Recitation 8

October 3, 2024

1 Recap

1.1 Orthogonal Decomposition Revisited

Given a matrix $A \in \mathbb{R}^{m \times n}$. Any vector $b \in \mathbb{R}^m$ can be uniquely expressed as $b = p + e$ for which $p \in C(A)$ and $e \in N(A^T)$. In particular, p and e are the orthogonal projections of b onto $C(A)$ and $N(A^T)$ respectively.

When A is tall ($m \geq n$) and has linearly independent columns, we can write

$$p = A(A^TA)^{-1}A^Tb \quad \text{and} \quad e = (I - A(A^TA)^{-1}A^T)b.$$

In other words, the projection matrix onto $C(A)$ is $P = A(A^TA)^{-1}A^T$ and onto $N(A^T) = Q = I - P$.

It is a good exercise to verify that such decomposition makes sense; that is to explain why $(A^TA)^{-1}$ exists, $p \in C(A)$, $e \in N(A^T)$, and $b = p + e$.

1.2 Overdetermined System

Consider the linear system $Ax = b$, where we have more equations than variables; i.e. A is tall with more rows than columns. The system may not have a solution that satisfies all equations.

Least Squares Approximate Solution: Assume linearly independent columns

1. Orthogonal projection: Project b onto the column space of A, i.e. $p = proj_{C(A)}b = A(A^TA)^{-1}A^Tb$. Then solve $Ax = p \Rightarrow x = (A^TA)^{-1}A^Tb$. This vector minimizes the norm of the residual $r = Ax - b$.

2. Calculus: Want to find x that minimizes $\|Ax-b\|^2$. Taking the gradient $\nabla_x\|Ax-b\|^2$ and setting it to zero gives $x = (A^TA)^{-1}A^Tb$.

There are many approaches to obtain the same result – the least-squares approximate solution $\hat{x} = (A^TA)^{-1}A^Tb$. (In optimization, this solution, the minimizer, is often denoted "$x_\star$" or "$x^\star$". We will use $x_\star$ below.)

The matrix $(A^TA)^{-1}A^T$ is sometimes called a **"left inverse"** of such a tall A, because if you multiply it on the left of A you get $(A^TA)^{-1}A^TA = I$. If A is non-square, however, it is *not* an ordinary matrix inverse because if you multiply it on the right of A you get $P = A(A^TA)^{-1}A^T$, a projection instead of an identity. (In the same way that $Q^TQ = I$ but QQ^T is a projection.)

1.3 Underdetermined System

More variables than equations; i.e. A has more columns than rows. The system has infinitely many solutions, and we need to pick a specific one.

Minimum Norm Solution: Assuming linearly independent rows, a common choice is to pick the "smallest" solution, i.e. we minimize $\|x\|^2$ subject to the constraint $Ax = b$. The solution of minimum norm is $x_\star = A^T(AA^T)^{-1}b$.

The matrix $A^T(AA^T)^{-1}$ is sometimes called a **"right inverse"** of such a wide A, because if you multiply it on the right of A you get $AA^T(AA^T)^{-1} = I$.

1.4 Regularization

Our goal remains the same: to solve the system $Ax = b$; however, the solution we want now is the one that *minimizes* $T(x) := \|Ax-b\|^2 + \lambda\|x\|^2$ where $\lambda > 0$ is a regularization parameter. The unique optimal solution is given by $x_\star = (A^TA + \lambda I)^{-1}A^Tb$. It can be shown that the inverse exists for any A (regardless of rank) for any $\lambda > 0$.

(This particular regularization is often called a **"ridge" or "Tikhonov" regularization**. We will later see that other regularizations are possible too; which regularization is the best depends on the specific application and what is known about the desired solution.)

2 Exercises

You can use Julia to help you with calculations if you want, or leave the answer in a form where you could plug it into Julia.

1. Let's look more closely at the minimum-norm solution $x_\star = A^T(AA^T)^{-1}b$ when A is a "wide" matrix with full row rank (underdetermined).

   (a) We already saw that if A has full *column* rank, then A^TA is invertible. Why does this mean that AA^T is invertible when A has full *row* rank?

   (b) Show that $x_\star$ is a solution to $Ax = b$.

   (c) $x_\star$ is in what subspace of A? Any *other* solution to $Ax = b$ must be of the form $x = x_\star + v$ where v is a vector in what subspace of A? From this, explain why $\|x_\star + v\| \geq \|x_\star\|$ for any such v, without using calculus. Hence, $x_\star$ is the minimum-norm solution!

