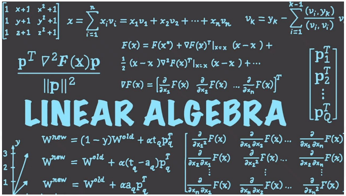
Linear Algebra, Optimization, and Linear Regression

Dana Golden, Lilia Maliar



Data Science and Machine Learning - **November 30, 2024**

Presentation Outline

1. Introduction and Background
2. Review of Linear Algebra
3. Optimization
4. Linear Regression
5. Conclusion

Why linear algebra is important?

- Linear algebra is at the heart of machine learning
- Many advanced linear algebra techniques are important to machine learning algorithms
- Matrices are how computers make sense of data

Why optimization is important?

- Most machine learning frameworks focus on optimization
- As economists, we often want to view algorithms through the lens of optimization

Why re-introduce linear regression?

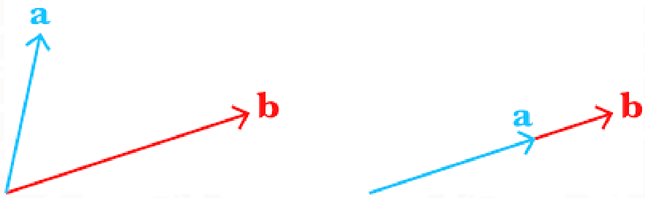
- Machine learning view on linear regression focuses on optimization
- Linear regression is a common framework in econometrics and provides a lens through which to see machine learning
- Most undergrad econometric classes don't focus on matrix algebra

Matrix Multiplication

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & a_9 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix}$$

Linear Independence

- A set of vectors $\{v_i\}_{i=1}^{n}$ is linearly independent if the vector equation $x_1 v_1 ... x_n v_n = 0$ has only the trivial solution x=0



Linear Independence Example

- Are the following vectors linearly independent?

$$\begin{bmatrix} 2 & -4 & 1 \\ 2 & 6 & 0 \\ 1 & 5 & 0 \end{bmatrix} \tag{1}$$

Linear Independence Example

- Are the following vectors linearly independent?

$$\begin{bmatrix} 2 & -4 & 1 \\ 2 & 6 & 0 \\ 1 & 5 & 0 \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} 2 & 15 & 3 \\ 5 & 7 & 9 \\ 4 & 30 & 6 \end{bmatrix} \tag{2}$$

Rank

- A matrix's rank is the number of linearly independent rows
- The rank of a matrix can be found by row-reducing and finding number of pivot points
- Only matrices of full rank are invertible. Why is this important?

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{2R_1+R_2\to R_2} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{-3R_1+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & -1 & -3 \end{bmatrix}$$

$$\xrightarrow{R_2+R_3\to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{-2R_2+R_1\to R_1} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

Inverse Definition

- A square matrix's inverse is the matrix that when multiplied by the matrix is the identity
- While most matrix multiplication is not communitive, inverse multiplication is
- Singular matrices have no inverse

$$AA^{-1} = A^{-1}A = I \tag{3}$$

Finding an Inverse

- Method 1: Augment matrix with identity matrix, and row reduce original matrix while applying steps to augmented matrix
- Method 2: Multiple inverse of absolute value of determinant by adjoint matrix

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{13} & a_{12} \\ a_{33} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ \begin{vmatrix} a_{23} & a_{21} \\ a_{33} & a_{31} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & \begin{vmatrix} a_{13} & a_{11} \\ a_{23} & a_{21} \end{vmatrix} \\ \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{12} & a_{11} \\ a_{32} & a_{31} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{bmatrix}$$

Usefulness of Inverses

- Matrix inverses can be used to solve systems of equations
- Crucial for econometrics and specific machine learning tasks

