Linear Algebra Review

- A **vector** is an ordered list of values. It is often denoted using angle brackets: $\langle a, b\rangle$, and its variable name is often written in bold ($\mathbf{z}$) or with an arrow ($\vec{z}$). We can refer to an individual element of a vector using its index: for example, the first element of $\mathbf{z}$ would be z_1 (or z_0, depending on how we're indexing). Each element of a vector generally corresponds to a particular dimension or feature, which could be discrete or continuous; often you can think of a vector as a point in Euclidean space.

- The **magnitude** (also called **norm**) of a vector $\mathbf{x} = \langle x_1, x_2, ..., x_n\rangle$ is $\sqrt{x_1^2 + x_2^2 + ... + x_n^2}$, and is denoted $|\mathbf{x}|$ or $\|\mathbf{x}\|$.

- The **sum** of a set of vectors is their elementwise sum: for example, $\langle a, b\rangle + \langle c, d\rangle = \langle a + c, b + d\rangle$ (so vectors can only be added if they are the same length). The **dot product** (also called **scalar product**) of two vectors is the sum of their elementwise products: for example, $\langle a, b\rangle \cdot \langle c, d\rangle = ac + bd$. The dot product $\mathbf{x} \cdot \mathbf{y}$ is also equal to $\|\mathbf{x}\|\|\mathbf{y}\| \cos\theta$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$.

- A **matrix** is a generalization of a vector: instead of having just one row or one column, it can have m rows and n columns. A **square** matrix is one that has the same number of rows as columns. A matrix's variable name is generally a capital letter, often written in bold. The entry in the ith row and jth column of a matrix $\mathbf{A}$ is referred to as $a_{i,j}$, or sometimes $\mathbf{A}_{i,j}$ or $\mathbf{A}[i, j]$.

- The **sum** of two matrices is their elementwise sum. The **product AB** of two matrices $\mathbf{A}$ and $\mathbf{B}$ is defined as $(\mathbf{AB})_{i,k} = \sum_j \mathbf{A}_{i,j}\mathbf{B}_{j,k}$. To multiply two matrices, the number of columns in the first matrix must equal the number of rows in the second matrix; the product of an m-by-n matrix and an n-by-p matrix will be an m-by-p matrix. Matrix multiplication is associative ($(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$) but not commutative (usually $\mathbf{AB} \neq \mathbf{BA}$).

- The **transpose** of a matrix $\mathbf{A}$, denoted $\mathbf{A}^T$ (or $\mathbf{A}^t$ or $\mathbf{A}^\top$), swaps the rows with the columns: $(\mathbf{A}^T)_{j,i} = \mathbf{A}_{i,j}$. The transpose of an m-by-n matrix will be an n-by-m matrix. The n-by-n **identity** matrix $\mathbf{I}_n$ (or just $\mathbf{I}$ when it's unambiguous) is a square matrix with 1's on the diagonal (entries where $i = j$) and 0's everywhere else. For any m-by-n matrix $\mathbf{A}$, $\mathbf{AI}_n = \mathbf{A}$, and for any n-by-p matrix $\mathbf{B}$, $\mathbf{I}_n\mathbf{B} = \mathbf{B}$. The **inverse** of a square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix for which $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}(= \mathbf{AA}^{-1})$. A non-invertible square matrix is called **singular**.

- A **hyperplane** is a higher-dimensional generalization of lines and planes. The equation of a hyperplane is $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is a vector normal to the hyperplane and b is an offset. Note that we can multiply by any constant and preserve the equality; if we multiply by $1/\|\mathbf{w}\|$, we get a new equation $\hat{\mathbf{w}} \cdot \mathbf{x} + b' = 0$, where $\hat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$ is the **unit normal vector** and $b' = b/\|\mathbf{w}\|$ is the distance from the hyperplane to the origin.

- For any vector $\mathbf{x}$ we can compute $y = \mathbf{w} \cdot \mathbf{x} + b$. If $y = 0$, then $\mathbf{x}$ is on the hyperplane. If $y > 0$, then $\mathbf{x}$ is on one side of the hyperplane, and if $y < 0$, then $\mathbf{x}$ is on the other side of the hyperplane. This will be useful when we are developing linear classifiers.

- The **gradient** of a surface at a point is a vector whose magnitude is the highest rate of increase from that point and whose direction is the direction of that rate of increase. Formally, the gradient of a function $f(x_1, x_2, ..., x_n)$ is $\nabla f = \langle \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n} \rangle$. If we are trying to find a maximum value, we can use a method called **gradient ascent** (or gradient descent if we're trying to find a minimum), in which we take steps in the direction of the gradient until the value of the function stops increasing.

Exercises

1. Given the following matrix, A:

$$\begin{pmatrix} -1 & 1 \\ 0 & 2 \end{pmatrix}$$

   find $(A^{\mathrm{T}}A)^{-1}$. Show that any matrix $A^{\mathrm{T}}A$ is always symmetric.

2. Given the equation for a hyperplane, $\mathbf{w} \cdot \mathbf{x} + b$, find the equation for the unit normal vector to the plane.

3. Derive the formula for distance from a plane to an arbitrary point in R^3. Generalize to R^n.

4. You are given the following equation:

$$L(\mathbf{w}) = \sum_{j=1}^{N} (y_j - (\mathbf{w}^{\mathrm{T}}\mathbf{x}_j))^2$$

   where each y_j and $\mathbf{x}_j$ is constant, and $\mathbf{x}_j$ and $\mathbf{w}$ are vectors with the same number k of components. Find $\frac{\partial L}{\partial w_i}$, where w_i is the ith component of $\mathbf{w}$.

5. Apply gradient descent to find the minimum of the function:

$$f(x) = (x - 3)^2$$

   starting at $x = 0$ and with $\alpha = \frac{1}{3}$. Do this again for $\alpha = 1$.