CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS 6

1 Least-Squares Regression: Solving Normal Equations

In linear regression, we seek a model that captures a linear relationship between input data and output data. The simplest variant is the least-squares formulation. In this scenario, we are given a data matrix $X \in \mathbb{R}^{n \times d}$, where each row represents a datapoint $X_i \in \mathbb{R}^d$. We are also given an associated vector of output values $y \in \mathbb{R}^n$. We define the problem to be

$$\arg\min_w \|Xw - y\|^2.$$

By finding the gradient of the objective equation (with respect to w) and setting it equal to zero, we arrive at the normal equations

$$X^\top X\hat{w} = X^\top y.$$

Solving these normal equations will yield an optimal choice of weights $\hat{w}$ for our linear model. One question that arises is, is there always a solution to the problem? When is the solution unique? We will answer these questions in this exercise.

(a) Prove that a solution always exists to the normal equations, regardless of the choice of X and y. **Hint**: Consider the normal equations to be a usual matrix-vector system of equations, $Aw = b$. What is the range of values that the right-hand side can take on? What about the left-hand side?

Solution: Recall from discussion 2 that for any matrix X, we have that $\text{range}(X^\top X) = \text{range}(X^\top)$. Now, we know that the right side of the normal equations is $z = X^\top y$, which is by definition in the range of $X^\top$. It must therefore also be in the range of $X^\top X$. By definition, this means that there exists a vector $\hat{w}$ such that $X^\top X\hat{w} = z$, so we have shown that a solution to the normal equations does exist. Here, we made no assumptions about X or y.

(b) When the matrix $X^\top X$ is invertible, there exists a unique solution ($\hat{w} = (X^\top X)^{-1}X^\top y$). What conditions need to be true about $X^\top X$ and X for this statement to be true? Express your answer in terms of *rank*.

Solution: When $X^\top X$ is invertible, it must be a full rank matrix. Since $X^\top X \in \mathbb{R}^{d \times d}$, we must have that $\text{rank}(X^\top X) = d$. By extension, since $\text{range}(X^\top X) = \text{range}(X^\top)$, we must have that $\text{rank}(X^\top X) = \text{rank}(X^\top) = d$. The rank of a matrix and its transpose are the same, so we must have that $\text{rank}(X) = d$.

(c) If the matrix $X^\top X$ is *not* invertible, there will be infinitely many solutions to the normal equations. One such solution can be defined in terms of the *Moore–Penrose pseudoinverse* of the matrix X.

We define the pseudoinverse of A to be the matrix

$$A^+ = V\Sigma^+ U^\top = \sum_{i:\sigma_i>0} \frac{1}{\sigma_i} v_i u_i^\top,$$

where Σ^+ is computed from Σ by taking the transpose A and inverting the nonzero singular values on the diagonal.

Verify that $\hat{w} = X^+ y$ is a solution to the normal equations

Solution: We simply plugin $w = X^+ y$ to verify that it is a solution.

$$\begin{aligned} X^\top X w &= X^\top X X^+ y = (U\Sigma V^\top)^\top U\Sigma V^\top V\Sigma^+ U^\top y \\ &= V\Sigma^2 V^\top V\Sigma^+ U^\top y = V\Sigma U^\top y = X^\top y \end{aligned}$$

So, the pseudo-inverse does give a solution! Note that there is some funny shape business with the Σ matrix. You should be able to convince yourself that everything works out when we take the transposes etc. as the main components of the matrix are diagonal.
A little more in depth about the SVD manipulations here: (in this derivation, we ignore the components of u_i and v_i that have $\sigma_i = 0$, this is called the **Compact SVD**) we have that $X \in \mathbb{R}^{n\times d}$, which implies that we can write the singular value decomposition in terms of $U_r \in \mathbb{R}^{n\times r}$, orthonormal $V_r \in R^{d\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}$ (where r is the rank of the matrix). Observe that U_r and V_r are orthonormal matrices truncated to the first r columns) . Note that this allows us to write $A = U\Sigma V^T$. Now, we can write $A^+ = V\Sigma^+ U$, where $\Sigma^+ = \text{diag}(\frac{1}{\sigma_1}, ...\frac{1}{\sigma_d})$.

