

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

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

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

- (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 $X^\top X$.

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

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 row space of X . **Hint:** Use the normal equations and the fundamental theorem of linear algebra.

- (b) Show that the solution w_0 in the row space is unique.

- (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 $\widehat{w}_{LS, LN}$).

- (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 question, we showed that there was only one solution to the normal equations in the row space of X and in part b) that this solution in the row space is the solution of least norm. Thus, if the pseudo-inverse is in the row space 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 row space of X .

3 Softmax Regression

Logistic regression directly models the probability of a data point x belonging to class 1, or $P(Y = 1|X = x) = g(w^\top x)$ 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, $P(Y = 1|X = x), P(Y = 2|X = x), \dots, P(Y = k|X = x)$. It does so by leveraging a different linear model w_i 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.
- (b) In it's 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.
- (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?