Mathematics of Big Data I

Review Session 0: Linear Algebra

September 1st, 2016

1 Linear Systems and Matrix Representation

What is the big picture of linear algebra? When we solve a linear system, for example

$$\begin{aligned} x + 2y &= 1 \\ 3x + 4y &= -1 \end{aligned}$$

we want to represent this system in the simplest way possible. If we write this as a matrix equation we get

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}}_{A}\underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}_{y}$$

so we can write this as $Ax = y$. The way we traditionally solve an equation like this is by diagonalizing the matrix A using elementary row operations. We can multiply the top row by 3 and subtract it from the second row to obtain

$$\begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \end{bmatrix}.$$

We can encode this operation into a matrix by using the elementary matrix

$$E_1 = \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}$$

and thecomputation above is identical to

$$\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The matrix E_1A is upper triangular. A theorem shows that the inverse of each elementary matrix is again an elementary matrix. We can easily find that

$$E_1^{-1} = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}$$

and it can be easily verified that $E_1^{-1}E_1 = E_1E_1^{-1} = I$ where I is the identity matrix. Observe that E_1 and E_1^{-1} are both lower triangular. Observe now that since

$$E_1A = \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}$$

we have that

$$A = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix} = LU$$

and therefore A can be decomposed into a lower triangular matrix multiplied on the right by an upper triangular matrix. We can actually find a unique decomposition of this form for A where the upper triangular matrix has ones along the diagonal. Another elementary row operation we find to make the upper triangular have units along the diagonal is the matrix

$$E_2 = \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}.$$

and therefore we find

$$E_2BA = \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \tilde{U}$$

and since

$$(E_2E_1)^{-1} = E_1^{-1}E_2^{-1} = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 3 & -2 \end{bmatrix} = \tilde{L}$$

then A has a unique decomposition $A = \tilde{L}\tilde{U}$. This is called the **Cholesky decomposition**.

Recall that a matrix A is **symmetric** if $A = A^T$. We say that a symmetric matrix is **positive-definite** if all of the eigenvalues of A are strictly positive. A symmetric positive-definite matrix A has a (non-unique) decomposition as $A = M^TM$. Question: is the matrix M invertible? One way we can tell is by is by taking the determinant of both sides of this equation. We know that since $\det A > 0$ we have that

$$0 < \det A = \det(M^TM) = \det(M^T)\det(M) = (\det M)^2.$$

and therefore $\det M \neq 0$ and M is invertible. How can we prove that A has this form?

Proof: We know that $A = \tilde{L}\tilde{U}$ is a unique Cholesky decomposition. Note however that since $A^T = A$ we have

$$A = A^T = \tilde{U}^T\tilde{L}^T.$$

by the uniqueness we must have that $\tilde{U}^T\tilde{L}^T = \tilde{L}\tilde{U}$ and therefore $\tilde{U}^T = \tilde{L}$. and therefore we can write $A = \tilde{U}^T\tilde{U}$. □

1.1 Rank and Determinant

It is important to have useful characteristics that are easy to compute for matrices. The first two that we look at are the determinant and rank of a matrix. The determinant of a matrix immediately gives information about the invertibility of a matrix. We can think of the determinant of a matrix geometrically in terms of a parallelpiped. The determinant of a matrix gives the (oriented) volume of the parallelpiped formed by the columns of the matrix. We can immediately see that if the determinant is zero, then some of the columns are linearly dependent.

1.2 Vector Spaces

When we discuss vector spaces we can "mimic" $\mathbb{R}^n$. Concepts that we can asbtractly define based on our notions of $\mathbb{R}^n$ are **subspaces, linear span, linear dependence/independence, basis and dimension**. We can define vector spaces in the following way.

Definition 1 *A **real vector space** V is a collection of elements called **vectors** such that for any two vectors $u, v \in V$ we have $u + v \in V$ and for any scalar $a \in \mathbb{R}$ we have $av \in V$.*

The abstract notion of a vector space allows us to discuss a broader range of examples than simpy $\mathbb{R}^n$. For instance, the collection $P_2(x)$ of all polynomials of degree at most 2 is a vector space. The set of all functions is also a vector space. Unlike $\mathbb{R}^n$, the collection of all functions with the same domain and codomain is infinite-dimensional.

1.3 Linear Transformations

The key to linear algebra is the concept of *linearity*. The reason why linearity is so powerful is that a linear transformation allows us to understand an infinite amount of information with only finitely many parameters. Tensors are an extension linear transformations which play an important role in data analysis.

