LEAST SQUARES

1. Introduction to Least Squares

We are now at the final topic that we'll be reviewing in Linear Algebra: The solution for the equation $\mathbf{Ax} = \mathbf{b}$ when we have an over-constrained system. That is, we have too many equations and the system of equations has no solution that satisfies all these. We'll examine what's the best we can do in such a situation. This situation arises frequently in Data Analytics. For instance, we might have several dozen measurement variables but several thousand equations with those dozen or so variables. This is a fairly common setting in Linear Regression. For instance, think FICO$^{\text{TM}}$ score. The Credit Scoring company has several dozen measurements about any individual but has several million consumers that it has to grade for credit risk. The system is over-constrained. Further, the measurements are noisy. That is the constraints $\mathbf{b}$ include some noise. We can think of $\mathbf{b} = \mathbf{p} + \mathbf{e}$ where the measurement is impacted by some unknown noise that is impossible to remove from the measurement. How can we solve equations in such a setting? This is where Least Squares comes in. We'll strive to come up with the best solution that minimizes our error the most.

Before we look at Least Squares, we need to familiarize ourselves with a few more concepts such as the various sub-spaces that can be inferred from a general matrix $\mathbf{A}$ which leads neatly to projections that a matrix induces and finally to Least Squares. Like Singular Value Decomposition, this is a rich topic. We can look at Least Squares from at least 3 perspectives. At this time, we'll look at it from the perspective of Algebra and Geometry. Later in the course, we'll visit the same topic from the prospective of Calculus.

2. The Four Sub-spaces of a Matrix

We saw during factorization by Elimination that the rank of a matrix is the number of pivots. When we look at an $m \times n$ matrix $\mathbf{A}$, we can think of four sub-spaces connected with it: Two come from $\mathbf{A}$ and two from $\mathbf{A^T}$.

- The row space, represented as $C(\mathbf{A^T})$ which contains vectors $\mathbf{y}$ such that $\mathbf{A^T y} = \mathbf{c}$
- The column space, represented as $C(\mathbf{A})$ which contains vectors $\mathbf{x}$ that satisfy $\mathbf{Ax} = \mathbf{b}$
- The null space $N(\mathbf{A})$, the null-space of $\mathbf{A}$ which contain vectors that satisfy $\mathbf{Ax} = \mathbf{0}$ and
- The left null space $N(\mathbf{A^T})$ which contain vectors that satisfy $\mathbf{A^T y} = \mathbf{0}$

We know that the row space and the column space both have dimension r, the rank of the matrix. From this it follows that the null space $N(\mathbf{A})$ has dimension $n - r$ and the

left null space $N(\mathbf{A^T})$ has dimension $m - r$, to make the full dimensions of the matrix we started with: $\mathbf{A}$.

3. Fundamental Theorem of Linear Algebra

This connection between the 4 sub-spaces forms Part I of the Fundamental Theorem of Linear Algebra: The column space and row space both have dimension r. The null spaces have dimensions $n - r$ and $m - r$.

Now let's look at the second part. We know that when two vectors are orthogonal, their dot product is zero. That is $\mathbf{v}.\mathbf{w} = \mathbf{v^T w} = \mathbf{0}$ for orthogonal vectors. They are at right angles with each other. In any given subspace, the basis vectors are orthogonal with respect to each other. Likewise, vectors that lie in two different sub-spaces of a given vector space are orthogonal with each other. Let's look at the null space of $\mathbf{A}$. By definition, the rows of $\mathbf{A}$ are orthogonal to the null space. Otherwise, it cannot be the null space of $\mathbf{A}$. Since any vector in the row space of $\mathbf{A}$ can be specified as a linear combination of the rows of $\mathbf{A}$, all vectors in this row space are orthogonal with the null space. Therefore the space spanned by the row vectors of $\mathbf{A}$ – the row space of $\mathbf{A}$ is orthogonal to the null space. Similarly, we can reason that the column space of $\mathbf{A}$ is orthogonal to the left null space or $N(\mathbf{A^T})$.

