ECE595 / STAT598: Machine Learning I
Lecture 1.2: Geometry

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

Lecture 1: Linear Regression

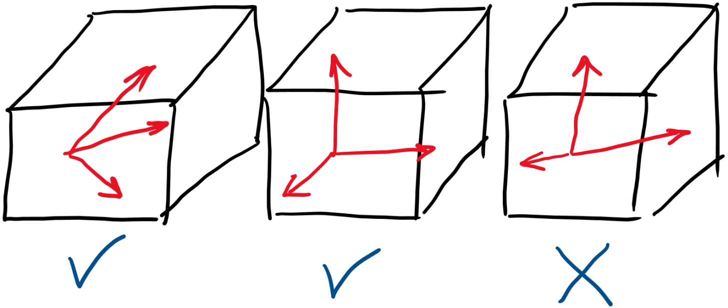
- Linear Regression
  - Notation
  - Loss Function
  - Solving the Regression Problem
- Geometry
  - Projection
  - Minimum-Norm Solution
  - Pseudo-Inverse

Linear Span

Given a set of vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d\}$, the **span** is the set of all possible linear combinations of these vectors.

$$\text{span}\left\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d\right\} = \left\{\boldsymbol{z} \mid \boldsymbol{z} = \sum_{j=1}^{d} \alpha_j \boldsymbol{a}_j\right\} \tag{1}$$

Which of the following sets of vectors can span $\mathbb{R}^3$?

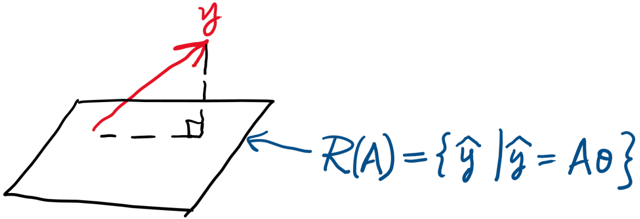


Geometry of Linear Regression

Given $\boldsymbol{\theta}$, the product $\boldsymbol{A\theta}$ can be viewed as

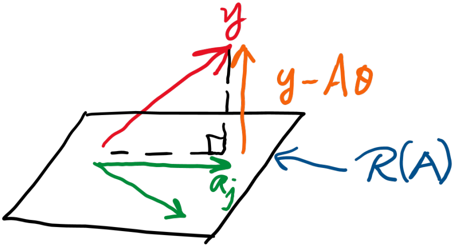
$$\boldsymbol{A\theta} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_d \\ | & | & & | \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} = \sum_{j=1}^{d} \theta_j \boldsymbol{a}_j.$$

So the set of all possible $\boldsymbol{A\theta}$'s is equivalent to $\text{span}\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_d\}$. Define the **range** of $\boldsymbol{A}$ as $\mathcal{R}(\boldsymbol{A}) = \{\widehat{\boldsymbol{y}} \mid \widehat{\boldsymbol{y}} = \boldsymbol{A\theta}\}$. Note that $\boldsymbol{y} \notin \mathcal{R}(\boldsymbol{A})$.



Orthogonality Principle

- Consider the error $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}$.
- For the error to minimize, it must be **orthogonal** to $\mathcal{R}(\boldsymbol{A})$, which is the span of the columns.
- This **orthogonality principle** means that $\boldsymbol{a}_j^T \boldsymbol{e} = 0$ for all $j = 1, \ldots, d$, which implies $\boldsymbol{A}^T \boldsymbol{e} = 0$.



Normal Equation

- The orthogonality principle, which states that $\boldsymbol{A}^T\boldsymbol{e} = \boldsymbol{0}$, implies that $\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}) = \boldsymbol{0}$ by substituting $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}$.
- This is called the **normal equation**:

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{A}^T\boldsymbol{y}. \tag{2}$$

- The predicted value is

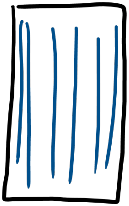
$$\widehat{\boldsymbol{y}} = \boldsymbol{A}\widehat{\boldsymbol{\theta}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

- The matrix $\boldsymbol{P} \overset{\text{def}}{=} \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ is a projection onto the span of $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d\}$, i.e., the range of $\boldsymbol{A}$.
- $\boldsymbol{P}$ is called the **projection matrix**. It holds that $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$.
- The error $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ is

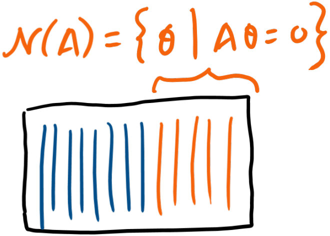
$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.\end{aligned}$$