2 The Least-Norm Solution of a Least-Squares Problem

Some least-squares linear regression problems are under-determined and have infinitely many solutions. In the last problem, we showed that the pseudo-inverse provided one such solution, but we don't want just any solution to this system.

In this problem, our goal is to provide an explicit expression for the *least-norm* least-squares estimator, defined to be

$$\widehat{w}_{LS,LN} = \arg\min_{w}\{\|w\|^2 : w \text{ is a minimizer of } \|Xw - y\|^2\},$$

where $X \in \mathbb{R}^{n\times d}$, $w \in \mathbb{R}^d$, and $y \in \mathbb{R}^n$.

(a) Show that there exists a solution to the least-squares problem (to minimize $\|Xw - y\|^2$) that lies in the rowspace of X. **Hint:** Use the normal equations and the fundamental theorem of linear algebra.

Solution: The minimizers of the least-squares objective are the solutions w to the normal equations

$$X^\top Xw = X^\top y.$$

By the fundamental theorem of linear algebra, we know that $\text{nullspace}(X^T X) = \text{nullspace}(X) \perp \text{rowspace}(X) = \text{range}(X^\top) = \text{range}(X^\top X)$. As these spaces are orthogonal, for any single solution $\overline{w}$, we can write $\overline{w} = w_0 + \Delta$, where $w_0 \in \text{rowspace}(X)$ and $\Delta \in \text{nullspace}(X^\top X)$.

We can then observe that the Δ component contributes nothing to the solution as by the argument above Δ is in the nullspace of $X^\top X$.

$$X^\top X\overline{w} = X^\top X(w_0 - \Delta) = X^\top Xw_0 = X^\top y,$$

Thus, w_0, the component of the solution $\overline{w}$ in the rowspace, is also a solution to the normal equations and thus a minimizer of the least squares objective.

(b) Show that the solution w_0 in the rowspace is unique.

Solution: Assume there exists some other solution $w_1 \in \text{rowspace}(X)$. Then by definition $X^\top Xw_1 = X^\top y$ as well. By subtracting both solutions from eachother, we have:

$$X^\top Xw_0 - X^\top Xw_1 = X^\top X(w_0 - w_1) = X^\top y - X^\top y = 0$$

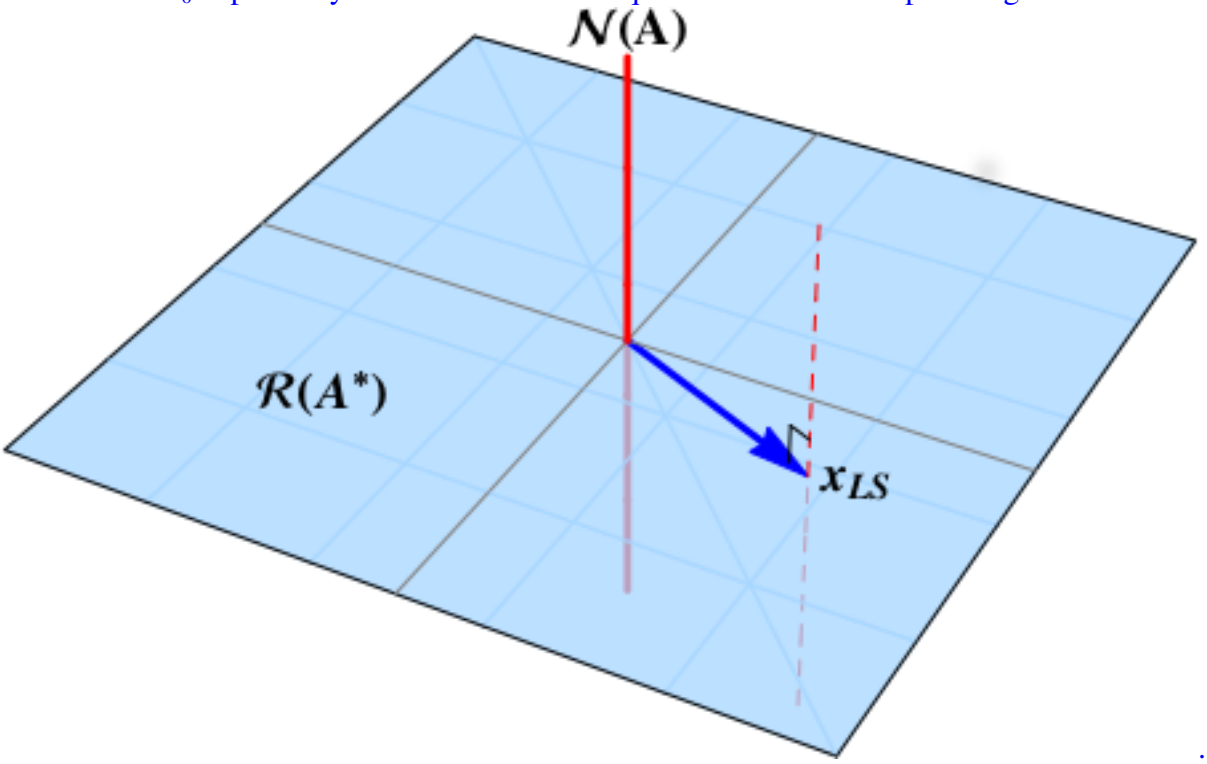
This means that $w_0 - w_1$ is by definition in the $\text{nullspace}(X^\top X) = \text{nullspace}(X)$. However, $w_0 - w_1$ is a linear combination of vectors in the row space of X, and is thus in the rowspace of X as well! This means that $w_0 - w_1 \in \text{nullspace}(X)$ and $w_0 - w_1 \in \text{range}(X^\top)$. But, these two spaces are orthogonal, and thus the only element they share is the zero vector. Consequently, it must be that $w_0 - w_1 = 0$, or $w_0 = w_1$.

(c) Show that the solution we identified in part (a) is in fact the solution with the smallest ℓ_2 norm (i.e., the solution to the least norm problem $\widehat{w}_{LS,LN}$).

Solution: We will show that w_0 is the solution to the least norm problem. Note that all possible solutions to the least squares problem are of the form $w_0 + \Delta$, where $\Delta \in \text{nullspace}(X)$. Thus, for any other minimizer $w = w_0 + \Delta$,

$$\begin{aligned} \|w\|^2 &= \|w_0 + \Delta\|^2 \\ &= \|w_0\|^2 + \|\Delta\|^2 + 2w_0^\top \Delta \\ &= \|w_0\|^2 + \|\Delta\|^2, \end{aligned}$$

where we use the fact that $w_0 \perp \Delta$, because the nullspace and rowspace of X are orthogonal. Hence, we conclude that $\|w\|^2$ is strictly greater than $\|w_0\|^2$ unless $\Delta = 0$, i.e., $w = w_0$. It follows that w_0 is precisely the least norm least-squares solution. A helpful diagram is this:



We see that the set of least squares solutions is an affine subspace parallel to the nullspace of A (represented by the red-dashed line). By the Pythagorean Theorem, we see that choosing a least squares solution that is not orthogonal to the nullspace can only increase the norm of the solution (which is the distance to the origin).

(d) Show that $\widehat{w}_{LS,LN}$ is the pseudoinverse solution (from the last problem)

$$\widehat{w}_{LS,LN} = X^+ y = \sum_{i:\sigma_i > 0} \frac{1}{\sigma_i} v_i (u_i^\top y).$$

In problem 1 we showed that the pseudo inverse was a solution to the normal equations. In part a) of this quesiton, we showed that there was only one solution to the normal equations in

the rowspace of X and in part b) that this solution in the rowspace is the solution of least norm. Thus, if the pseudo-inverse is in the rowspace of X it is the solution of least norm. Show this directly by checking that the above expression for $\widehat{w}_{LS,LN}$ is in the rowspace of X.

