, $\frac{df}{dx}$	For a function $f : \mathbb{R} \to \mathbb{R}$, then the slope at x (if it exists) is $$f'(x) = \frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$
Simple Derivatives		e.g. $\frac{d}{dx}c = 0$ for constant c $\frac{d}{dx}x^a = a \cdot x^{a-1}$ $\frac{d}{dx}e^x = e^x$ $\frac{d}{dx}\ln x = \frac{1}{x}$ $\frac{d}{dx}\sin(x) = \cos(x)$ $\frac{d}{dx}\cos(x) = -\sin(x)$ What are the product, quotient and chain rules?
Integration	$\int_a^b f(x)dx$, $\int f(x)dx$	To know what is meant by a definite and indefinite integral (Riemann integration), and to have some idea of the requirements on $f(x)$, e.g. finite etc.
Common Integrals		e.g. $\int k dx = kx + C$ $\int x^a dx = \frac{x^{a+1}}{(a+1)} + C$ $\int e^{ax} dx = \frac{1}{a}e^{ax} + C$ etc
Taylor Series		A representation of a function as an infinite sum of terms that are calculated from the values of the function's derivatives at a single point. For function f (with appropriate properties) $$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots$$ Do you know the Taylor series for e^x?

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

Prerequisite	Symbols	Examples & Notes
Partial derivative	$\frac{\partial f}{\partial x}(x,y)$, $\frac{\partial}{\partial x}(f(x,y))$	The partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant.
Gradient	∇f, $\nabla_x f$	The gradient is a multi-variable generalization of the derivative. Like the derivative, the gradient represents the slope of the tangent of the graph of the function. For function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient $$\nabla f(\mathbf{x}) = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^T$$ points in the direction of the greatest rate of increase of the function f at point $\mathbf{x}$.
Using Derivatives		How do you find the maxima, minima and points of inflection of a function? E.g. set $$f'(x) = 0$$ How do you determine what type of point you have found? What if your function takes 2 or more inputs e.g. $f : \mathbb{R}^m \to \mathbb{R}$? How do you find the maxima, minima and other stationary points of this function?
Statistics and Probability		
Mean		The mean of some set of values $x_1, x_2 \ldots x_N$ is $$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$ If these values are the outcomes (samples) of N repetitions of some experiment, this might be called the *sample mean*.
Probability distribution		A random variable X with a finite number of outcomes $x_1, x_2, \ldots, x_k$ can be represented by finite number of weights $p_1, p_2, \ldots p_k \in [0,1]$ such that $\sum_{i=1} p_i = 1$, where p_i is the probability of X taking value x_i.

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

Prerequisite	Symbols	Examples & Notes
		For a 1-dimensional random variable, X, which can take any value, x, in a continuous range, its distribution is typically described by a *probability density function*, $p(x)$. The probability that $a < X < b$ for some a and b is given by: $$p(X \in [a, b]) = \int_a^b p(x)\, dx$$
Cumulative Distribution Function (CDF)	$F(x)$, $F_X(x)$	For a random variable X the CDF specifies the probability that the random variable takes a value less than or equal to x, i.e. $$F_X(x) = p(X \leq x)$$
Expectations	$\mathbf{E}$	For a random variable X with a finite number of outcomes $x_1, x_2, \ldots, x_k \in \mathbb{R}$ occuring with respective probabilities $p_1, p_2, \ldots p_k$, the expectation of X is $$\mathbf{E}[X] = \sum_{i=1}^{k} p_i x_i \quad (1)$$ For a continuous random variable $a < X < b$ with probability density function $p(x)$ the expectation of X is $$\mathbf{E}[X] = \int_a^b xp(x)dx$$ We sometimes call $\mathbf{E}[X]$ the mean of X.
Variance		The variance of a finite random variable X, with outcomes $x_1, x_2, \ldots, x_k \in \mathbb{R}$ and respective probabilities $p_1, p_2, \ldots p_k$ is $$\text{Var}[X] = \sum_{i=1}^{k} p_i (x_i - \mu)^2$$ where μ is the mean of X (i.e. $\mathbf{E}[X]$). What is the variance of a random variable X with probability density function $p(x)$?

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.

Prerequisite	Symbols	Examples & Notes
Normal Distribution		The normal (Gaussian) distribution is a very commonly encountered continuous probability distribution, with probability density function: $$f(x\|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$ where μ is the mean, and σ^2 is the variance.
Linear Regression		In the 1-dimensional case, we assume that we observe n pairs of values $(x_i, y_i) \in \mathbb{R}^2$ for $i = 1, \ldots, n$, and we want to find the straight line $$y = f(x) = \beta_0 + \beta_1 x$$ that *best fits* these values, i.e. $y_i \approx f(x_i)$ for each i. The meaning of this can vary, but in the simplest case we want to *minimise the sum of squared errors*, i.e. find the parameters (β_0, β_1) that minimises $$\sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i))^2$$ How do we find the *best values* for (β_0, β_1)? We might call the x_is the input variables or independent variables, and the y_is the output variables or dependent variables. What changes if we want to find the best fit for this function $$y = \beta_0 + \beta_1 x + \beta_2 x^2 \quad ?$$ What if we allow each input variable to be a k-dimensional vector, i.e. $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T \in \mathbb{R}^k$, and want to ensure that $$y_i \approx \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \quad ?$$ What assumptions are we typically making when we use these techniques or interpret their results?

Table 1: You should be relatively familiar with the majority of the mathematical notation, results and techniques outlined in this table.