week 1: preliminaries

http://mlvu.github.io

February 26, 2019

This week, the homework will be a review of subjects you should be familiar with already. Machine learning is built mostly on three types of math: *Linear Algebra*, *Calculus* and *Probability Theory*. Luckily, we only use a very limited number of concepts from each, so even if you don't know these subjects or don't know them well, it shouldn't take too much work to get up to speed.

Here we'll review some linear algebra and some calculus, saving the probability for later. We'll also review three concepts that you should become very familiar with to follow machine learning theory: sums, expectations and logarithms. Even if you know these in principle, you should take the time to get comfortable with them.

This homework should give you an indication of whether you have a sufficient grasp of the preliminaries. If you struggle with anything, please follow the links provided to brush up, before continuing with next week's homework.

1 Sums, expectations and logarithms

You've probably encountered all three of these before. However, in the lectures, we will often see long derivations which require you to to be very familiar with these in order to follow all the steps. If you find yourself baffled by some of the math in the lecture, you can probably get a lot closer by just practising your sums, expectations and logarithms a little.

Sums Sigma (Σ) notation is simply a concise way of writing down long sums. Assume we have a list of numbers a_1, a_2, a_3, a_4, a_5. We can write:

$$a_1 + a_2 + a_3 + a_4 + a5$$

but we can also write

$$\sum_{i=1}^{5} a_i \quad \text{or} \quad \textstyle\sum_{i=1}^{5} a_i$$

The bit below the sigma tells you which symbol is used as an index (i) to iterate over the elements to sum over, and what the first value is (1). The bit above the sigma tells you what the last value of the index is. [1]

Since this is a very dense notation, it's often simplified by leaving out the start and end values, assuming that they can be figured out from context:

$$\sum_{i} a_i$$

Sums can be nested as well (using a sum within a sum). This is not a special notation, it simply follows from the definition. You can 'unpack' these by turning the sigmas into a regular sum one by one:

$$\begin{aligned}
&\sum_{i=1}^{2} \sum_{j=1}^{3} a_i b_j \\
&= \sum_{j=1}^{3} a_1 b_j + \sum_{j=1}^{3} a_2 b_j \\
&= (a_1 b_1 + a_1 b_2 + a_1 b_3) + (a_2 b_1 + a_2 b_2 + a_2 b_3)
\end{aligned}$$

For nested sums, the indices are often combined into a single sigma:

$$\sum_{i,j} a_i b_j = \sum_{i} \sum_{j} a_i b_j$$

Exercise 1

Let $x_1 = 5, x_2 = 1, x_3 = 4, x_4 = 1, x_5 = 3$. Compute the following values:

1. $\sum_i x_i$
2. $\sum_i 2 \cdot x_i$
3. $2 \cdot \sum_i x_i$

[1] If you're an experienced programmer, it may be helpful to think of this as a `for` loop: something like `for(int i = 1; i < 6; i++)` ....

4. $\sum_{i=1}^{3} x_i$
5. $\sum_{i=1}^{x_5} x_i$
6. $\sum_{i=1}^{3} i \cdot x_i$

Exercise 2

Which of the following are always true? If you struggle, try rewriting as a normal sum, and using what you already know about sums. The variable c is a constant.

1. $\sum_i c y_i = c \sum_i y_i$
2. $\sum_i (c + y_i) = c + \sum_i y_i$
3. $\sum_{i=1}^{n} (c + y_i) = nc + \sum_{i=1}^{n} y_i$
4. $\sum_i (x_i + y_i) = (\sum_i x_i) + (\sum_i y_i)$
5. $\sum_i \sum_j a_i b_j = \sum_j \sum_i a_i b_j$
6. $\sum_i a_i b_i = (\sum_i a_i)(\sum_i b_i)$

It can be a little ambiguous where the sum operator ends. For instance, in $\sum_i x_i + c$, we don't know whether the c should be added once, or for every term in the sum. There's no official standard, and authors should make sure to write sums in an unambiguous way. For instance $\sum_i c + x_i$ or $c + \sum_i x_i$ or simply $\sum_i (c + x_i)$.

