REGRESSION

Dr. Matthieu R Bloch

Wednesday October 06, 2021
LOGISTICS

Assignment 4 assigned Tuesday, October 5, 2021

- Includes a (small) programming component
- Due October 14, 2021 (soft deadline, hard deadline on October 16)
WHAT’S ON THE AGENDA FOR TODAY?

Last time: Least-square regression

Today

- Solving linear least-square regression
- Extension to infinite dimension

Reading: Romberg, lecture notes 8
SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \| y - X\theta \|^2_2$ must satisfy

$$X^T X \theta^* = X^T y$$

This system is called *normal equations*.
**Proposition.** Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \| y - X\theta \|_2^2$ must satisfy

$$X^TX\theta^* = X^Ty$$

This system is called *normal equations*

**Facts:** for any matrix $A \in \mathbb{R}^{m \times n}$

- $\ker A^T A = \ker A$

$$\begin{align*}
\ker(A) &\triangleq \text{Nul}(A) \triangleq \{ x \in \mathbb{R}^n : A x = 0 \} \subset \mathbb{R}^n \\
\text{Im}(A) &\triangleq \text{Col}(A) \triangleq \{ A x : x \in \mathbb{R}^n \} \subset \mathbb{R}^m
\end{align*}$$
**Proposition.** Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2$ must satisfy

$$X^T X \theta^* = X^T y$$

This system is called *normal equations*.

**Facts:** for any matrix $A \in \mathbb{R}^{m \times n}$

- $\ker A^T A = \ker A$
- $\text{col}(A^T A) = \text{row}(A) \subset \mathbb{R}^n$

\[ A = \begin{bmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{pmatrix} c_1 & \cdots & c_n \end{pmatrix} = \begin{pmatrix} \vdots \\ \vdots \\ \vdots \end{pmatrix} \begin{pmatrix} c_1 & \cdots & c_n \end{pmatrix} \begin{pmatrix} \vdots \\ \vdots \\ \vdots \end{pmatrix} \begin{pmatrix} m \end{pmatrix} \]
SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution \( \theta^* \) to the problem \( \min_{\theta \in \mathbb{R}^d} ||y - X\theta||_2^2 \) must satisfy

\[
X^TX\theta^* = X^Ty
\]

This system is called *normal equations*

**Facts:** for any matrix \( A \in \mathbb{R}^{m \times n} \)

- \( \ker A^TA = \ker A \)
- \( \text{col}(A^TA) = \text{row}(A) \)
- \( \text{row}(A) \) and \( \ker A \) are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
**Proposition.** Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \| y - X\theta \|_2^2$ must satisfy

$$X^T X \theta^* = X^T y$$

This system is called *normal equations.*

**Facts:** for any matrix $A \in \mathbb{R}^{m \times n}$

- $\ker A^T A = \ker A$
- $\text{col}(A^T A) = \text{row}(A)$
- $\text{row}(A)$ and $\ker A$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If $\text{rank}(X) = d$, there is a unique solution: $\left( A^T A \right)^{-1} A^T y$  \[ \text{(b/c } X^TX \text{ is invertible)} \]
**Proposition.** Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \| y - X\theta \|_2^2$ must satisfy

$$X^T X \theta^* = X^T y$$

This system is called **normal equations**

**Facts:** for any matrix $A \in \mathbb{R}^{m \times n}$

- $\ker A^T A = \ker A$
- $\text{col}(A^T A) = \text{row}(A)$
- $\text{row}(A)$ and $\ker A$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If $\text{rank}(X) = d$, there is a unique solution: $(A^T A)^{-1} A^T y$
3. If $\text{rank}(X) < d$ there are infinitely many non-trivial solutions

If $\text{rank}(X) < d$, $\ker(X) \neq \emptyset$

$\exists \theta_0 \neq 0$ and $X\theta_0 = 0$

For any solution $\theta^*$ of the normal equations, $\theta^* + \theta_0$ is also a solution

$X^T X (\theta^* + \theta_0) = X^T X \theta^* + X^T \theta_0 = X^T y$

**Remark:** the space of solutions is $\theta^* + K$
Remark: Assume $\tilde{\Theta}$ a solution of $X^T\Theta = X^T y$ (in addition $\Theta^*$)

then $\tilde{\Theta} = \Theta^* + (\Theta - \Theta^*)$

$X^T X (\tilde{\Theta} - \Theta^*) = X^T y - X^T y = 0$ so that $\tilde{\Theta} - \Theta^* \in \text{Ker}(X^T X) = \text{Ker}(X)$

