Notes on slides 12 and 21-25 of lecture 6

Thibault Vatter

September 28, 2017

1 Slide 12

We say that U has orthonormal columns if its columns $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ form an orthonormal set. For instance, if $U = \begin{bmatrix} 1/\sqrt{2} & 2/3 \\ 1/\sqrt{2} & -2/3 \\ 0 & 1/3 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} \sqrt{2} \\ 3 \end{bmatrix}$, we have

$$U^\top U = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 2/3 & -2/3 & 1/3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 2/3 \\ 1/\sqrt{2} & -2/3 \\ 0 & 1/3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\text{and } U\mathbf{x} = \begin{bmatrix} 1/\sqrt{2} & 2/3 \\ 1/\sqrt{2} & -2/3 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} \sqrt{2}3 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} \implies \|U\mathbf{x}\| = \|x\| = \sqrt{11}.$$

The proof of the first part of theorem 8 for $n = 3$ (i.e., $U = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{bmatrix}$) is sufficient to understand what is going on. Note that we have:

$$U^\top U = \begin{bmatrix} \mathbf{u}_1^\top \\ \mathbf{u}_2^\top \\ \mathbf{u}_3^\top \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^\top \mathbf{u}_1 & \mathbf{u}_1^\top \mathbf{u}_2 & \mathbf{u}_1^\top \mathbf{u}_3 \\ \mathbf{u}_2^\top \mathbf{u}_1 & \mathbf{u}_2^\top \mathbf{u}_2 & \mathbf{u}_2^\top \mathbf{u}_3 \\ \mathbf{u}_3^\top \mathbf{u}_1 & \mathbf{u}_3^\top \mathbf{u}_2 & \mathbf{u}_3^\top \mathbf{u}_3 \end{bmatrix}$$

As such, $U^\top U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \iff \begin{cases} \mathbf{u}_i^\top \mathbf{u}_j = 1 & \text{if } i = j \\ \mathbf{u}_i^\top \mathbf{u}_j = 0 & \text{otherwise} \end{cases} \iff \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ form an orthonormal set.

The second part of theorem 8 essentially says that the the linear transformation resulting from multiplying a vector by U (i.e., a matrix "acting" on a vector) preserves length and orthogonality. The proof of a):

$$\begin{aligned} U\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{u}_i \implies \|U\mathbf{x}\|^2 &= \left(\sum_{i=1}^{n} x_i \mathbf{u}_i\right) \cdot \left(\sum_{i=1}^{n} x_i \mathbf{u}_i\right) \\ &= \sum_{i=1}^{n} \sum_{n=1}^{n} x_i x_j \qquad \underbrace{\mathbf{u}_i \cdot \mathbf{u}_j}_{1 \text{ if } i=j, 0 \text{ otherwise}} \\ &= \sum_{i=1}^{n} x_i^2 \\ &= \|\mathbf{x}\|^2 \end{aligned}$$

(the second equality because the columns of U are orthonormal.) The proof of b) is similar.

Theorem 8 is especially useful when it is applied to square and invertible matrices: defining a matrix as orthogonal if $U^{-1} = U^\top$, theorem 8 implies directly that such a matrix has orthonormal columns (i.e., the equivalence between 1 and 2 in theorem 9). The fact that such a matrix also has orthonormal rows can be proven as follows (i.e., equivalence between 1, 2 and 3 in theorem 9): by definition, U satisfies $U^\top U = I$ and $UU^\top = I$. Since $U = (U^\top)^\top$, we have $U^\top (U^\top)^\top = I$ et $(U^\top)^\top U^\top = I$, which shows that $U^\top$ is also orthogonal. By the second part of theorem 9, this implies that the columns of $U^\top$ are orthonormal, and since the columns of $U^\top$ are the rows of U, we conclude that the rows of U are orthonormal.

2 Slide 21 (intro)

I think that the wikipedia page provides a nice introduction:

`https://en.wikipedia.org/wiki/Linear_least_squares_(mathematics)`

Essentially, the idea is to express a response variable b as a linear function of a vector explanatory variables $\mathbf{a} \in \mathbb{R}^n$ plus some "noise" z:

$$b = \mathbf{a} \cdot \mathbf{x} + z, \tag{1}$$

where $\mathbf{x}$ is a vector of coefficients expressing the linear relationship between the response and explanatory variables. Note that, most of the time, people use y for the response, $\mathbf{x}$ for the explanatory variables and β for the coefficients. However, I used the notation above to stay consistent with the lectures on linear systems.