2. Two points in $\mathbb{R}^3$ have (x, y, z) coordinates as follows.

$$a = (1, 0, 0), \quad b = (0, 1, 1),$$

   (a) Find the plane $z = C + Dx + Ey$ that gives the best fit to the two points a and b that minimizes $C^2 + D^2 + E^2$.

   (b) What is the least squares error?

   (c) Predict the value of z when $(x, y) = (2, -1)$.

3. It was claimed in class that the ridge-regularized least-squares problem, minimizing $\|Ax - b\|^2 + \lambda\|x\|^2$ for $\lambda > 0$, is solved by $x_\star = (A^TA + \lambda I)^{-1}A^Tb$. This can be easily derived without calculus by showing that it is equivalent to an ordinary least-squares problem.

   (a) $\|Ax - b\|^2 + \lambda\|x\|^2 = \left\| \begin{pmatrix} Ax - b \\ ??? \end{pmatrix} \right\|^2$ for what "???"?

   (b) Hence, $\|Ax - b\|^2 + \lambda\|x\|^2 = \|Bx - d\|^2$ where $B = \begin{pmatrix} A \\ ?? \end{pmatrix}$ and $d = \begin{pmatrix} b \\ ?? \end{pmatrix}$ for what ??'s?.

   (c) Hence the minimizer is the ordinary least-square solution $x_\star = (B^TB)^{-1}B^Td$. Explain why this gives $x_\star = (A^TA + \lambda I)^{-1}A^Tb$.

   (d) *Optional:* Show that $A^TA + \lambda I$ is always invertible. $x^T(A^TA + \lambda I)x = (Ax)^T(Ax) + \lambda x^Tx = \|Ax\|^2 + \lambda\|x\|^2 \geq 0$, and is only $= 0$ if $x =$??. Why does this imply that $N(A^TA + \lambda I) = \{\vec{0}\}$ (hence invertible since it is square)?

4. Consider the function values

$$f(-2) = 0, \quad f(-1) = 0, \quad f(0) = 1, \quad f(1) = 0, \quad f(2) = 0.$$

   (a) Find the straight line $f(t) = C + Dt$ that is closest (in the least squares sense) to these values.

   (b) Find the parabola $f(t) = C + Dt + Et^2$ that is closest (in the least squares sense) to these values. *Hint: Write down the system of equations $A\mathbf{x} = \mathbf{b}$ in three unknowns $x = (C, D, E)$ for the parabola $f(t)$ to go through the points.*

   (c) Find the closest 4th degree polynomial for these points. What is the least squares error?

3 Solutions

1. Let's look more closely at the minimum-norm solution $x_\star = A^T(AA^T)^{-1}b$ when A is a "wide" matrix with full row rank (underdetermined).

   (a) If A has full row rank, then $B = A^T$ has full column rank, and hence $B^TB = AA^T$ is invertible.

   (b) Easy: $Ax_\star = \cancel{AA^T(AA^T)^{-1}}b = b$.

   (c) $x_\star$ is A^T(something), so it is in $C(A^T)$. Any *other* solution to $Ax = b$ must be of the form $x = x_\star + v$ where v is a vector $N(A)$ (so that $Ax = Ax_\star + \cancel{Av} = b$). But these are orthogonal complements, so $x_\star \perp v$! Hence $\|x_\star + v\|^2 = (x_\star + v)^T(x_\star + v) = \|x_\star\| + \|v\|^2 \geq \|x_\star\|^2$ (the cross terms are zero by orthogonality). This is very similar to the derivation from class of the least-square solution to the overdetermined problem!

2. (a) We have

$$\underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{pmatrix} C \\ D \\ E \end{pmatrix}}_{u} = \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{b}$$

   and $b = (0, 1)$ we wish to solve $Au = b$ which minimizes $\|u\|^2$. The solution is $u_\star = A^T(AA^T)^{-1}b = (1/3, -1/3, 2/3) = (C, D, E)$. (This a 2×2 system of equations: solve $AA^Tv = b$ followed by $u = A^Tv$.)

   (b) least squares error is 0 because the system is underdetermined.

   (c) $C + 2D - 1E = 1/3 + 2(-1/3) - 1(2/3) = -1$

3. It was claimed in class that the ridge-regularized least-squares problem, minimizing $\|Ax - b\|^2 + \lambda\|x\|^2$ for $\lambda > 0$, is solved by $x_\star = (A^TA + \lambda I)^{-1}A^Tb$. This can be easily derived without calculus by showing that it is equivalent to an ordinary least-squares problem.