Definition 2 *Given two vector spaces V and W, a **linear transformation** or **linear map** is a mapping $L : V \to W$ such that for all $u, w \in V$ and for all $k \in \mathbb{R}$ we have*

$$\begin{aligned} L(u+w) &= L(u) + L(w) \\ L(ku) &= kL(u). \end{aligned}$$

For every linear transformation there is an associated *matrix representation* which we can define as follows. Given vectors spaces V and W with linear map $L : V \to W$, we can define a basis $v_1, \ldots, v_n$ of V. Given a vector $w = a_1 v_1 + \ldots + a_n v_n$ we have that

$$\begin{aligned} L(w) &= L(a_1 v_1 + \ldots + a_n v_n) \\ &= a_1 L(v_1) + \ldots + a_n L(v_n) \\ &= \underbrace{\begin{bmatrix} L(v_1) & \ldots & L(v_n) \end{bmatrix}}_{matrix\ representation} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}. \end{aligned}$$

The **kernel** of a linear map is the set of all elements $v \in V$ such that $L(v) = 0 \in W$ and is a subspace of V. The **range** of L is the image $L(V) \subseteq W$. The range-nullity theorem guarantees that

$$\dim \ker(L) + \dim \text{range}(L) = \dim V.$$

We know this since if $\dim V = n$, $\dim W = m$, then if the rank of the matrix representation of L is k we know that $\dim \ker L = n - k$ and $\dim \text{range}\, L = k$.

2 Convex Optimization

Definition 3 *A **convex combination** of two vectors x and y in a vector space V is a vector $\theta x + (1-\theta)y$ where $0 \le \theta \le 1$. A **convex set** is a subset $A \subseteq V$ such that for all $x, y \in A$ the convex combination $\theta x + (1-\theta)y \in A$ for all $\theta \in [0,1]$.*

Some examples of convex sets include the collection of all Hermitian matrices H such that $H^* = H$ where $H^* = \overline{H}^T$ and the overbar denotes complex conjugation. Another nontrivial convex set is the collection of probability distributions on $\mathbb{R}^n$. We can think of this as the set

$$\mathbb{P}(\mathbb{R}^n) = \left\{ \mathbb{P}(x) \;\middle|\; \int_{\mathbb{R}^n} \mathbb{P}(x)dx = 1\ ,\ \text{and} \mathbb{P}(x) \right\}.$$

Given $f, g \in \mathcal{P}(\mathbb{R}^n)$ and $\theta \in [0,1]$ we have

$$\int_{\mathbb{R}^n} (\theta f + (1-\theta)g)dx = \theta \int f dx + (1-\theta) \int g dx = \theta + (1-\theta) = 1.$$

We also have that $\theta f(x) + (1-\theta)g(x) \ge 0$. For example, we can take $f(x) = e^{x^T x} = e^{x_1^2 + \ldots + x_n^2}$. Taking the partial derivatives of this, we can show that the Hessian $\nabla^2 f$ will be positive definite. Observe that

$$\begin{aligned}\frac{\partial^2 f}{\partial x_i \partial x_j} &= \frac{\partial f}{\partial x_i} 2x_j e^{x^T x} \\ &= 4x_i x_j e^{x^T x}\end{aligned}$$

and we obtain that $\nabla^2 f = 4xx^T e^{x^T x}$. We can show this is positive-definite by taking any arbitrary $z \in \mathbb{R}^n$ and observing that

$$z^T \nabla^2 f z = 4e^{x^T x}(z^T x)(x^T z) = 4(z^T x)^2 e^{x^T x} \geq 0.$$

It so turns out that the only multivariable functions that are convex are the ones that have positive semi-definite Hessians.

Theorem 1 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the Hessian $\nabla^2 f$ is positive semi-definite.*

Theorem 2 *Given $f : \mathbb{R}^n \to \mathbb{R}$ which is convex and an arbitrary matrix A on $\mathbb{R}^n$, we also have that $f(Ax+b)$ is convex.*

We can use convexity for optimization problems. Given a function $f_0(x)$ subject to the constraints $f_i(x) < 0$ and $h_j(x) = 0$, we say that a point x^* is **optimal** if there exists $R > 0$ such that $\|x^* - x\| < R$. Oftentimes we want to minimize $f(x)$ such that $f_i(x) \leq 0$ and $Ax = 0$ with both f and f_i convex, then we say that this problem is a **convex optimization problem**.

Theorem 3 *Given a convex problem and a local optimum x^* such that*

$$f_0(x^*) = \inf\{f_0(z) \mid z \text{ is reasonable and } \|z - x\|_2 \leq R\}$$

then x^* is the global optimum.

The most basic type of convex optimization problem is a linear program where we want to minimize the function $f(x) = c^T x$ for some vector c subject to $Ax = 0$ and $Gx \leq h$. Another type is a quadratic program.