Assuming that one observes m data points, Equation 1 can be rewritten in matrix form as

$$\mathbf{b} = A\mathbf{x} + \mathbf{z},$$

where

$$\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}, A = \begin{bmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix} \text{ and } \mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}.$$

Think of $A\mathbf{x}$ as an approximation of $\mathbf{b}$. The smaller the distance between $A\mathbf{x}$ and $\mathbf{b}$, given by $\|\mathbf{b} - A\mathbf{x}\|$, the better the approximation. The least-squares problem then amounts at finding the "best approximation" (i.e., the one that minimizes $\|\mathbf{b} - A\mathbf{x}\|$), namely $\widehat{\mathbf{x}}$ such that

$$\|\mathbf{b} - A\widehat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|,$$

$\forall \mathbf{x} \in \mathbb{R}^n$. The adjective "least-squares" arise from the fact that $\|\mathbf{b} - A\mathbf{x}\|$ is the square root of a sum of squares.

Linear regression is (arguably) the most important statistical tool most people ever learn. However, the way its usually taught makes it hard to see the essence of what regression is really doing. Most of the time, people focus on the "calculus" view: they start with $\|\mathbf{b} - A\mathbf{x}\|$, minimize the expression by taking the first derivative with respect to $\mathbf{x}$, set it to zero, and do a ton of manipulations to solve for $\mathbf{x}$. In slides 22-25, I show that there is a simpler and more direct way to understand the least-squares problem.

3 Slide 22

Recall that multiplying a vector of $\mathbb{R}^n$ by A gives a vector of $\mathbb{R}^m$, and that $\operatorname{Col} A = \{\mathbf{x} \in \mathbb{R}^m \mid \text{ such that } \mathbf{x} \text{ is a linear combination of the colums of } A\}$.

The important aspect of the least-squares is that, no matter which $\mathbf{x}$ we select, $A\mathbf{x} \in \operatorname{Col} A$. As such, we seek $\mathbf{x}$ that makes $A\mathbf{x}$ the "closer" point to $\mathbf{b}$ in $\mathbb{R}^m$. Obviously, if $\mathbf{b} \in \operatorname{Col} A$, then the system $A\mathbf{x} = \mathbf{b}$ is consistent and any solution if also a least-squares solution.

If $\widehat{\mathbf{b}} = A\widehat{\mathbf{x}}$, then by definition $\widehat{\mathbf{b}} \in \operatorname{Col} A$. Since $\left\|\mathbf{b} - \widehat{\mathbf{b}}\right\| \leq \|\mathbf{b} - A\mathbf{x}\|$ for each $\mathbf{x} \in \operatorname{Col} A$, Theorem 10 implies that $\widehat{\mathbf{b}} = \operatorname{proj}_{\operatorname{Col} A}\mathbf{b}$, namely that $\widehat{\mathbf{b}}$ is the orthogonal projection of $\mathbf{b}$ on the columns of A. Furthermore, because one can always find such a projection, it means that the linear system $A\mathbf{x} = \widehat{\mathbf{b}}$ is consistent with $\widehat{\mathbf{x}}$ its (potentially non-unique) solution.

4 Slide 23

Let $A = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{bmatrix}$ (i.e., here $\mathbf{a}_1, \cdots, \mathbf{a}_n$ denote the columns of A and not the rows as above).

Now, since $\widehat{\mathbf{b}} = \text{proj}_{\text{Col}\,A}\mathbf{b} \in \text{Col}\,A$, then $\mathbf{b} - \widehat{\mathbf{b}} \in (\text{Col}\,A)^{\perp}$, that is the $\mathbf{b} - \widehat{\mathbf{b}}$ is in the orthogonal complement of $\text{Col}\,A$, which implies that $\mathbf{b} - \widehat{\mathbf{b}} = \mathbf{b} - A\widehat{\mathbf{x}}$ is orthogonal to every column of A (i.e., $\mathbf{a}_j \cdot (\mathbf{b} - A\widehat{\mathbf{x}})\,\forall j$). This last condition can be rewritten in matrix form as $A^{\top}(\mathbf{b} - A\widehat{\mathbf{x}}) = \mathbf{0}$ (i.e., $(\mathbf{b} - A\mathbf{x}) \in \text{Nul}\,A^{\top}$ by theorem 4), which can again be rewritten to obtain the normal equations.

Concerning the statement of theorem 12:

The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ is equal to the (nonempty) set of solutions of $A^{\top}A\mathbf{x} = A^{\top}\mathbf{b}$.