   (a) $\|Ax - b\|^2 + \lambda\|x\|^2 = \left\| \begin{pmatrix} Ax - b \\ \sqrt{\lambda}x \end{pmatrix} \right\|^2.$

   (b) Hence, $\|Ax - b\|^2 + \lambda\|x\|^2 = \|Bx - d\|^2$ where $B = \begin{pmatrix} A \\ \sqrt{\lambda}I \end{pmatrix}$ and $d = \begin{pmatrix} b \\ \vec{0} \end{pmatrix}$.

   (c) $B^Td = A^Tx$ (since the other components are zero), and $B^TB = A^TA + \lambda I$, so this gives $x_\star = (B^TB)^{-1}B^Td(A^TA + \lambda I)^{-1}A^Tb$ as desired.

   (d) *Optional:* $x^T(A^TA + \lambda I)x = (Ax)^T(Ax) + \lambda x^Tx = \|Ax\|^2 + \lambda\|x\|^2 \geq 0$, and is only $= 0$ if $x = 0$ (since otherwise the $\lambda\|x\|^2$ term is > 0). If $(A^TA + \lambda I)x = \vec{0}$, then $x^T(A^TA + \lambda I)x = 0$, but we just showed that this is true if and only if $x = 0$. Hence $N(A^TA + \lambda I) = \{\vec{0}\}$ (hence invertible since it is square).

4. (a) We want to solve an *overdetermined* system $A\mathbf{x} = b$ for which

$$A = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ with variables } \mathbf{x} = \begin{bmatrix} C \\ D \end{bmatrix}.$$

The least square answer is given by

$$\begin{bmatrix} C \\ D \end{bmatrix} = \mathbf{x} = (A^\top A)^{-1} A^\top b = \begin{bmatrix} 1/5 \\ 0 \end{bmatrix}$$

which means the closest line is $f(t) = 1/5$.

(b) We want to solve an *overdetermined* system $A\mathbf{x} = b$ for which

$$A = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ with variables } \mathbf{x} = \begin{bmatrix} C \\ D \\ E \end{bmatrix}.$$

The least square answer is given by

$$\begin{bmatrix} C \\ D \\ E \end{bmatrix} = \mathbf{x} = (A^\top A)^{-1} A^\top b = \begin{bmatrix} 17/35 \\ 0 \\ -1/7 \end{bmatrix}$$

which means the closest parabola is $f(t) = \frac{17}{35} - \frac{t^2}{7}$.

(c) Suppose that we want to solve for $f(t) = C + Dt + Et^2 + Ft^3 + Gt^4$. We want to solve an *overdetermined* system $A\mathbf{x} = b$ for which

$$A = \begin{bmatrix} 1 & -2 & 4 & -8 & 16 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ with variables } \mathbf{x} = \begin{bmatrix} C \\ D \\ E \\ F \\ G \end{bmatrix}.$$

We note that A is square and invertible, which means that we can solve x exactly and uniquely. By solving $\mathbf{x} = A^{-1}b$, we derive $\mathbf{x} = [1\ 0\ -5/4\ 0\ 1/4]^\top$ which gives $f(t) = 1 - \frac{5}{4}t^2 + \frac{1}{4}t^4$.

Since the system can be solve to the exact, we must have $Ax = b$ which means that the residual $r = Ax - b$ is zero. The least square error is thus $\|r\|^2 = 0$.

(d) Construct the system of equations $A\mathbf{x} = b$ in a similar fashion to previous parts. When the degree is 5, we have 5 equations with 6 variables which make the system *underdetermined*. This means we have infinitely many answers and the smallest answer is given by $\mathbf{x} = A^\top (AA^\top)^{-1} b = [1\ 0\ -5/4\ 0\ 1/4\ 0]^\top$ which gives $f(t) = 1 - \frac{5}{4}t^2 + \frac{1}{4}t^4$.

(e) With degree at least 4, we get the exact fit. On the other hand, the lower the degree is, the more general the best-fit line is. Note that there is no absolute best model/degree for data fitting. In this case, one might argue that linear fitting is the best because every point but $f(1) = 1$ yields value 0 which may lead us into thinking that $f(1) = 1$ is an *outlier*. Others may argue that it is absolutely needed to fit every point (or almost every point) onto the line so they tend to choose higher degree fitting. This; however, may lead to the *overfitting* problem.