Determinants

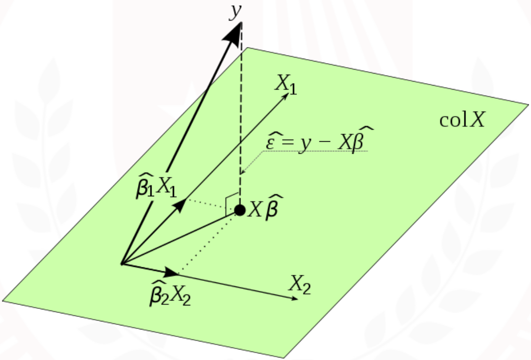
- Determinants have four properties:
    - The determinant of the identity matrix is 1
    - Exchange of two rows multiplies determinant by -1
    - Multiplying a row by a number multiplies the determinant by this number
    - Adding to a row a multiple of another row does not change the determinant

Eigenvalues and Eigenvectors

- Eigenvectors are vectors that when multiplied by a matrix produce themselves times a constant
- The constant is the eigenvalue
- Eigendecomposition is incredibly useful for PCA

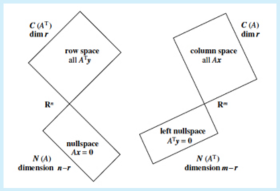
$$A\vec{v} = \vec{v} \tag{4}$$

Projection



Spaces

- a vector is said to be in space V if for scalar c, $c\vec{a} \in V$ and for $\vec{a} \in V$ and $\vec{b} \in V$, $\vec{a} + \vec{b} \in V$



Norms

- Norms have three properties
  - Subadditivity: $p(x+y) \leq p(x) + p(y) \forall x, y \in X$
  - Absolute homogeneity: $p(sx) = |s|p(x)$
  - Positive definiteness: $p(x) = 0 \Leftrightarrow x = 0$
- Why are these useful? What might a function that is a norm look like?

Norms

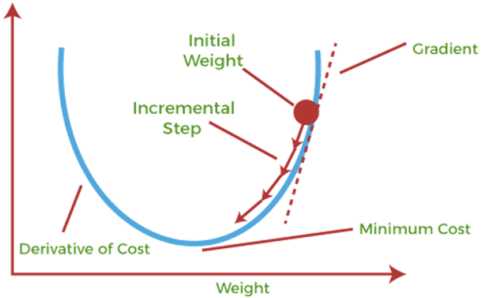
- Norms have three properties
  - Subadditivity: $p(x+y) \leq p(x) + p(y) \forall x, y \in X$
  - Absolute homogeneity: $p(sx) = |s|p(x)$
  - Positive definiteness: $p(x) = 0 \Leftrightarrow x = 0$
- Why are these useful? What might a function that is a norm look like?
- Euclidean Norm: $||x||_2 = \sqrt{x_1^2 + ... + x_n^2}$
- Taxicab Norm: $||x||_1 = \sum_{i=1}^{n} |x_i|$
- P-norm $||x||_p = (_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$

Analytic Optimization

- Analytic optimization is the most well known to economists
- It involves finding the maximum of a convex function
- Analytic optimization can only be done for functions with analytic maximums

Gradient Descent

Understanding Gradient Descent



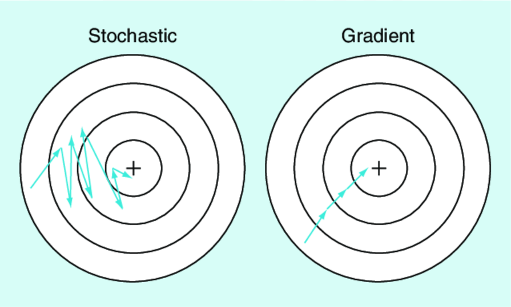
$$f(x) = x^2 - 4x + 3$$

x = x - learning_rate * grad

Stochastic Gradient Descent

- Take gradient for random observation i and take step in that direction

$$\theta' = \theta - \alpha \nabla f_i(\theta) \tag{5}$$



Newton's Method

$$\theta' = \theta - \frac{f'(\theta)}{f''(\theta)} \quad (6)$$

Deriving Least Squares with Matrix

$$y = \beta X + u \tag{7}$$

$$\min_{\beta} \Sigma_{t=1}^{T}[y_t - \Sigma_{i=1}^{n}\beta_1 x_{ti}]^2 \tag{8}$$