The capital pi Π works in exactly the same way as the Σ operator, but describes a *product*:

$$\prod_{i=1}^{4} i = 24$$

More practice? Try these links:

- https://www.khanacademy.org/math/algebra2/sequences-and-series/alg2-sigma-notation/e/evaluating-basic-sigma-notation
- https://www.mathsisfun.com/algebra/sigma-notation.html
- https://www.youtube.com/watch?v=TjMLzklnn2c

Expectations An expectation is nothing more than a sum, with each term weighted by a probability. Let's say you are playing a game where you roll a die, and you receive the number of eyes rolled in euros. Your expected reward is:

$$\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = 3.5$$

In general, if you have a random variable V (an experiment with a numeric outcome) which has possible outcomes $1, 2, \ldots$ with associated values $v_1, v_2, \ldots$ and associated probabilities $p_1, p_2, \ldots$, the **expected value** of the experiment is written as

$$\mathbb{E}(V) = \sum_i p_i v_i \tag{1}$$

The brackets can be omitted at the author's discretion. If multiple probability distributions are applicable, the choice of distribution may be indicated in the subscript $\mathbb{E}_p V$.

We can also extend the experiment by applying a function to the outcome, and computing the expected value over the result. Say that V describes a betting game with winnings v_i, but you will have to pay a 20% tax over whatever you win. Then, the expectation for the amount of money you actually get to keep is

$$\mathbb{E}(0.8 \cdot V) = \sum_i p_i \cdot 0.8 \cdot v_i$$

Exercise 3

Because the expectation is basically a sum, most of the things that hold for sums, also hold for expectations. Show that the following hold by applying what you know about sums. If you get stuck, try expanding the expectation into a sum, as in (1).

1. $\mathbb{E}(c \cdot V) = c \cdot \mathbb{E}(V)$ (homogeneity)
2. Let V and W have different numeric values, v_i and w_i, for outcomes with the same probability p_i. Then $\mathbb{E}(V) + \mathbb{E}(W) = \mathbb{E}(V + W)$ (additivity)
3. $\mathbb{E}(V) + \mathbb{E}(\sin(V)) = \mathbb{E}(V + \sin(V))$
4. $\mathbb{E}\left((V - \mathbb{E}V)^2\right) = \mathbb{E}(V^2) - (\mathbb{E}V)^2$

If your outcomes form a *continuum*, like real-valued numbers (e.g. if you sample from a normal distribution), the expectation is defined with an integral instead of a sum ($\mathbb{E}_V = \int_x p(x)v(x)dx$). Since the integral has most of the same properties as the sum, you can use both types of expectation in the same way. [2]

More practice? Try these links:

- https://www.khanacademy.org/math/probability/probability-geometry/expected-value-geo/a/expected-value-basic

Logarithms The *logarithm* is nothing more than the inverse of the exponent. If $y = b^x$ then $x = \log_b y$.

Since $100 = 10 \cdot 10 = 10^2$, $1000 = 10 \cdot 10 \cdot 10^3$ and so on, $\log_{10} x$ gives us an indication of the *order of magnitude* (essentially the number of digits) of x. For instance $\log_{10}(314) \approx 2.5$ and $\log_{10}(31400) \approx 4.5$. In other words, if two numbers have the same order or magnitude, the difference between their logarithms will be at most 1. This also works for numbers smaller than one: $\log_{10}(0.1) = -1$, $\log_{10}(0.01) = -2$, $\log_{10}(0.001) = -3$ and so on. For bases other than 10, it's not quite so neat, but the same principle holds: all numbers of a single order of magnitude are squeezed into a constant-size interval.

There are two main reasons why it's such a useful function:

- It allows us to use the limited-precision representations available in computers to accurately manipulate very small and very large numbers. This is useful, for instance, in representing probabilities. Since our probabilities are usually guesses anyway, we don't care as much about the difference between 0.0004 and 0.00035 as we do about the difference between 10^{-21} and 10^{-24}. This means that it's often much better to store the logarithm of the probability rather than the probability itself.

- Taking the logarithm **turns a product into a sum**:

$$\log(abc) = \log a + \log b + \log c$$

  Sums are much easier to analyze and manipulate. This can be a tremendous help in working out derivates, or probabilities.

[2] In fact, we use this trick in the course to hide some of the complexity: by talking about expectations without unpacking the notations, we can discuss many aspects of distributions on continuous spaces, without ever using integrals.

Exercise 4

Which of the following are true for all a, b? If true show why, if not, give a counterexample.

1. $\log_b(b^a) = a$
2. $\log a + \log b = \log(ab)$
3. $\log(a^b) = b\log(a)$
4. $\log(a + b) = \log(b) \cdot \log(a)$
5. $\log(a/b) = \log(a) - \log b$

More practice? Try these links:

- http://www.oxfordmathcenter.com/drupal7/node/532
- https://www.khanacademy.org/math/algebra2/exponential-and-logarithmic-functions
- https://study.com/academy/lesson/exponentials-and-logarithms-graphing-and-property-review.html

2 Linear Algebra

In all homework and lectures, bold lowercase letters like $\boldsymbol{x}$ indicate a vector, bold uppercase letters like $\boldsymbol{W}$ indicate a matrix and non-bold lowercase letters like x indicate a scalar (that is, a number).

