DS-GA 3001.03: Extended Syllabus Lecture 3

Optimization and Computational Linear Algebra for Data Science (Fall 2016)

Afonso S. Bandeira
bandeira@cims.nyu.edu
http://www.cims.nyu.edu/~bandeira

September 27, 2016

These are not meant to be Lecture Notes. They are simply extended syllabi with the most important definitions and results from the lecture. As such, they lack the intuition and motivation and so they are not a good place to learn the material the first time, just to briefly review it. These extended syllabi will also have references.

There are many amazing books about linear algebra and virtually all of them will contain the material for this particular lecture, examples include the book suggested for the course [2]. Another place you can read about some of this is the Lecture Notes from last years DSGA1002 [1].

Please let me know if you find any typos!

- If $x, y \in \mathbb{R}^n$ and $x^T y = 0$ then the angle between them is $\frac{\pi}{2}$ and we say that the vectors are orthogonal. The observation that in that case $\|x - y\|^2 = \|x\|^2 + \|y\|^2$ is Pythagoras theorem.
- An orthogonal basis of a vector space is a set of vectors that span the vector space and are pairwise orthogonal. If, moreover, each vector has norm 1 it is called an orthonormal basis.
- Given an orthonormal basis $v_1, \ldots, v_n \in \mathbb{R}^n$, the matrix $V \in \mathbb{R}^{n \times n}$ whose columns are $v_1, \ldots, v_n$ satisfies $V^T V = I$ (and $V V^T = I$).
- When a matrix $V \in \mathbb{R}^{n \times n}$ satisfies $V^T V = I$ it is called an orthogonal matrix.
- Given an orthonormal basis $v_1, \ldots, v_n \in \mathbb{R}^n$, a vector $u \in \mathbb{R}^n$ is easily writable as a linear combination of the basis as $u = \langle u, v_1 \rangle v_1 + \cdots + \langle u, v_n \rangle v_n$. Recall that $\langle u, v_k \rangle = u_k^v$.
- An orthogonal basis can be constructed using the Gram-Schmit process (see, for example [2]). Also important, given a set of orthogonal vectors in a subspace, it is also possible to find an orthogonal basis for that subspace containing the original vectors (by making use of the Gram-Schmit process).
- Two subspaces $U, V \subset \mathbb{R}^n$ are said to be orthogonal ($U \perp V$) if, for all $u \in U$ and all $v \in V$ we have $u^T v = 0$.
- Given a matrix $A \in \mathbb{R}^{n \times m}$ we have $\mathrm{Im}(A) \perp \ker(A^T)$.
- Given a subspace $U \subset \mathbb{R}^n$, its orthogonal complement $U^\perp$ is the subspace containing all vectors v such that $v^T u = 0$ for all $u \in U$.

- Given a matrix $A \in \mathbb{R}^{n\times m}$ we have $\mathrm{Im}(A)^\perp = \ker(A^T)$.
- Given a subspace $U \subset \mathbb{R}^n$, $\dim(U) + \dim(U^\perp) = n$.
- Notice that, since $\dim \mathrm{Im}(A) = \dim \mathrm{Im}(A^T)$ the two statements above imply the fundamental Theorem of Linear Algebra (just take $A = B^T$): For $B \in \mathbb{R}^{p\times q}$

$$\dim \mathrm{Im}(B) + \dim \ker(B) = q.$$

- Given a vector $v \in \mathbb{R}^n$ and a subspace $U \subset \mathbb{R}^n$ the projection of x in U, also called orthogonal projection of x in U, and denoted by $P_U(x)$ is such that $x - P_U(x)$ is orthogonal to all elements of U, meaning that $x - P_U(x) \in U^\perp$. (draw a picture!).
- Least Squares: For $m < n$, let $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. The solution $x^\natural$ of $\min_{x\in\mathbb{R}^m} \|Ax - b\|_2$ satisfies $A^TAx^\natural = A^Tb$.

  There are two nice ways of proving this, one "by calculus") simply differentiate $\|Ax - b\|_2^2$, and a second one "by geometry") notice that the point Ax should be the (orthogonal) projection of b on $\mathrm{Im}(A)$ and so $b - Ax$ should be orthogonal to $\mathrm{Im}(A)$ which is the same as being in $\ker(A^T)$ and thus $A^T(Ax - b) = 0$.
- Notice that this system always has a solution, since $\mathrm{Im}(A^T) = \mathrm{Im}(A^TA)$.
- A good example of a least squares problem is linear regression. Let's say I have a function $f\mathbb{R} \to \mathbb{R}$ and that I have (possibly noisy) measurements $f_1, \ldots, f_n$ of f at the points $t_1, \ldots, t_n \in \mathbb{R}$. For simplicity let us assume $\sum_{k=1}^n t_k = 0$. If one believe that the function should be linear $f(t) = \mu + \alpha t$ then it makes sense to solve:

$$\min_{\mu,\alpha} \sum_{k=1}^n |f_k - (\mu + \alpha t_k)|^2.$$

  This is the same as

$$\min_{\mu,\alpha} \left\| \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{n-1} \\ 1 & t_n \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \end{bmatrix} - \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix} \right\|,$$

  which we know is given by the solution of

$$\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{n-1} \\ 1 & t_n \end{bmatrix}^T \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{n-1} \\ 1 & t_n \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{n-1} \\ 1 & t_n \end{bmatrix}^T \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix}.$$

  which, because of the assumption is the same as

$$\begin{bmatrix} \mu \\ \alpha \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & \sum_{k=1}^n t_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{k=1}^n f_k \\ \sum_{k=1}^n t_k f_k \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n f_k \\ \\ \frac{\sum_{k=1}^n t_k f_k}{\sum_{k=1}^n t_k^2} \end{bmatrix},$$

  provided that $\sum_{k=1}^n t_k^2 \neq 0$. Notice that if $\sum_{k=1}^n t_k^2 = 0$ then all the t_k's must be 0 and so there is data about only one point.

References

[1] Carlos Fernandez-Granda, *Lecture Notes of DSGA1002*, available at `http://www.cims.nyu.edu/~cfgranda/pages/DSGA1002_fall15/notes.html`, 2015

[2] Gilbert Strang, *Introduction to Linear Algebra*, Fifth Edition, 2016