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 design 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 $y \in \mathbb{R}^n$. Our problem is to

$$\text{find } w \text{ that minimizes } \|Xw - y\|^2.$$

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? Did you learn anything in Discussion 2 that might apply here?

(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 the *rank* of X.

(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.

You may or may not be familiar with the *singular value decomposition* (SVD) of a matrix. We will review it briefly late in the semester. For now, all you need to know is that **every** matrix A can be written in the form

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

where

- Σ is a diagonal matrix (but not necessarily square) with diagonal entries σ_i;
- The columns u_i of U have unit length and are pairwise orthogonal; that is, $U^\top U = I$;
- The columns v_i of V have unit length and are pairwise orthogonal; that is, $V^\top V = I$.

Note that none of U, Σ, nor V is necessarily square.

Give a diagonal matrix Σ, we define its *psuedoinverse* Σ^+ to be a diagonal matrix found by taking the transpose of Σ (if Σ is $i \times j$ then Σ^+ is $j \times i$) and then replacing every nonzero value σ_i on the diagonal with its reciprocal $1/\sigma_i$ (but every zero on the diagonal remains a zero). Observe that $\Sigma\Sigma^+$ and $\Sigma^+\Sigma$ are square, diagonal matrices with 1's and 0's on their diagonals. So Σ^+ is as close to an "inverse" of Σ as we can expect.

The *pseudoinverse* of a general matrix A with singular value decomposition $A = U\Sigma V^\top$ is the matrix

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

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 underdetermined and have infinitely many solutions. In the last problem, we showed that the pseudoinverse 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 w_0 to the least-squares problem (to minimize $\|Xw - y\|^2$) that lies in the rowspace of X. **Hint:** Use the normal equations and the fact that the nullspace of X is orthogonal to the rowspace of X (the fundamental theorem of linear algebra).

(b) Show that the solution w_0 in the rowspace 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 Question 1 we showed that the pseudoinverse was a solution to the normal equations. In parts a) and b) of this question, we showed that there was only one solution to the normal equations in the rowspace of X, and in part c) we showed that this solution in the rowspace is the solution of least norm. Thus, if the pseudoinverse 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.

3 Softmax Regression

Logistic regression directly models the probability of a data point x belonging to class 1, or $P(Y = 1|X = x) = s(w^\top x)$ where s is the sigmoid function $s(\gamma) = 1/(1 + e^{-\gamma})$. 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, $s_i(z) = e^{-z_i}/(\sum_{j=1}^{k} e^{-z_j})$. Concretely,

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

This assumes each classes probability is proportional to $e^{-w_i^\top x}$. To make all the probabilities sum to 1, the denominator is the sum of the numerators.

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

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

(c) Recall the logistic (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?