Over-determined and Under-determined Systems

- Assume $\boldsymbol{A}$ has full column rank.
- Over-determined $\boldsymbol{A}$: Tall and skinny. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$.
- Under-determined $\boldsymbol{A}$: Fat and short. $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{y}$.
- If $\boldsymbol{A}$ is under-determined, then there exists a non-trivial **null space** $\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{\theta} \mid \boldsymbol{A}\boldsymbol{\theta} = 0\}$.
- This implies that if $\widehat{\boldsymbol{\theta}}$ is a solution, then $\widehat{\boldsymbol{\theta}} + \boldsymbol{\theta}_0$ is also a solution as long as $\boldsymbol{\theta}_0 \in \mathcal{N}(\boldsymbol{A})$. (Why?)



over-determined



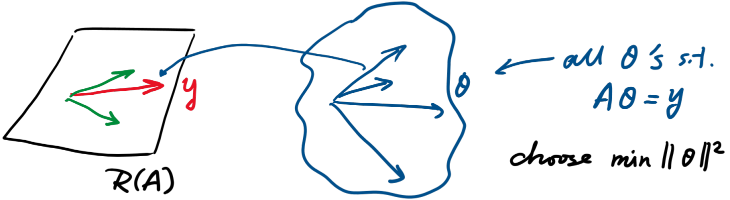
under-determined

Minimum-Norm Solution

- Assume $\boldsymbol{A}$ is fat and has full row rank.
- Since $\boldsymbol{A}$ is fat, there exists infinitely many $\widehat{\boldsymbol{\theta}}$ such that $\boldsymbol{A}\widehat{\boldsymbol{\theta}} = \boldsymbol{y}$.
- So we need to pick one in order to be unique.
- It turns out that $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{y}$ is the solution to

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \;\; \|\boldsymbol{\theta}\|^2 \;\; \text{subject to} \;\; \boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{y}. \tag{3}$$

  (You can solve this problem using Lagrange multiplier. See Appendix.)
- This is called the **minimum-norm** solution.

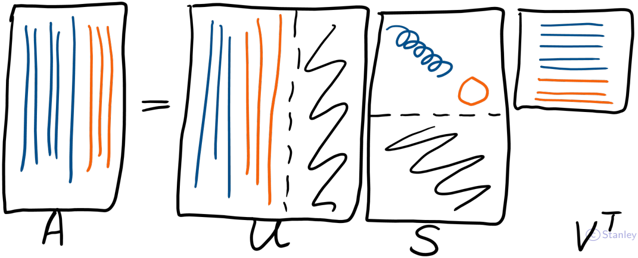


What if Rank-Deficient?

- If $\boldsymbol{A}$ is rank-deficient, then $\boldsymbol{A}^T\boldsymbol{A}$ is not invertible
- Approach 1: **Regularization**. See Lecture 2.
- Approach 2: **Pseudo-inverse**. Decompose $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$.
- $\boldsymbol{U} \in \mathbb{R}^{N\times N}$, with $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$. $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, with $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$.
- The diagonal block of $\boldsymbol{S} \in \mathbb{R}^{N\times d}$ is $\text{diag}\{s_1, \ldots, s_r, 0, \ldots, 0\}$.
- The solution is called the **pseudo-inverse**:

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{V}\boldsymbol{S}^+\boldsymbol{U}^T\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{S}^+ = \text{diag}\{1/s_1, \ldots, 1/s_r, 0, \ldots, 0\}$.



Reading List

Linear Algebra

- Gilbert Strang, Linear Algebra and Its Applications, 5th Edition.
- Carl Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, 2000.
- Univ. Waterloo Matrix Cookbook. `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

Linear Regression

- Stanford CS 229 (Note on Linear Algebra) `http://cs229.stanford.edu/section/cs229-linalg.pdf`
- Elements of Statistical Learning (Chapter 3.2) `https://web.stanford.edu/~hastie/ElemStatLearn/`
- Learning from Data (Chapter 3.2) `https://work.caltech.edu/telecourse`

Appendix

Solving the Minimum Norm problem

Consider this problem

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \|\boldsymbol{\theta}\|^2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{y}. \tag{5}$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \|\boldsymbol{\theta}\|^2 + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}).$$

Take derivative with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ yields

$$\nabla_{\boldsymbol{\theta}}\mathcal{L} = 2\boldsymbol{\theta} + \boldsymbol{A}^T\boldsymbol{\lambda} = 0, \quad \nabla_{\boldsymbol{\lambda}}\mathcal{L} = \boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y} = 0$$

- First equation gives us $\boldsymbol{\theta} = -\boldsymbol{A}^T\boldsymbol{\lambda}/2$.
- Substitute into second equation yields $\boldsymbol{\lambda} = -2(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{y}$.
- Therefore, $\boldsymbol{\theta} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{y}$.