Mathematical Notation for Machine Learning

06/03/2020

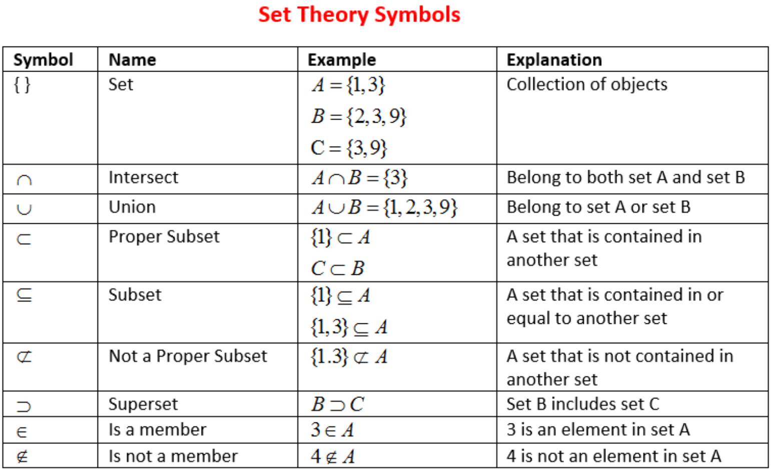
Introduction

- We will go over some of the mathematical notation needed to understand machine learning (ML) papers.
- This is a very broad area in general, but in particular, we'll be looking at set theory, linear algebra, and statistics notational conventions.

Set Theory

- Sets are the basis of all collections of mathematical objects and so are crucial structures that we must be able to communicate with fluently
- We denote a **set**, simply a collection of mathematical objects, using the $\{\}$ curly braces (e.g., $\{2, 3, 5, 7\}$).
- We can also define elements of a set within these braces, e.g., $\{p|p$ is a prime number less than $10\}$
- The following table illustrates a variety of set notations and their meanings:

Set Theory



Set Theory Symbols

Symbol	Name	Example	Explanation
$\{\,\}$	Set	$A=\{1,3\}$ $B=\{2,3,9\}$ $\mathrm{C}=\{3,9\}$	Collection of objects
$\cap$	Intersect	$A\cap B=\{3\}$	Belong to both set A and set B
$\cup$	Union	$A\cup B=\{1,2,3,9\}$	Belong to set A or set B
$\subset$	Proper Subset	$\{1\}\subset A$ $C\subset B$	A set that is contained in another set
$\subseteq$	Subset	$\{1\}\subseteq A$ $\{1,3\}\subseteq A$	A set that is contained in or equal to another set
$\not\subset$	Not a Proper Subset	$\{1.3\}\not\subset A$	A set that is not contained in another set
$\supset$	Superset	$B\supset C$	Set B includes set C
$\in$	Is a member	$3\in A$	3 is an element in set A
$\notin$	Is not a member	$4\notin A$	4 is not an element in set A

Linear Algebra

- Linear algebra notation gives us succinct ways to write down otherwise complicated interactions between datasets.
- Linear algebra deals with the set $\mathbb{R}^n$, which simply means the set of column vectors with n elements, all belonging to the real numbers (e.g., $a = (2, 3, 5, 7)^T \in \mathbb{R}^4$, where the T operator denotes the transpose of the vector)
- Note also that vectors are usually denoted with lowercase variables, e.g., a or x.
- the argmax() operator on a vector returns the index of the element of the vector with the highest value, so for a which we defined above, $\text{argmax}(a) = 4$

Linear Algebra

- Elements of $\mathbb{R}^{mxn}$ are matrices with m rows and n columns, e.g., $A = \begin{pmatrix} 2 & 3 & 5 \\ 7 & 11 & 13 \end{pmatrix} \in \mathrm{R}^{3x2}$
- We usually denote matrices with uppercase variable names like A or X
- Matrix multiplication is written just like multiplication of regular variables: A times $B = AB, A, B \in \mathbb{R}^{nxn}$
- Note that for matrix multiplication, the number of columns in the first matrix **must** match the number of rows in the first
- The forward propagation step of a neural network is nothing but matrix multiplication under activation functions (see example papers for details).

Statistics

- With statistical notation, we are dealing with **random variables** which each follow **distributions**, which each have their own set of parameters that define them
- For example, if a random variable X follows a normal distribution with mean μ and variance σ^2, we write $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Each random variable has an **expectation** (or mean) denoted as $\mathrm{E}(X)$ and a **variance** denoted as $\mathrm{Var}(X)$.
- Many optimizers take basic gradient descent a step further by adding random noise to the steps of the gradient descent to avoid getting stuck in local minima. The selection of distribution for this random noise is therefore crucial to being able to train a model as fast as possible.

Conclusion

- This was a very brief overview of the mathematical notation you might see in a machine learning paper
- This was in no way comprehensive, but it should give you a pretty good idea of what to expect
- Other notations (like those related to multivariate calculus) I have withheld here in the interest of accessibility
- The best way to become more comfortable with notation is to read often and use it even more often!