$$\min_{\beta} \Sigma_{i=1}^{N}[y - X\beta]^2 \tag{9}$$

Deriving Least Squares with Matrix

$$y = \beta X + u \tag{7}$$

$$\min_{\beta} \Sigma_{t=1}^{T}[y_t - \Sigma_{i=1}^{n}\beta_1 x_{ti}]^2 \tag{8}$$

$$\min_{\beta} \Sigma_{i=1}^{N}[y - X\beta]^2 \tag{9}$$

- Take the matrix derivative

$$X'(y - X\hat{\beta}) = 0 \tag{10}$$

Deriving Least Squares with Matrix

$$y = \beta X + u \tag{7}$$

$$\min_{\beta} \Sigma_{t=1}^{T}[y_t - \Sigma_{i=1}^{n}\beta_1 x_{ti}]^2 \tag{8}$$

$$\min_{\beta} \Sigma_{i=1}^{N}[y - X\beta]^2 \tag{9}$$

- Take the matrix derivative

$$X'(y - X\hat{\beta}) = 0 \tag{10}$$

$$X'y - X'X\hat{\beta} = 0 \tag{11}$$

$$X'y = X'X\hat{\beta} = 0 \tag{12}$$

$$\hat{\beta} = (X'X)^{-1}(X'y) \tag{13}$$

Least Squares Asymptotics

$$E(\hat{\beta}) = (X'X)^{-1}(X'y) = \tag{14}$$

$$(X'X)^{-1}(X'(X\beta + u)) = \tag{15}$$

$$(X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \tag{16}$$

$$\beta + (X'X)^{-1}X'u \tag{17}$$

- $(X'X)^{-1}X'u$ asymptotically goes to zero. Why?

Least Squares Asymptotics

$$E(\hat{\beta}) = (X'X)^{-1}(X'y) = \tag{14}$$

$$(X'X)^{-1}(X'(X\beta + u)) = \tag{15}$$

$$(X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \tag{16}$$

$$\beta + (X'X)^{-1}X'u \tag{17}$$

- $(X'X)^{-1}X'u$ asymptotically goes to zero. Why?

$$E(\hat{\beta}) = \beta \tag{18}$$

Least Squares Standard Error

$$D(\hat{\beta}) = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' = \tag{19}$$

Least Squares Standard Error

$$D(\hat{\beta}) = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' = \tag{19}$$

$$E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta}) = \tag{20}$$

Least Squares Standard Error

$$D(\hat{\beta}) = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' = \tag{19}$$

$$E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta}) = \tag{20}$$

$$E((X'X)^{-1}X'uu'X(X'X)^{-1}) = \tag{21}$$

Least Squares Standard Error

$$D(\hat{\beta}) = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' = \tag{19}$$

$$E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta}) = \tag{20}$$

$$E((X'X)^{-1}X'uu'X(X'X)^{-1}) = \tag{21}$$

$$(X'X)^{-1}X'E(uu')X(X'X)^{-1} = \tag{22}$$

Least Squares Standard Error

$$D(\hat{\beta}) = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' = \tag{19}$$

$$E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta}) = \tag{20}$$

$$E((X'X)^{-1}X'uu'X(X'X)^{-1}) = \tag{21}$$

$$(X'X)^{-1}X'E(uu')X(X'X)^{-1} = \tag{22}$$

$$(X'X)^{-1}X'\sigma^2X(X'X)^{-1} \tag{23}$$

Least Squares Standard Error

$$D(\hat{\beta}) = E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' = \tag{19}$$

$$E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta}) = \tag{20}$$

$$E((X'X)^{-1}X'uu'X(X'X)^{-1}) = \tag{21}$$

$$(X'X)^{-1}X'E(uu')X(X'X)^{-1} = \tag{22}$$

$$(X'X)^{-1}X'\sigma^2X(X'X)^{-1} \tag{23}$$

$$\sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \tag{24}$$

$$Var\hat{\beta}_i = \sigma^2(X'X)^{-1}_{ii} \tag{25}$$

Assumptions and Violations of Least Squares Asymptotics

- What happens if the x-values are correlated with the error term?