Exercise 5

Explain in words what the following notations represent:

1. $f : \mathbb{R}^3 \to \mathbb{R}^2$
2. $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$
3. $z = \boldsymbol{y}^\top \boldsymbol{x}$
4. $\boldsymbol{W} \in \mathbb{R}^{5\times 4}$

Hint: if you're not sure, see if you can find the symbols in this page: https://en.wikipedia.org/wiki/List_of_mathematical_symbols. Even if that doesn't explain it fully, it may provide you with some keywords to google.

Exercise 6

Which of the following operations are possible, and which aren't?

1. Multiplying a 5×4 matrix by another 5×4 matrix.
2. Element-wise multiplying a 5×4 matrix by another 5×4 matrix.
3. Multiplying a 5×4 matrix by a 4×5 matrix.
4. Multiplying a 5×4 matrix by a 4×1 matrix.
5. Element-wise multiplying a 5×4 matrix by a 4×5 matrix.
6. Multiplying any matrix by its transpose.
7. Element-wise multiplying any matrix by its transpose.
8. Element-wise multiplying a square matrix by its transpose.

Exercise 7

Which of the following is true?

1. Let $U = VW$ for matrices U, V, W. U_{ij} is the dot product of the i-th column of V and the j-th column of W.
2. Let $U = VW$for matrices U, V, W. U_{ij} is the dot product of the i-th row of V and the j-th column of W.
3. Matrix multiplication is *commutative*.
4. Matrix multiplication is *distributive*.
5. Matrix multiplication is *associative*.
6. There exist matrices U and V such that $UV = VU$.

Exercise 8

1. The linear function $f(a, b, c) = \alpha a + \beta b + \gamma c$ can be written as the dot product of two vectors. What are the vectors?

2. The linear function $f(a, b, c) = \alpha a + \beta b + \gamma c + \delta$ can be also be written as the dot product of two vectors. What are the vectors?

3. The *quadratic* function $f(a, b) = \alpha a^2 + \beta ab + \gamma ba + \delta b^2$ can be written as $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W} \boldsymbol{x}$. What should $\boldsymbol{x}$ and $\boldsymbol{W}$ be?

If you found these exercises completely impossible, you should brush up on your Linear Algebra a little. You don't need much, the basics of vectors, matrices and matrix multiplication should do the trick. The following articles may help. If not, hunt around for one that explains things at your level:

- https://betterexplained.com/articles/linear-algebra-guide/
- https://www.khanacademy.org/math/linear-algebra
- http://samples.jbpub.com/9781556229114/chapter7.pdf

If you find anything that helps you, let us know on the discussion board.

3 Calculus

To denote the derivative of $f(x)$ we will use the notation $\frac{df(x)}{dx}$ (with ∂ taking the place of d for partial derivatives). You may be more comfortable with the notation $f'(x)$ for the derivative. I sympathize, but when it comes to machine learning, the former notation makes things much simpler down the line, so I recommend getting used to it.

Exercise 9

Let $f(x) = 3x^2 + 5x + 1$ with x a scalar.

1. What is the derivative of $f(x)$?

2. For which x is $f(x)$ at its minimum?

3. Let $h(x) = g(f(x))$, with f defined as above. Let $\frac{dg(x)}{dx} = \frac{\sin(x)}{x}$. Without knowing what $g(x)$ is (or working it out), can we find the derivative of $h(x)$?

Exercise 10

Let $\boldsymbol{x} \in \mathbb{R}^2$ and let $f(\boldsymbol{x}) = 3{x_1}^2 + 4x_1x_2 - {x_2}^2$

1. What is the *partial* derivative of $f(\mathbf{x})$ with respect to x_1?
2. What is the partial derivative of $f(\mathbf{x})$ with respect to x_2?
3. What is the *gradient* of $f(\mathbf{x})$?
4. The gradient is a function derived from f, just like the derivative is a function. What are the domain and range of the gradient of $f(\mathbf{x})$?

As before, if you found these exercises tricky, even after reading the answers, you should probably brush up a little on your calculus. Here are some links you may find helpful.

- https://betterexplained.com/articles/vector-calculus-understanding-the-gradient/
- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient-articles/a/introduction-to-partial-derivatives
- http://tutorial.math.lamar.edu/Classes/CalcIII/PartialDerivatives.aspx