Solution: The singular value decomposition of X is

$$X = \sum_{i:\sigma_i>0} \sigma_i u_i v_i^\top.$$

Notice that the set $\{v_i : \sigma_i > 0\}$ define an orthonormal basis for the rowspace of X, as weightings of the v_i's define the rows of X. We can prove this formally by showing that the v_i's are orthogonal to any vector in the null space of X, and are thus all in rowspace of X. Note that by the construction of SVD, the v_i's are already assumed to be orthonormal.

Consider an arbitrary element $z \in \text{nullspace}(X)$. We have that

$$Xz = \sum_{i:\sigma_i>0} \sigma_i u_i (v_i^\top z) = 0.$$

Since the $\sigma_i u_i$ are all linearly independent, $Xz = 0$ if and only if $v_i^\top z = 0$ for all i. Therefore, z is orthogonal to $\{v_i : \sigma_i > 0\}$, and we have thus shown this set is the orthogonal complement to the nullspace.

To see that $\widehat{w}_{LS,LN}$ is in the rowspace of X, observe that $\widehat{w}_{LS,LN}$ is a linear combination of v_i for $i : \sigma_i > 0$.

3 Softmax Regression

Logistic regression directly models the probability of a data point x belonging to class 1, or $P(Y = 1|X = x) = \mathbf{g}(w^\top x)$ where g is the sigmoid function $g(z) = \frac{1}{1+e^(-z)}$. This is however limited to modeling binary classification problems. While logistic regression can be extended to the multi-class setting using many-to-one or one-to-one approaches, there exists a more elegant solution.

Rather than only modeling $P(Y = 1|X = x)$, softmax regression models the entire categorical distribution over k classes, $P(Y = 1|X = x), P(Y = 2|X = x), ..., P(Y = k|X = x)$. It does so by leveraging a different linear model w_i (weight vectors) for each of the k classes and the softmax function, $s(z)_i = \frac{e^{-z_i}}{\sum_{j=1}^{k} e^{-z_j}}$. Concretely:

$$P(Y = i|X = x) = \frac{e^{-w_i^\top x}}{\sum_{j=1}^{k} e^{-w_j^\top x}}$$

This essentially assumes each classes probability is proportional to $e^{-w_i^\top x}$ and normalizes by the sum of total values.

(a) Show that in the case where $k = 2$, softmax regression is the same as logistic regression.

Solution:

$$P(Y = 1|X = x) = \frac{e^{-w_1^\top x}}{e^{-w_1^\top x} + e^{-w_2^\top x}} = \frac{1}{1 + e^{-(w_2 - w_1)^\top x}}$$

(b) In its default form given above, softmax regression is actually overparameterized – there are more parameters than needed for the same model. This should be evident in your answer to part a). Reformulate softmax regression such that it requires fewer parameters.

Solution: We can divide out by one of the classes to remove it from the equation.

$$P(Y = k|X = x) = \frac{e^{-w_k^\top x}}{\sum_{j=1}^{k} e^{-w_j^\top x}} = \frac{1}{1 + \sum_{j=1}^{k-1} e^{-(w_j - w_k)^\top x}}$$

Thus, we only need to learn weights for $k - 1$ of the k classes. The implicit weights in the above equation are $w_j - w_k$.

(c) Recall binary cross-entropy loss:

$$L(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

How would you design the analogous loss function for softmax regression?

Solution: The generalization is just called cross-entropy loss!

$$L(\hat{y}, y) = \sum_{i=1}^{k} -1\{y = i\} \log \hat{y}_i$$

where $\hat{y}$ is a vector of the predicted probabilities such that $\hat{y}_i = P(Y = i|X = x)$.