Hence $\tilde{\Theta} \in \text{Ker}(X)$

space of solutions

Remark: if $\Theta \in \text{Ker}(X)$ then $\forall x \in \mathbb{R}$ $x \Theta \in \text{Ker}(X)$
Proposition. Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2$ must satisfy

$$X^T X \theta^* = X^T y$$

This system is called normal equations

Facts: for any matrix $A \in \mathbb{R}^{m \times n}$

- $\ker A^T A = \ker A$
- $\text{col}(A^T A) = \text{row}(A)$
- $\text{row}(A)$ and $\ker A$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If $\text{rank}(X) = d$, there is a unique solution: $(A^T A)^{-1} A^T y$
3. If $\text{rank}(X) < d$ there are infinitely many non-trivial solutions
4. If $\text{rank}(X) = n$, there exists a solution $\theta^*$ for which $y = X\theta^*$

Note: $\text{rank}(X) = n$
- Then $XX^T$ is invertible
- So that $XX^T X\theta^* = XX^T y$
- and $X\theta^* = y$
- $\|y - X\theta^*\|_2^2 = 0$
**Proposition.** Any solution $\theta^*$ to the problem $\min_{\theta \in \mathbb{R}^d} \| y - X\theta \|^2_2$ must satisfy

$$X^T X \theta^* = X^T y$$

This system is called *normal equations*.

**Facts:** for any matrix $A \in \mathbb{R}^{m \times n}$

- $\ker A^T A = \ker A$
- $\text{col}(A^T A) = \text{row}(A)$
- $\text{row}(A)$ and $\ker A$ are orthogonal complements

We can say a lot more about the normal equations:

1. There is always a solution
2. If $\text{rank}(X) = d$, there is a unique solution: $(A^T A)^{-1} A^T y$
3. If $\text{rank}(X) < d$ there are infinitely many non-trivial solutions
4. If $\text{rank}(X) = n$, there exists a solution $\theta^*$ for which $y = X\theta^*$

In machine learning, there are often infinitely many solutions.
One reasonable to choose a solution among infinitely many is the *minimum energy* principle

\[
\min_{\theta \in \mathbb{R}^d} \|\theta\|_2^2 \text{ such that } X^T X \theta = X^T y
\]

- We will see the solution is always unique using the SVD

For now, assume that \(\text{rank}(X) = d\), so that the problem becomes

\[
\min_{\theta \in \mathbb{R}^d} \|\theta\|_2^2 \text{ such that } X \theta = y
\]
One reasonable to choose a solution among infinitely many is the *minimum energy* principle

\[
\min_{\theta \in \mathbb{R}^d} ||\theta||_2^2 \text{ such that } X^T X \theta = X^T y
\]

- We will see the solution is always unique using the SVD

For now, assume that \( \text{rank}(X) = n \), so that the problem becomes

\[
\min_{\theta \in \mathbb{R}^d} ||\theta||_2^2 \text{ such that } X \theta = y
\]

**Proposition.** The solution is \( \theta^* = X^T (X X^T)^{-1} y \in \mathbb{R}^d \)

\[
X^T (X X^T)^{-1} y \text{ is in } \text{col}(X^T) = \text{row}(X) = \text{span}(\{x_i : i = 1\})
\]
Proof: A solution is in $\mathbb{R}^d = \ker(X) \oplus \text{row}(X)$ so that it can be written $\theta_1 + \theta_2$

\[ x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d \]

Hence we are looking for \[ \min_{\theta_1 \in \text{row}(X)} \| \theta_1 + \theta_2 \|_2 \quad \text{s.t.} \quad X(\theta_1 + \theta_2) = y = X\theta_1. \]

Since $\ker(X) = \text{row}(X)$ then \[ \| \theta_1 + \theta_2 \|_2 = \| \theta_1 \|_2 + \| \theta_2 \|_2. \]

Hence any solution of the problem must be such that $\theta_2 = 0$ and must be in $\text{row}(X)$.