In other words, $N(\mathbf{A})$ and $C(\mathbf{A^T})$ are orthogonal sub-spaces of R^n. Likewise, $N(\mathbf{A^T})$ and $C(\mathbf{A})$ are orthogonal sub-space of R^m. Further, not only are they orthogonal with each other, they are also complementary. For an $m \times n$ matrix, the null space and the row space are complementary in R^n, since the row space has dimension r and the null space has dimension $n - r$. Similarly, the column space has dimension r and the left null space has dimension $m - r$. This brings us to the second part of the Fundamental Theorem.

$N(\mathbf{A})$ is the orthogonal complement of the row space $C(\mathbf{A^T})$ and $N(\mathbf{A^T})$ is the orthogonal complement of the column space $C(\mathbf{A})$.

This suggests that any vector $\mathbf{x}$ can be split into two components, $\mathbf{x_r}$, which is the row space component and $\mathbf{x_n}$, which is the null space component. $\mathbf{x} = \mathbf{x_r} + \mathbf{x_n}$. We know $\mathbf{A x_r} = \mathbf{b}$ and $\mathbf{A x_n} = \mathbf{0}$. Therefore, $\mathbf{A x} = \mathbf{b}$ when we put the two equations together. We know this because $\mathbf{b}$ is in the column space of $\mathbf{A}$ and we saw back in the very beginning that the system of equations can be viewed as a linear combination of columns of $\mathbf{A}$. Therefore, multiplying a vector $\mathbf{x}$ by the matrix $\mathbf{A}$ cannot do anything else but project the vector $\mathbf{x}$ into the column space of $\mathbf{A}$, where $\mathbf{b}$ comes from.

4. Sub-space Projections

When we compute $\mathbf{A x}$, we get a vector in the column space of $\mathbf{A}$ as can be seen from the column space formulation of the system of equations. Let's now formally define projections. When we project a vector $\mathbf{x}$ onto a line, the projection $\mathbf{p}$ is the part of the vector $\mathbf{x}$ that lies along the line. Likewise, when we project $\mathbf{x}$ onto a plane, the projection $\mathbf{p}$ is the part of $\mathbf{x}$ that lies in the plane. So, the vector $\mathbf{x}$ can be viewed as an addition of two vectors, $\mathbf{p}$, the projection onto a sub-space and $\mathbf{a}$ a vector that is orthogonal to that sub-space. In R^3, the vector $\mathbf{x}$ would be realized as the triple: (x_1, x_2, x_3). The projection onto the

z-axis would be the vector $(0, 0, x_3)$. Likewise, the projection onto the xy-plane would be $(x_1, x_2, 0)$. Note that $\mathbf{x}$ is the sum of two vectors: $\mathbf{x} = (x_1, x_2, 0) + (0, 0, x_3)$ and that the two parts are orthogonal with respect to each other.

When we project a vector $\mathbf{b}$ onto a sub-space, we are essentially finding the point in that subspace that is closest to the vector. This is the projected vector $\mathbf{p}$. It is easy to visualize using geometry that this projection vector is orthogonal to the difference vector between $\mathbf{b}$ and $\mathbf{p}$. Let's look at the difference $\mathbf{v} = \mathbf{b} - \mathbf{p}$. If this were not orthogonal to $\mathbf{p}$ and the sub-space, then we can always split this difference into two components: $\mathbf{v'}$ that lies in the sub-space and $\mathbf{v}^{\perp}$, which is orthogonal to the sub-space. Consider the product $\mathbf{p}.\mathbf{v}$. This product is $\mathbf{p}.\mathbf{v'} + \mathbf{p}.\mathbf{v}^{\perp}$ and it minimizes only when $\mathbf{p}.\mathbf{v'}$ is zero. This happens only when $\mathbf{v} = \mathbf{v}^{\perp}$. For all other choices of the difference vector, there is a small component $\mathbf{v'}$ along the sub-space which is non-zero and therefore results in a projection that is not the closest it can be in the sub-space. Since by definition, the projection vector is the closest vector to $\mathbf{b}$ in the sub-space, the difference vector has to be orthogonal to $\mathbf{p}$.