Assumptions and Violations of Least Squares Asymptotics

- What happens if the x-values are correlated with the error term?

$$E(\hat{\beta}) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \qquad (26)$$

$$\beta + (X'X)^{-1}X'u \neq \beta \qquad (27)$$

Assumptions and Violations of Least Squares Asymptotics

- What happens if the x-values are correlated with the error term?

$$E(\hat{\beta}) = (X'X)^{-1} X'X\beta + (X'X)^{-1} X'u = \tag{26}$$

$$\beta + (X'X)^{-1} X'u \neq \beta \tag{27}$$

- What happens if the y values are correlated with the error term?

Derivation of Maximum Likelihood Estimator of Least Squares

$$y_t = X_t\beta + u_t, u_t \sim iidN(0, \sigma^2) \tag{28}$$

$$L(y|\beta,\sigma) =_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma}} exp\{\frac{-1}{2\sigma}(y - X_t\beta)^2\} \tag{29}$$

$$ln(L(y|\beta,\sigma)) = \sum_{t=1}^{T} \frac{-1}{2} ln(2\pi) - ln(\sigma) - \frac{1}{2\sigma}(y_t - X_t\beta)^2 \tag{30}$$

- This is maximized by minimizing the sum of squared errors

Algorithm for Solving Least Squares using Maximum Likelihood

- Start with cost function
- Minimize
- How to find standard error?

Algorithm for Solving Least Squares using Maximum Likelihood

- Start with cost function
- Minimize
- How to find standard error?
- Hessian matrix/ Information matrix
- Monte Carlo

Cost Functions

- A function you attempt to minimize within the machine learning context
- A way to measure how well your algorithm is performed
- Example: MSE, log loss
- Generally make log loss negative. Why?

LASSO

- L1 Norm
- Used to choose variables and prevent overfitting
- Sets value of some coefficients to zero

$$min_{\beta_0,\beta_1}\{\Sigma_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\} s.t. \Sigma_{j=1}^{p}|\beta_j| \leq t \tag{31}$$

Ridge

- L2 norm
- Scales all coefficients based on their value for prediction
- Can perform regression even when colinearity exists

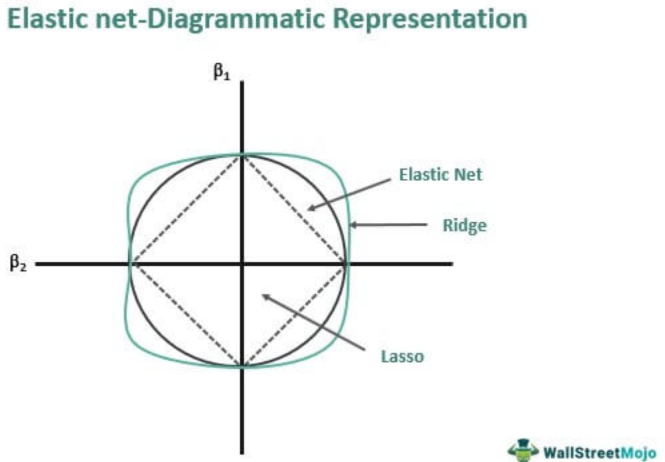
$$min\Sigma_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\}s.t.\lambda\Sigma_{j=1}^{p}|\beta_j^2| \leq t \tag{32}$$

Elastic Net

- Elastic Net uses penalties on both the L_1 and L_2 norm
- Compromise between Lasso and Ridge

$$min\Sigma_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\}s.t.\lambda_2||\beta||^2 \leq t_1, \lambda_1||\beta_1|| \leq t_2 \tag{33}$$

Visualization



Thank You So Much!

Sources