CS 189 Introduction to Machine Learning
Spring 2021 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

$X \in R^{n\times d}$

$$n\left\{\begin{bmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_n^T - \end{bmatrix}\right.$$

X_1
X_2
$\vdots$
X_i

$X_i \longrightarrow \hat{y}$

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

$2X^TXw - 2X^Ty = 0$
$\Rightarrow X^TX\hat{w} = X^Ty$

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

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

$A\hat{w} = b$

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?

range(X^TX) = range(X^T) $\quad$ $b \in$ range(X^T), $X^TX\hat{w} \in$ range(X^T)

(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*.

$\hat{w} = (X^TX)^{-1}X^Ty$

X has full row-rank

rank$(X) = d$

$\underbrace{X^TX}_{d\times d}$ & rank d $\Rightarrow$ invertible

rank$(X) < d$
X^TX non-inv

(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

$A \in \mathbb{R}^{m \times n}$

$A^+ \in \mathbb{R}^{n \times m}$

$A = U\Sigma V^T$

$X^TX = V\Lambda V^T$

$(X^TX)^{-1} = V\Lambda^{-1}V^T$

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

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

$\boxed{\text{SVD}}$

$w = X^+y$

$X^TXw = X^Ty$ $\quad U^TU = I$

$X^TXX^+y = (V\Sigma^T\cancel{U^T})(\cancel{U}\Sigma \cancel{V^T})(\cancel{V}\Sigma^+U^T)y$

$= V\Sigma^T\Sigma\Sigma^+U^Ty$

$= V\Sigma^TU^Ty$

$= X^Ty$

$$A = U\begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} V^T$$

$$A^{-1} = \underbrace{V\begin{bmatrix} 1/\sigma_1 & & & & \\ & 1/\sigma_2 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} U^T}_{\Sigma^+}$$

$$\Sigma^T \begin{bmatrix} \sigma_1 & & & & & 0 \\ & \sigma_2 & & & & 0 \\ & & \sigma_3 & & & 0 \\ & & & 0 & & 0 \\ & & & & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & & & & \\ & 1/\sigma_2 & & & \\ & & 1/\sigma_3 & & \\ & & & 0 & \\ & & & & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$= \Sigma^T \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} = \Sigma^T$$

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

$$\widehat{w}_{LS,LN} = \arg\min_{w}\{\|w\|_2^2 : w \text{ is a minimizer of } \underbrace{\|Xw-y\|^2}_{W}\},$$

$$W = \left\{ \begin{bmatrix} 1 \\ 2 \\ x \end{bmatrix} : x \in \mathbb{R} \right\}$$

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

$$w_{LS,LN} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

(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.

$$\text{null}(X^TX) = \text{null}(X) \perp \text{row}(X) = \text{range}(X^T) = \text{range}(X^TX)$$

$$X^TXw = X^Ty$$

$$\text{null}(X) \perp \text{range}(X^TX)$$

$$w = \underset{\in\,\text{row}(X)}{w_0} + \underset{\in\,\text{null}(X^TX)}{\Delta} \quad (\Delta \to 0)$$

$$X^TX(w_0+\Delta) = X^TXw_0 + \cancel{X^TX\Delta} = X^TXw_0 = X^Ty$$

$$X^TXw = X^Ty$$
$$\Downarrow$$
$$X^TXw_0 = X^Ty$$
$$\hookrightarrow \in \text{row}(X)$$

$$\|w\|_2^2 = \|w_0\|^2 + \|\Delta\|^2 + 2\cancel{w_0^T\Delta}$$
$$\geq \|w_0\|^2$$

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

$$w_0' \neq w_0 \ : \ w_0' \in \text{row}(X),\ X^TXw_0' = X^Ty$$

$$X^TXw_0' = X^Ty$$
$$-\ X^TXw_0 = X^Ty$$
$$X^TX(w_0'-w_0) = 0 \Rightarrow (w_0'-w_0) \in \text{null}(X) \quad \text{— (1)}$$

$$w_0',\ w_0 \in \text{row}(X)$$
$$\Rightarrow w_0' - w_0 \in \text{row}(X) \quad \text{— (2)}$$

$$\Rightarrow w_0' - w_0 = 0$$

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

$$\|w\|_2^2 = \|w_0\|^2 + \|\Delta\|^2 + \cancel{2w_0^T\Delta} \to 0$$
$$\geq \|w_0\|^2$$

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

$$\underbrace{\widehat{w}_{LS,LN}} = X^+y = \sum_{i:\sigma_i>0} \frac{1}{\sigma_i} \overset{\text{vector}}{v_i} \underbrace{(u_i^\top y)}_{\text{scalar}}. \quad = \text{some lin comb of the basis vecs of } row(X)$$

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.

need to show

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

$\{v_i : \sigma_i > 0\}$ forms an orthonormal basis for $row(X)$.

$z \in null(X)$

$$Xv = \sum_{i:\sigma_i>0} \sigma_i u_i (v_i^\top z) = 0 \qquad v_i^\top z = 0 \quad \forall i : \sigma_i > 0$$

$\{v_i : \sigma_i > 0\} \perp null(X)$

3 Softmax Regression

Logistic regression directly models the probability of a data point x belonging to class 1, or $P(Y = 1|X = x) = \mathbf{g}(w^\top x)$ 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, $P(Y = 1|X = x), P(Y = 1|X = x), ..., P(Y = k|X = x)$. 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?