Therefore we can parameterize any solution as $\theta = X^T \alpha$.

Our equivalent problem is then to find \[ \min_{\alpha \in \mathbb{R}^m} \| X^T \alpha \|_2 \quad \text{s.t.} \quad y = X^T \alpha. \]

The only solution is actually \[ \alpha = (X^T)^{-1} y \] so that \[ \theta = X^T (X^T)^{-1} y. \]
Recall the problem

\[
\min_{\theta \in \mathbb{R}^d} \|\theta\|_2^2 \text{ such that } X^T X \theta = X^T y
\]

- There are infinitely many solutions if \( \ker X \) is non-trivial
- The space of solutions is unbounded! (i.e., \( \|\theta\|_2 \) is unbounded)
Recall the problem

\[ \min_{\theta \in \mathbb{R}^d} \| \theta \|_2^2 \text{ such that } X^T X \theta = X^T y \]  

- There are infinitely many solutions if \( \ker X \) is non trivial
- The space of solution is unbounded!
- Even if \( \ker X = \{0\} \), the system can be poorly conditioned

**Regularization** with \( \lambda > 0 \) consists in solving

\[ \min_{\theta \in \mathbb{R}^d} \| y - X \theta \|_2^2 + \lambda \| \theta \|_2^2 \]
Recall the problem

\[
\min_{\theta \in \mathbb{R}^d} \|\theta\|_2^2 \text{ such that } X^T X \theta = X^T y
\]

- There are infinitely many solutions if \( \ker X \) is non-trivial.
- The space of solutions is unbounded.
- Even if \( \ker X = \{0\} \), the system can be poorly conditioned.

Regularization with \( \lambda > 0 \) consists in solving

\[
\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2
\]

- This problem always has a unique solution.

**Proposition.** The solution is \( \theta^* = (\underbrace{X^T X + \lambda I}_{\in \mathbb{R}^{d \times d}})^{-1} X^T y = \underbrace{X^T (XX^T + \lambda I)^{-1}}_{\in \mathbb{R}^{d \times d}} y \)
Proof. \[ \|y - x^\theta \|^2 + \lambda \|w\|^2 = (y - x^\theta)^T (y - x^\theta) + \lambda \theta^T \theta \]
\[ = y^T y - 2 y^T x^\theta + \theta^T X^T x^\theta + \lambda \theta^T \theta \]
\[ = \theta^T (X^T X + \lambda I) \theta \]

Taking the gradient and setting it to zero, we obtain \( (X^T X + \lambda I) \theta = X^T y \)

is always invertible!

Facts: 1. \( X^T X + \lambda I \) is symmetric
2. \( X^T X + \lambda I \) is positive definite: \( w^T (X^T X + \lambda I) w = w^T X^T X w + \lambda w^T w = \|Xw\|^2 + \lambda \|w\|^2 \geq 0 \)
   with equality iff \( w = 0 \)
3. \( X^T X + \lambda I \) is invertible \( \lambda > 0 \)

Hence \( \theta = (X^T X + \lambda I)^{-1} X^T y \)
We want to show that \( \Theta = X^T (XX^T + \delta I)^{-1} y \).

Since \( \Theta \) is unique, all we have to check is that \( \Theta \) is a solution of \( (XX^T + \delta I) \Theta = X^T y \).

Note that \( (XX^T + \delta I) \Theta = (XX^T + \delta I) y = (X^T X + \delta I) \Theta = X^T y \).
Recall the problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_2^2 \text{ such that } X^T X \theta = X^T y$$

- There are infinitely many solutions if \( \ker X \) is non-trivial
- The space of solutions is unbounded!
- Even if \( \ker X = \{0\} \), the system can be poorly conditioned

**Regularization** with \( \lambda > 0 \) consists in solving

$$\min_{\theta \in \mathbb{R}^d} \|y - X \theta\|_2^2 + \lambda \|\theta\|_2^2$$

- This problem **always** has a unique solution

**Proposition.** The solution is \( \theta^* = (X^T X + \lambda I)^{-1} X^T y = X^T (XX^T + \lambda I)^{-1} y \)

Note that \( \theta^* \) is the row space of \( X \)

$$\theta^* = X \alpha \text{ with } \alpha = (XX^T + \lambda I)^{-1} y$$