As mentioned above, the set of least-squares solutions is nonempty and $\widehat{\mathbf{x}}$ satisfies the normal equations. Now, we suppose that $\widehat{\mathbf{x}}$ satisfies $A^{\top}A\widehat{\mathbf{x}} = A^{\top}\mathbf{b}$ and we show that it is also a least-squares solution:

$$\begin{aligned} A^{\top}A\widehat{\mathbf{x}} = A^{\top}\mathbf{b} &\implies A^{\top}(\mathbf{b} - A\widehat{\mathbf{x}}) = \mathbf{0} \\ &\implies (\mathbf{b} - A\widehat{\mathbf{x}}) \in \text{Nul}\,A^{\top} \\ &\implies (\mathbf{b} - A\widehat{\mathbf{x}}) \in (\text{Col}\,A)^{\perp} \text{ (theorem 4)} \\ &\implies \mathbf{b} = A\widehat{\mathbf{x}} + (\mathbf{b} - A\widehat{\mathbf{x}}) \text{ is a decomposition such that} \\ &A\widehat{\mathbf{x}} \in \text{Col}\,A \text{ and } \mathbf{b} - A\widehat{\mathbf{x}} \in (\text{Col}\,A)^{\perp}. \end{aligned}$$

By the uniqueness of the orthogonal projection (theorem 10), we conclude that $A\widehat{\mathbf{x}} = \text{proj}_{\text{Col}\,A}\mathbf{b}$. In other words, $\widehat{\mathbf{x}}$ is a least-squares solution.

Concerning the second part of theorem 12:

Furthermore, the following statements are equivalent:

1. *$A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution $\forall \mathbf{b} \in \mathbb{R}^m$.*

2. *The columns of A are linearly independent.*

3. *$A^{\top}A$ is invertible.*

The proof involves many concepts from the previous lectures and is a good exercise.

- $1 \iff 3 : 1$ notice that $Q = A^{\top}A$ is an $n \times n$ matrix and $\mathbf{v} = A^{\top}\mathbf{b} \in \mathbb{R}^n$. As such, $1 \iff 3$ results from the fact that $Q\mathbf{x} = \mathbf{v}$ has a unique solution if an only if Q is invertible (from the theorem about invertible matrices).

- $3 \iff 2$: note that
  - A matrix has linearly independent columns if $\sum_{i=1}^{n} c_i \mathbf{a}_i = \mathbf{0} \iff c_i = 0 \forall i$, which can be rewritten as $A\mathbf{c} = \mathbf{0}$ with $\mathbf{c} \in \mathbb{R}^n$ the weights in vector form. In other words, A has linearly independent columns if and only if $\operatorname{Nul} A = \{\mathbf{0}\}$.
  - One can prove that $A^\top A\mathbf{x} = \mathbf{0} \iff A\mathbf{x} = \mathbf{0}$, which implies $\operatorname{Nul} A^\top A = \operatorname{Nul} A$.

  One then has $3 \iff \operatorname{Nul} A^\top A = \{\mathbf{0}\} \iff \operatorname{Nul} A = \{\mathbf{0}\} \iff 2$, where the first implication results from the property of invertible matrices, the second implication results from $\operatorname{Nul} A^\top A = \operatorname{Nul} A$, and the third implication from the fact that A has linearly independent columns if and only if $\operatorname{Nul} A = \{\mathbf{0}\}$.

- $1 \implies 2$:

Since $1 \iff 3 \iff 2$, the proof is complete, although showing that $A^\top A\mathbf{x} = \mathbf{0} \iff A\mathbf{x} = \mathbf{0}$ is left as an exercise.

Concerning the formula $\widehat{\mathbf{x}} = (A^\top A)^{-1} A^\top \mathbf{b}$, it is mainly useful for theoretical purposes, for hand calculations when $A^\top A$ is a 2×2 invertible matrix, and for actually implementing of linear regressions (in statistical software).

5 Slide 24

This is an example where $A^\top A$ is a 2×2 invertible matrix and the calculations can be done directly with $\widehat{\mathbf{x}} = (A^\top A)^{-1} A^\top \mathbf{b}$.

6 Slide 25

This is an example where computing $(A^\top A)^{-1}$ to use $\widehat{\mathbf{x}} = (A^\top A)^{-1} A^\top \mathbf{b}$ is harder. As such, one proceeds by solving the normal equations $A^\top A\widehat{\mathbf{x}} = A^\top \mathbf{b}$ as usual.