University of California, Berkeley Lecturer: Benjamin Recht

CS189 Introduction to Machine Learning

Spring 2018 Notes: Benjamin Recht

Some notes on pseudoinverses

Let $\boldsymbol{A}$ be a square, full rank, $d \times d$ dimensional matrix. We all know that the inverse of $\boldsymbol{A}$ is the matrix $\boldsymbol{B}$ such that $\boldsymbol{AB} = \boldsymbol{I}$. What's the natural notion of an "inverse" for a matrix that is not square or possibly not full rank?

Let $\boldsymbol{X}$ be a $n \times d$ dimensional matrix and assume that $n \geq d$. We say that a $d \times n$ dimensional $\boldsymbol{Z}$ is a *pseudoinverse* of $\boldsymbol{X}$ if the following properties hold

1. $\boldsymbol{XZX} = \boldsymbol{X}$.
2. $\boldsymbol{ZXZ} = \boldsymbol{Z}$.
3. $\boldsymbol{XZ}$ and $\boldsymbol{ZX}$ are both symmetric matrices.

We denote the pseudoinverse by $\boldsymbol{X}^\dagger$.

Note that all of these properties are trivially satisfied by the matrix inverse. So for a square matrix with full rank $\boldsymbol{A}^\dagger = \boldsymbol{A}$.

Let's first assume that $\boldsymbol{X}$ is full rank. Then the $\boldsymbol{X}^\dagger = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. To verify this, observe

$$\boldsymbol{X}\boldsymbol{X}^\dagger\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}\,.$$

Similarly,

$$\boldsymbol{X}^\dagger\boldsymbol{X}\boldsymbol{X}^\dagger = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{X}^\dagger\,.$$

Finally, we can see that

$$\begin{aligned}\boldsymbol{X}^\dagger\boldsymbol{X} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}\\ \boldsymbol{X}\boldsymbol{X}^\dagger &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,,\end{aligned}$$

which are both symmetric matrices. Interestingly, in the case that $\boldsymbol{X}$ is full rank, multiplying $\boldsymbol{X}$ by the pseudoinverse on the left yields the identity matrix, just as a normal inverse would. On the other hand, since $\boldsymbol{X}\boldsymbol{X}^\dagger$ is $n \times n$ it is necessarily rank deficient. In this case, $\boldsymbol{X}\boldsymbol{X}^\dagger$ is the projection onto the range of $\boldsymbol{X}$.

What about when the matrix is rank deficient? The pseudoinverse still exists and we can derive it using the singular value decomposition. Let

$$\boldsymbol{X} = \boldsymbol{USV}^T = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

be a singular value decomposition of $\boldsymbol{X}$ and

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$$

be the singular values. Let k denote the index of the *smallest* nonzero singular value. That is, for $j > k$, we should have$\sigma_j = 0$. If $\boldsymbol{X}$ is full rank, then $k = d$.

Now define

$$\boldsymbol{Z} := \sum_{j=1}^{k} \sigma_j^{-1} \boldsymbol{v}_j \boldsymbol{u}_j^T .$$

I claim that this matrix $\boldsymbol{Z}$ satisfies all of the properties of a pseudoinverse. To see this, let's just work with matrices. Define the $d \times n$ matrix $\boldsymbol{R}$ to be

$$\boldsymbol{R} = \left[\begin{array}{cc} \operatorname{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_k^{-1}, 0, \ldots, 0) & \mathbf{0}_{d\times(n-d)} \end{array} \right] .$$

Note that $\boldsymbol{R}$ is the pseudoinverse of $\boldsymbol{S}$! Moreover,

$$\begin{aligned} \boldsymbol{X}\boldsymbol{Z}\boldsymbol{X} &= \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{R}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T \\ &= \boldsymbol{U}\boldsymbol{S}\boldsymbol{R}\boldsymbol{S}\boldsymbol{V}^T = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T . \end{aligned}$$

The other identities can be similarly checked. Now, in this case, the one particularly interesting identity is

$$\boldsymbol{Z}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{R}\boldsymbol{S}\boldsymbol{V}^T = \sum_{i=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i^T$$

In this case, $\boldsymbol{X}^\dagger \boldsymbol{X}$ does not equal the identity matrix, but is instead the projection operator onto the range of $\boldsymbol{X}^T$.

Finally, consider the context of least-squares. In the case that $\boldsymbol{X}$ is rank deficient, the least-squares solution is not unique. However, we can define a unique solution in the range of $\boldsymbol{X}^T$ using the pseudoinverse. If we define $\boldsymbol{w}_{\text{pinv}} = \boldsymbol{X}^\dagger \boldsymbol{y}$ this corresponds to the vector

$$\boldsymbol{w}_{\text{pinv}} = \sum_{i=1}^{k} \sigma_i^{-1} \langle \boldsymbol{u}_i, \boldsymbol{y} \rangle \boldsymbol{v}_i$$

where k again denotes the index of the smallest, nonzero singular value. This is precisely the least-squares solution when we restrict the possible $\boldsymbol{w}$ to lie in the span of the data (i.e., the range of $\boldsymbol{X}^T$).