1678 Lecture 1: Intro

What is deep learning?

Deep learning is the optimization of flexible and differentiable multi-layered neural networks (which approximate functions).

Linear Algebra Review

The dot/inner product is transforming two vectors into a scalar. For example:

$$\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 1 * 3 + 2 * 4 = 11$$

Matrix multiplication be like:

$$\begin{bmatrix} A & B & C \\ D & E & F \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} = \begin{bmatrix} 1A + 3B + 5C & 2A + 4B + 6C \\ 1D + 3E + 5F & 2D + 4E + 6F \end{bmatrix}$$

Note that the first matrix is 2 by 3, the second is 3 by 2, and the output is 2 by 2. This is a general thing: if you're multiplying A by B and B by C, the output will be A by C. Also, that B term always needs to match.

Partial Derivatives

Take $f(x, y) = x^2 + y^3$. When taking the partial derivative of f with respect to, say, x, we just do a normal derivative and treat every y term as a constant. (Because that expression is describing how f changes as **only** x changes.) So

$$\frac{\partial}{\partial x} f(x, y) = 2x$$
$$\frac{\partial}{\partial x} f(x, y) = 3y$$

Gradients

A gradient the vector of a function's partial derivatives with respect to each of its inputs. It's denoted with the ∇ symbol.

You can think of it as a multi-dimensional derivative– it shows how a function changes as its input**s** change.

Using that same $f(x, y) = x^2 + y^3$ example (but this is obviously generalizable to a function of any number of inputs):

$$\nabla f = \begin{bmatrix} 2x \\ 3y \end{bmatrix}$$

Probability

Probability is the number of wanted outcomes, divided by the number of possible outcomes. It's scaled to be between 0 and 1 (inclusive), such that the probabilities of all possible outcomes sum to 1.

Expected Value

Expected value is the weighted average outcome. It's clearly somewhere between 0 and 1, so if you get a number bigger than 1, something's gone horribly wrong. Or in math-ese, if you think that way:

$$E[x] = \sum_{x \in X} Pr(X = x)x$$

and

$$E[f(x)] = \sum_{x \in X} Pr(X = x) f(x)$$

Lastly, some identities:

$$E[x + y] = E[x] + E[y]$$
$$E[\alpha x] = \alpha E[x]$$

Conditional Probability

The probability of A, given B, is the probability of A and B both happening, divided by the probability of B.

$$Pr(A|B) = \frac{Pr(A, B)}{Pr(B)}$$