Let's start with the n column vectors of matrix $\mathbf{A} = (\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n})$. When we want to find the projection of $\mathbf{b}$ in the sub-space spanned by these column vectors, we are interested in finding a vector $\mathbf{p}$ such that $\mathbf{p} = \hat{x_1}\mathbf{a_1} + \hat{x_2}\mathbf{a_2} + \ldots + \hat{x_n}\mathbf{a_n}$ and this $\mathbf{p}$ is closest to $\mathbf{b}$. The error vector $\mathbf{e} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$ is perpendicular to the sub-space, as the projection has to be closest vector to $\mathbf{b}$ in the column space of $\mathbf{A}$.

Since the projection vector $\mathbf{p}$ is orthogonal to the difference $\mathbf{e} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$, it implies that any vector in the column space of $\mathbf{A}$ is orthogonal to $\mathbf{e}$. That is, $\mathbf{a_i^T e} = \mathbf{0}\ \forall i \in (1, n)$. In other words, $\mathbf{A^T e} = \mathbf{A^T}(\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}) = \mathbf{0}$. Moving the terms around, we get $\mathbf{A^T A}\hat{\mathbf{x}} = \mathbf{A^T b}$ where $\hat{\mathbf{x}}$ is the best linear combination vector that minimizes the difference $\mathbf{e}$. In other words, the projection of $\mathbf{b}$ in the column space of $\mathbf{A}$ is given by $\mathbf{A}\hat{\mathbf{x}}$. These are equivalent statements.

5. Least Squares Approximations

Let's examine the case $\mathbf{Ax} = \mathbf{b}$ when the system of equations has no solutions. As we saw earlier, the frequent reason for this case is that there are too many equations and not enough unknowns. In practical situations, there is also the case where the measurement $\mathbf{b}$ is noisy and the noise vector can not be modeled completely. This implies that $\mathbf{b}$ lies outside the column space of $\mathbf{A}$. Otherwise, we could have solved the system of equations. In other words, there exists a non-zero error $\mathbf{e} = \mathbf{b} - \mathbf{Ax}$. We cannot solve for $\mathbf{x}$ exactly. But, we can find the best $\hat{\mathbf{x}}$ that minimizes the error. We call it the *least squares* solution.

We saw earlier that the closest vector to $\mathbf{b}$ is the projection vector $\mathbf{p} = \mathbf{A}\hat{\mathbf{x}}$ as this minimizes the error $\mathbf{e} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$. The $\hat{\mathbf{x}}$ that minimizes the error is the *least squares* solution. Let's see why it called so. Recall that

$$\mathbf{Ax} = \mathbf{b} = \mathbf{p} + \mathbf{e} \tag{1}$$

The squared length of the difference $\mathbf{Ax} - \mathbf{b}$ is

$$\|\mathbf{Ax} - \mathbf{b}\|^2 = \|\mathbf{Ax} - \mathbf{p}\|^2 + \|\mathbf{e}\|^2 \tag{2}$$

We know that when we choose $\mathbf{x} = \hat{\mathbf{x}}$, the first term in the difference vanishes leaving us only with the orthogonal error term $\|\mathbf{e}\|^2$. This is the smallest error we can get. When we get the smallest error, the *squared* length of $\mathbf{Ax} - \mathbf{b}$ is minimized. Hence the *least squares* name. The least squares solution for an unsolvable system of equations is given by $\hat{\mathbf{x}}$ that makes the error as small as possible. From before, we know that this is given by the vector $\hat{\mathbf{x}}$ that solves the modified equation $\mathbf{A^T A\hat{x}} = \mathbf{A^T b}$.

When we find the least squares solution to the problem, we are finding the solution to the modified problem where $\mathbf{A}$ is pre-multiplied by $\mathbf{A^T}$ and from projections, we know that this will result in the solution that minimizes the distance between $\mathbf{b}$ and the column space of $\mathbf{A}$.

This is the algebraic approach to least squares. Later, we'll visit least squares from the perspective of calculus and see that they are completely equivalent formulations.